

South-East  
European University



Faculty of Contemporary  
Sciences and Technologies

Third Cycle of Studies

Doctoral Dissertation Topic:

*“Rule discovery for exploratory causal reasoning and the ethical consideration in observational studies with public data”*

Candidate:  
MSc. Shkurte Luma-Osmani

Mentor:  
Asst. Prof. Dr. Florije Ismaili

Tetovo, 2022




## *AUTHOR'S DECLARATION*

I herewith confirm and declare that the dissertation presented in this manuscript, is my own independent and original research and the same is drafted for the purpose of getting a PhD degree in the doctoral program E-Technologies, Faculty of Contemporary Sciences and Technologies at the South East European University.

Furthermore, I want to certify that to the best of my knowledge, the used bibliography has been cited accordingly and indicated in the list of references.

Tetovo, February 2022



Shkurte Luma-Osmari

## *ACKNOWLEDGEMENTS*

Roughly five years ago, this journey was set in motion and is ending with a group of people who have supported me and made this initiative possible. This PhD thesis constitutes an incredibly important step in my academic career.

Therefore, today, with much gratitude, I want to convey the most superior thanks to my scientific leader, Assoc. Prof. Florije Ismaili who guides and inspires young researchers like me. I express gratitude toward her for believing in me, for the perseverance, and for bringing out the best of me.

Neither to forget making excerpt to in gratification for my colleagues and co-authors as well, Asst. Prof. Parashu Ram Pal, Assoc. Prof. Pankaj Pathak, Assoc. Prof. Xhemal Zenuni and Assoc. Prof. Bujar Raufi for their support and scientific guidelines in the realization of this dissertation.

My beloved family, immense thanks and gratitude to just about every one of you! My parents, my in-laws, and my sisters, everything they have given me, for the emotional support, encouragement and empathy shown.

Lastly, I want to express a heartfelt thanks for my husband, Blerim and my daughter Amelia for the calmness and positivity they consistently give me. I want to apologize them for all the missing family moments and for the sacrifice they had to undergo jointly with me, as well as to express the endless and unconditional love I feel for them. Amelia, I'll make up those days, I promise!

<b>AUTHOR'S DECLARATION .....</b>	<b>3</b>
<b>ACKNOWLEDGEMENTS .....</b>	<b>4</b>
<b>LIST OF FIGURES .....</b>	<b>8</b>
<b>LIST OF TABLES .....</b>	<b>10</b>
<b>LIST OF ACRONYMS .....</b>	<b>11</b>
<b>ABSTRACT .....</b>	<b>13</b>
<b>ABSTRAKTI .....</b>	<b>15</b>
<b>АННОТАЦИЯ .....</b>	<b>17</b>
<b>CHAPTER 1 .....</b>	<b>19</b>
<b>INTRODUCTION .....</b>	<b>19</b>
1.1 Overview .....	19
1.2 Problem Statement .....	20
1.3 Dissertation Structure .....	20
1.4 Research Fields.....	21
1.5 Publications.....	22
1.5.1 Conferences .....	22
1.5.2 Journals .....	22
<b>CHAPTER 2 .....</b>	<b>23</b>
<b>FUNDAMENTALS .....</b>	<b>23</b>
2.1 Knowledge Discovery in Databases .....	23
2.2 Data Mining.....	25
2.3 Data Mining Techniques .....	27
2.3.1 Anomaly Detection .....	27
2.3.2 Association Rule Mining.....	27
2.3.3 Classification .....	28
2.3.4 Clustering .....	28
2.3.5 Regression .....	29
2.4 Association rules .....	29
2.4.1 Apriori Algorithm .....	32
2.5 Causal Learning .....	34
2.5.1 Directed Acyclic Graphs .....	34
2.5.2 Conditional Probability .....	35
2.5.3 Conditional Independence.....	36

2.5.4 Bayes Theorem .....	36
2.5.5 Naïve Bayes Classification .....	37
2.5.6 Bayesian Networks .....	37
2.6 Chapter Discussion .....	40
<b>CHAPTER 3 .....</b>	<b>41</b>
<b>LITERATURE REVIEW .....</b>	<b>41</b>
3.1 Causal Discovery .....	41
3.2 Causal Rules Definition .....	44
3.3 Related Work .....	45
3.3.1 Association Rule Mining .....	45
3.4 Systematic Literature Review Steps .....	46
3.5 WordCloud Analysis .....	52
3.6 Chapter Discussion .....	54
<b>CHAPTER 4 .....</b>	<b>56</b>
<b>RESEARCH METHODOLOGY .....</b>	<b>56</b>
4.1 Objectives of the Study .....	57
4.2 Research Questions .....	57
4.3 Hypothesis .....	75
4.4 Unexplored Causality Areas .....	75
4.4.1 Smart Agriculture .....	76
4.5 Limitations of the Study .....	81
4.6 Chapter Discussion .....	82
<b>CHAPTER 5 .....</b>	<b>83</b>
<b>ANALYSIS OF RESULTS .....</b>	<b>83</b>
5.1 Tools and programming language used .....	83
5.2 Method .....	89
5.2.1 Ethics .....	89
5.2.2 Participants .....	89
5.2.3 Procedure .....	92
5.2.4 Limitations .....	92
5.2.5 Data Preprocessing .....	93
5.3 Cyberstalking Dataset .....	93
5.4 Correlation .....	97

5.5 Discovering causality.....	100
5.6 Multiple logistic regression model.....	107
5.7 Model Evaluation using Confusion Matrix.....	110
5.8 Proposed Algorithm.....	115
5.9 Chapter Discussion.....	119
<b>CHAPTER 6 .....</b>	<b>120</b>
<b>ETHICAL ISSUES IN PUBLICLY AVAILABLE DATA .....</b>	<b>120</b>
6.1 Ethics in Data Science .....	120
6.2 Publicly available data ethics .....	122
6.3 Methodology.....	127
6.4 Interpretation of Results.....	129
6.4.1 The current status of the Data Ethics publications.....	129
6.4.2 Co-authorship, Keywords and Title Analysis.....	132
6.5 Chapter Discussion.....	139
<b>CHAPTER 7 .....</b>	<b>140</b>
<b>CONCLUSIONS, CONTRIBUTIONS AND FUTURE WORK.....</b>	<b>140</b>
7.1 Conclusions .....	140
7.2 Thesis Contributions .....	142
7.3 Future Work .....	142
<b>APPENDIX A .....</b>	<b>143</b>
<b>APPENDIX B .....</b>	<b>145</b>
<b>APPENDIX C .....</b>	<b>175</b>
<b>REFERENCES.....</b>	<b>182</b>

## *LIST OF FIGURES*

Figure 2.1: Graphs Types.....	35
Figure 2.2: Dependent sets .....	35
Figure 2.3: Independent sets .....	35
Figure 2.4: Bayesian Network .....	38
Figure 2.5: Possible causal structures .....	39
Figure 3.1: Creswell’s five steps to Literature Review .....	47
Figure 3.2: Digital Libraries Used .....	49
Figure 3.3: Literature Evaluation and Selection Process.....	50
Figure 3.4: The Literature Map .....	51
Figure 3.5: Title WordCloud Analysis .....	52
Figure 3.6: Problem Definition WordCloud Analysis .....	53
Figure 3.7: Conclusion WordCloud Analysis .....	54
Figure 4.1: Types of Causal Relationships .....	64
Figure 4.2: Datasets used in Literature .....	72
Figure 4.3: Types of Datasets .....	73
Figure 4.4: Data Repositories .....	73
Figure 4.5: Unexplored Areas .....	76
Figure 5.1: Top 10 Programming Languages.....	83
Figure 5.2: Digitalization around the world .....	86
Figure 5.3: Participant’s data .....	89
Figure 5.4: Gender .....	91
Figure 5.5: Age .....	92
Figure 5.6: Form of harassment.....	96
Figure 5.7: Cyberstalking frequency .....	96
Figure 5.8: Age based cyberstalking frequency .....	97
Figure 5.9: Pearson correlation coefficient.....	98
Figure 5.10: Heatmap Correlation .....	100
Figure 5.11: Victim of Cyberstalking variable .....	103
Figure 5.12: Receiver operating characteristics.....	111
Figure 5.13: COJEC Algorithm Pseudocode.....	116
Figure 5.14: Cyberstalking causal rules.....	118



Figure 6.1: Linking anonymized data .....	122
Figure 6.2: Mostly used datasets .....	125
Figure 6.3: Publications per year .....	129
Figure 6.4: Number of Publications .....	131
Figure 6.5: Sum of Citations .....	131
Figure 6.6: H-Index of Data Ethics Publications .....	131
Figure 6.7: Co-authorship Analysis .....	133
Figure 6.8: Author's density visualization map .....	133
Figure 6.9: Keyword co-occurrence network.....	135
Figure 6.10: Keyword co-occurrence timeline .....	136
Figure 6.11: Title terms timespan .....	138

## *LIST OF LISTINGS*

Listing 2.1: Apriori Algorithm .....	33
Listing 5.1: System Properties .....	85
Listing 5.2: Participant's data plotting code .....	90
Listing 5.3: Gender plotting code .....	91
Listing 5.4: Participants Age plotting code.....	92
Listing 5.5: Python Libraries .....	109
Listing 5.6: Variable Definition .....	109
Listing 5.7: Logistic Regression.....	110
Listing 5.8: Prediction and Confusion matrix .....	110
Listing 5.9: ROC .....	111
Listing 5.10: Model Coefficients.....	112
Listing 5.11: Logit Model .....	112

## *LIST OF TABLES*

Table 2.1: PC Hardware Transaction.....	30
Table 2.2: Binary transaction matrix.....	31
Table 2.3: Number of DAGs compared to the number of nodes.....	40
Table 3.1: Keyword Search Statement.....	47
Table 4.1: CAR, CCC and CCU Causal Rules .....	62
Table 5.1: Cyberstalking Dataset Information's.....	89
Table 5.2: Variable's description.....	93
Table 5.3: Two variables ratio .....	102
Table 5.4: Three variables ratio .....	103
Table 5.5: Four variables ratio .....	104
Table 5.6: Cyberstalking Bayesian Network.....	106
Table 5.7: Coefficients and p-value of the main variables.....	112
Table 5.8: Logit Regression Results.....	113
Table 5.9: Marginal effect of the main variables .....	114
Table 6.1: Retrieved documents types .....	127
Table 6.2: Stages of bibliometric analysis on data ethics research .....	128
Table 6.3: Top 10 Authors of data ethics publications .....	133
Table 6.4: Top 10 Cited Papers of data ethics publications.....	134
Table 6.5: Data ethics keywords .....	137

## *LIST OF ACRONYMS*

ADR - Adverse Drug Reactions

AI - Artificial Intelligence

ANN - Artificial Neural Networks

AR - Association Rule

ARM - Association Rule Mining

CBN - Causal Bayesian Network

CDT - Causal Decision Tree

CI - Conditional Independence

CPT - Conditional Probability Table

CR - Causal Rule

CRE - Causal Rule Explorer

DAG - Directed Acyclic Graph

DT - Decision Trees

DV - Dependent Variable

FCI - Fast Causal Inference

ICD - International Classification of Diseases

IEEE - Institute of Electrical and Electronics Engineers

IMF - International Monetary Fund

IoT - Internet of Things

IV - Independent Variable

KDD - Knowledge Discovery in Databases

KNN - K-Nearest Neighbor

LHV - Left Hand Value

ML - Machine Learning

MLE - Maximum Likelihood Estimation

MLR - Multiple Linear Regression

NBC - Naive Bayes Classifier

NLP - Natural Language Processing

OLS - Ordinary Least Square

PDF - Probability Density Function

POC - Pre Order Coding

RCT - Randomized Controlled Trials

RHV - Right Hand Value

ROC - Receiver Operating Characteristic

SEM - Structural Equation Model

SLR - Simple Linear Regression

SVM - Support Vector Machines

WTO - World Trade Organization

## ***ABSTRACT***

Data science is opening new avenues for academia in the global research trend. Considering last years' time span, there has been noticed an archetype alteration on causal reasoning, along with the disclosure of causal rules among the changeable factors, and the same approach has been seen as a great potential to assist in understanding and solving diverse intricate real-life difficulties. Internet, as a massive repository of data, has significantly increased the phenomenon of using public datasets in various surveys, and consequently has captured the attention of numerous data miners. This study proposes to provide an explicit synopsis on the key features of the ethical concerns related to public data on a bibliometric analysis. Likewise, state-of-the-art approaches and studies of machine learning methods to causal inference techniques, as well as unexplored causality areas such as smart agriculture and ethics, has been presented in a form of systematic literature review related to the digital libraries that are among the most prominent in the field. This dissertation presents an etiological cyberstalking study, meaning the use of various technologies and internet in general to harass or to stalk someone. However empirical study of cyberstalking victimization has received less attention from the research community. In most of the studies, a priority is given on a single causation identification, whereas the data examination used for mining causal relationships in this paper presents a novel and great potential to detect combined or multiple cause factors, in this case, trajectories of the factors of cybercrime. Moreover, the dissertation focuses in the impact that variables such as age, gender and the fact whether the participant has ever harassed someone, is related to the fact of being victim of cyberstalking. The research aims to find the causes and effects of cyberstalking in high school's teenagers in the city of Tetova, North Macedonia. Furthermore, an exploratory data analysis has been performed. Correlation between the factors on the dataset is considered. The odds ratio among the variables has been calculated, which implies that girls are twice as likely as boys to be cyberstalked. Similarly, concerning outcomes related to cyberstalking frequency recidivism is noticed. A logistic model was built and evaluated by utilizing the confusion matrix. A novelty of the paper is presented through newly proposed causal algorithm COJEC based on joint entropy.

**Keywords:** Data science, causality, association rules, data mining, ethics, causal discovery, observational data, cyberstalking;

## **ABSTRAKTI**

Shkenca e të dhënave po hap shtigje të reja për akademinë në trendin global të hulumtimit. Duke konsideruar vitet e fundit, është vërejtur një ndryshim paradigme në arsyetimin kauzal, së bashku me zbulimin e relacioneve kauzale midis variablave dhe potencialit të tij për të ndihmuar në kuptimin dhe zgjidhjen e problemeve të ndryshme komplekse të jetës reale. Interneti, si një depo masive e të dhënave, ka rritur ndjeshëm dukurinë e përdorimit të të dhënave publike në kërkime të ndryshme, dhe për rrjedhojë ka tërhequr vëmendjen e shumë minatorëve të të dhënave. Studimi i paraqitur në këtë disertacion, synon të përshkruaj një pasqyrë të qartë mbi tiparet kryesore të shqetësimeve etike që lidhen me të dhënat publike bazuar në një analizë bibliometrike. Po kështu, qasjet dhe studimet më të fundit të metodave të të mësuarit e makinës për teknikat e konkluzioneve kauzale, si dhe fusha të pa eksploruara të shkakësisë, të tilla si smart agrikultura dhe etika, janë paraqitur në një formë të rishikimit sistematik të literaturës në lidhje me libraritë dixhitale më të spikatura në këtë fushë. Ky disertacion paraqet një studim etiologjik të sulmit kibernetik, që nënkupton përdorimin e teknologjive të ndryshme dhe të internetit në përgjithësi për të ngacmuar ose për të përndjekur dikë. Megjithatë, studimi empirik i viktimizimit të sulmeve kibernetike ka marrë pak vëmendje nga komuniteti hulumtues. Në shumicën e studimeve, një përparësi i jepet një identifikimi të vetëm shkakësor, ndërkaq ekzaminimi i të dhënave të përdorura për gjurmimin e relacioneve shkakësore në këtë punim paraqet një potencial të ri dhe të madh për të zbuluar faktorët shkaktarë të kombinuar ose të shumëfishtë, në këtë rast, trajektoret e faktorëve të krimit kibernetik. Për më tepër, disertacioni fokusohet në ndikimin që variabilet si mosha, gjinia dhe fakti nëse pjesëmarrësi ka ngacmuar ndonjëherë dikë, lidhet me faktin e të qenit viktimë e sulmit kibernetik. Hulumtimi synon të gjejë shkaqet dhe efektet e sulmit kibernetik tek adoleshentët e shkollave të mesme në qytetin e Tetovës, Maqedoninë e Veriut. Për më tepër, është kryer një analizë eksploruese e të dhënave. Është marrë parasysh korrelacioni ndërmjet faktorëve në datasetin e të dhënave, gjithashtu është llogaritur raporti i gjasave midis variablave, që na len të nënkuptojmë se vajzat kanë dy herë më shumë gjasa se djemtë për t'u ngacmuar në internet. Në mënyrë të ngjashme, në lidhje me rezultatet që lidhen me frekuencën e sulmit kibernetik vërehet recidivizëm. Modeli logjistik u

ndërtua dhe i njëjti u vlerësua duke përdorur matricën e konfuzionit. Risja e punimit është paraqitur përmes algoritmit kauzal të propozuar COJEC bazuar në entropinë e përbashkët.

**Fjalët kyçe:** Shkenca e të dhënave, kauzaliteti, rregullat e shoqërimit, minimi i të dhënave, etika, zbulimi shkakësor, të dhënat vëzhguese, përndjekja kibernetike;



## **АПСТРАКТ**

Науката за податоци отвора нови патишта за академијата во глобалниот тренд на истражување. Со оглед на последните години, имаше промена на парадигмата на каузалното расудување, заедно со откривање на причинско-последична врска помеѓу варијаблите и нејзиниот потенцијал да помогне во разбирањето и решавањето на различни сложени проблеми од реалниот живот. Интернетот, како огромно складиште на податоци, значително го зголеми феноменот на користење на јавни податочни сетови во различни истражувања, и следствено го привлече вниманието на бројни податочни рудари. Оваа студија има намера да обезбеди експлицитен преглед на главните карактеристики на етичките прашања поврзани со јавните податоци во форма на библиометриска анализа. Слично на тоа, најсовремените пристапи и студии за методите на машинско учење за техниките за каузално заклучување, како и неистражените области на каузалноста, како што се паметното земјоделство и етика, се претставени во форма на систематски преглед на литература поврзана со најистакнатите дигитални библиотеки на бази на податоци на областа. Оваа дисертација претставува етиолошка студија за кибернетско следење, што значи употреба на различни технологии и интернет воопшто за да се вознемири или да се следи некого. Сепак, емпириското проучување на виктимизацијата на сајбер следење доби малку внимание од истражувачката заедница. Во повеќето студии, приоритет е даден на единствена каузална идентификација, додека испитувањето на податоците што се користи за рударење на причински врски во овој труд претставува нов и голем потенцијал за откривање комбинирани или повеќекратни причинители, во овој случај, траектории на факторите на компјутерски криминал. Дополнително, дисертацијата се фокусира на влијанието што варијаблите како што се возраста, полот и фактот дали учесникот некогаш малтретирал некого, е поврзан со фактот дека е жртва на сајбер-протекување. Истражувањето има за цел да ги открие причините и ефектите од сајбер демантирањето кај тинејџерите од средни училишта во Тетово, Северна Македонија. Понатаму, извршена е истражувачка анализа на податоци. Се разгледува корелацијата помеѓу факторите на сетот на податоци. Пресметан е соодносот на шансите меѓу променливите, што имплицира дека девојчињата имаат двојно поголема веројатност од момчињата да бидат сајбер

нападнати. Слично на тоа, во однос на исходите поврзана со оваа појава фреквентен рецидив е забележан. Беше изграден логистички модел и истиот е оценет со користење на матрицата за конфузија. Новитетот на трудот е претставен преку предложениот COJES каузален алгоритам базиран на заедничка ентропија.

**Клучни зборови:** Наука за податоци, каузалност, правила за здружување, податочно рударство, етика, каузално откритие, набудувачки податоци, сајбер прогонство;

## CHAPTER 1

# INTRODUCTION

---

### 1.1 Overview

Every dissertation tries to tell a story, and mine tries to tell a causal one. Identification of causal effects of diverse factors, variables data or events marks very vital step in understanding and clearing up dissimilar singularities in healthcare, culture, education, or agriculture amid other areas (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020). Marching toward such data-driven world, to find as well to understand the existing relations in data presents the key element to novel knowledge discovery explanation (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

In a world described by random variables, some of which may have causal inference on others, (Pearl, 2010) discusses the underlying mathematical framework of causal inference through three fundamental concepts: causation, intervention, mechanisms. Apart from being a fundamental philosophical topic, causal reasoning can be studied and analyzed in almost all disciplines, out of which artificial intelligence in general and machine learning in particular become important in modeling and solving causal reasoning in data (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

A standard systematic literature review process presented in the study, utilizes to the mostly known digital computer science database libraries in last years. It aims the investigation and identification of certain research questions in order to broadly recapitulate and discuss all the necessary points of view vis-à-vis causal reasoning: technical, application in real-world problems and ethical issues (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

## 1.2 Problem Statement

In order to address this issue of data causal discovery several approaches and algorithms were presented like the **LDC** algorithm (Cooper G. F., 1997) **CCC** and **CCU** rules (Silverstein, Brin, Motwani, & Ullman, 1998), **FCI** along with **PC** algorithms (Spirtes, Glymour, & Scheines, 2000), variantt of LDC such as **LCDa**, **LCDb** and **LCDc** (Mani & Cooper, 2001), Total Conditioning **TC**, and a variant, **TC<sub>bw</sub>** (Pellet & Elisseeff, 2008), **CRM** (Bhoopathi & Rama, 2016), **CDT** (Li, Ma, Le, Liu, & Liu, 2016) etc.

Anyway, each method has its own drawbacks. Therefore, we propose a hybrid approach using probabilistic models and structural causal models utilized in causal relationships discovery. The main motivation for this work is the fact that there is no such approach available.

## 1.3 Dissertation Structure

This dissertation comprises of seven chapters. The first one begins with an overviewed introduction to the research topics elaborated.

The second chapter proceeds further with the fundamental aspects pertaining to the process of KDD in addition to detailed description to data mining techniques with an emphasis to association rules along with Apriori algorithm as the best approach in discovering such relationships among data.

A summary of the literature supporting our dissertation related to causal discovery is discussed in the third chapter. Chapter four outlines the research methodology in consort with the research questions and hypothesis raised. It closes with the not-explored areas of causality such as smart agriculture.

The fifth chapter consists of the analysis of the results obtained. Indeed, it explores a certain dataset and investigates the relationships between variables such as their association and correlation. A graphical visual data representation has been made. An added value is the logit model for discovering causal relationships, along with the newly proposed algorithm named COJEC.

A description the ethical concerns and principles of the data science is elaborated in detail in chapter six. Text mining approach through a bibliometric analysis has been utilized in ascertaining the last trends of the research area.

As usually, the last chapter highlights the main conclusions based on the data analysis and forthcoming avenues that could outspread the current research.

## **1.4 Research Fields**

- Data Science: multidisciplinary field based on algorithms, scientific approaches, procedures, and systems to excerpt acquaintance and perceptions from various kinds of existing data.
- Association Rules: a significant data mining technique with an objective of unfolding motivating relations among the attributes in enormous sets of data.
- Causal Relationships: also referred as causal reasoning or causal association relationship, is represented by two events or two variables and it is especially focused on the cases when one of them is causing the other.
- Public Data Ethics: refers to data that can be found in the internet, in an easy manner, and retrieved readily and for free as well as their usage for the consideration of significant ethical aspects. Newfangled technologies present precious and valuable tools but anyway, serious ethical penalties must be considered (Reynolds G. W., 2015).

## 1.5 Publications

Parts of the thesis were presented and published in several scientific conferences and publications, as listed below:

### 1.5.1 Conferences

- **Luma-Osmani S.**, Ismaili, F., Ram Pal, P., *“Building a Model in Discovering Multivariate Causal Rules for Exploratory Analyses”*, DATA21: International Conference on Data Analytics for Business and Industry, 25-26 October 2021, pp. 272-276, DOI: 10.1109/ICDABI53623.2021.9655981, Sakheer, Kingdom of Bahrain.
- **Luma-Osmani S.**, Ismaili, F., Zenuni, X. & Raufi, B. *“A Systematic Literature Review in Causal Association Rules Mining”* - IEMCON 2020: 11th Annual IEEE Information Technology, Electronics and Mobile Communication Conference, November 4-7, 2020, DOI: 10.1109/IEMCON51383.2020.9284908, eISBN:978-1-7281-8416-6, Vancouver, Canada.
- **Luma-Osmani S.**, Ismaili, F. & Raufi B., *“Bibliometric Analysis and Visualization of Ethical Concerns on Publicly Accessible Data Sets”* - ISCBIT 2020: 4th International Scientific Conference on Business and Information Technologies, ISSN: 2671-373X, pp. 168-179, September 17-18, 2020, Tetovo, Republic of North Macedonia.

### 1.5.2 Journals

- **Luma-Osmani S.**, Ismaili, F., Pathak P. & Zenuni, X., *“Identifying Causal Structures from Cyberstalking: Behaviors Severity and Association”*, Journal of Communications Software and Systems, 2021, (Scopus Indexed), ISSN: 1845-6421, Vol. 18, No. 1, pp. 1-8, DOI: 10.24138/jcomss-2021-0139, January 2022, Split, Croatia.
- **Luma-Osmani S.**, Ismaili, F., Raufi B. & Zenuni, X., *“Causal Reasoning Application in Smart Farming and Ethics: A Systematic Review”*, Annals of Emerging Technologies in Computing (AETiC). Vol. 4, No. 4, 2020, pp. 10-18, DOI: 10.33166/AETiC.2020.04.002 (Scopus Indexed), ISSN: 2516-029X, October 1, 2020, London, UK.

## CHAPTER 2

# FUNDAMENTALS

---

### 2.1 Knowledge Discovery in Databases

We are aware that we are coexisting in the age of information, therefore having data is a must. (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020). It is being generated through multiple sources this is not just coming from social media, it comes from sectors such as healthcare sector, financial sector, telecom sector, and so on and this data is also available in different formats.

The part dealing with considerable number of repetitively presented data, such as in database is defined as the process of knowledge discovery in databases - KDD (Mazlack, 2001). Whereas when those mining techniques are functional to web-based data, then it's talked about web mining (van Wel & Royakkers, 2004).

Data mining in reality refers to an actual stage in discovering knowledge from the database. The whole end-to-end process was firstly proposed by (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). It consists of several important techniques like: selection of data, pre-processing, data transforming, data mining, evaluation/interpretation of obtained patterns and representation of knowledge (DataFlair, 2019).

#### 2.1.1 Data Selection

The whole process for extracting knowledge, also known as KDD, starts at the bottom left corner with raw data that users have access on. It is an important phase since as we already know that data available in repositories comes from multiple sources and that is also presented in multiple

formats therefore, it can be selected, organized and extracted to only include the data that user will need to query and that's relevant to the problem that he/she is trying to address.

So, the first task marks the integration of this data and its storage in appropriate location such as data warehouse, and after those processes are done, we can select one particular dataset, targeted data on which we would like to do the analytical tasks.

### ***2.1.2 Pre-Processing***

The target data goes through a process called data pre-processing. Pre-processing also involves tasks such as data cleaning, which would convert the target data into pre-processed actionable processable data set. Data cleaning includes imputing the missing values, removing noise missing data, inconsistent or irrelevant data. Generally speaking, the stage of cleaning the data serves as a decent filter in reducing the errors and improving quality of the data.

Pre-processing benefits to understand the structure of data, there can be also used visualization techniques, on meantime it provides a clear picture about the relationship that variables have with one another in the dataset, the existing correlation midst them, where can also be applied simple operations like summarizing aggregation and normalization.

### ***2.1.3 Transformation***

Data transformation routine means consolidation and converting of data as per different mining techniques. Transform or into forms appropriate for mining techniques, need to convert them into one particular form in order that process may be more efficient for mining. With methods such as transformation, a significant reduction number of variables is noticed. Finalized transformed data is pushed into diverse data mining algorithms.

### ***2.1.4 Mining of Data***

In the KDD process data mining is captured and depicted as one step where pre-processed data is converted into patterns knowledge. As mentioned above, data miners apply various intelligent operations or algorithms aiming the extraction of hidden data patterns. Then data mining algorithms do their mathematical derivation or techniques such as clustering, classification,



regression and so on, intelligent operations are applied with the aim to obtain patterns. Intelligent methods will be covered in detail on the next section.

### ***2.1.5 Pattern Interpretation/Evaluation***

Pattern evaluation is the upcoming phase. Data mining produces noteworthy evaluated patterns aiming to obtain the desired outcomes. This phase includes representation of mined patterns into some meaningful form or into some useful form, then only they will turn into knowledge or what exactly we seek. After it it's time to check for the validity of these patterns that as we do control all of the three-parameters, it is requested to be sure that gained information is useful, correct and innovative. Since the obtained information is validated, now it's time finally to present the information using graphs or diagrams.

### ***2.1.6 Knowledge Discovery***

Finally, the last phase is the knowledge representation. In this step there are used several visualization and knowledge representation techniques, integrating the knowledge into other structures for supplementary actions, or merely documenting the same and reporting or presenting it to the parties that are interested (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Manual analysis and interpretation characterize the traditional method of converting data into knowledge (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

## **2.2 Data Mining**

Anyway, as a subject of data mining typically are large datasets. A simple way to illustrate this would be from the everyday life at workplace with our colleagues, where we take a photo or record a short video and upload it to one of the mostly used social media platforms, let's say Instagram. On meantime, if we look at the bigger picture, it's not just us, but there are millions of people uploading millions of such files every single day.

In simple terms data mining means mining of information from data and discover patterns which are novel, with some degree of certainty and convenient (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Why it is important? Let's try to explain it through illustrations.

If we have a database containing data related to children technology usage, when we apply some data mining techniques and finally, we find out a pattern that tells us that technology overuse is considered as a factor that children more likely display aggressive behavior. Even we've invested time, money probably, and invested effort and we found out information which we already know. It's not something new since many other researchers have already come to same conclusion.

Let's look at the next parameter the accuracy of the information, let's again take an example of Apple smart phones dataset which encompasses of the information's about the expectations that iPhone 11 will provide a Qi standard for charging other devices also, such as the Air Pods. It has been found out taking in consideration that the Android powered competition had this sorted out years ago, but on the contrary, what actually happens is that this feature about wireless charging is not provided yet. We found out information which eventually proved to be wrong and that is why and is always necessary to check the validity of the information before we reach to any sort of decision.

As we saw in the first example, we found out an information which told us that children aggressive behavior usually comes from technology overuse, since that information did not serve a purpose, instead of finding that information what we can do is you can find out patterns which would tell us what is the allowed screen time for children on daily basis. So, this information will be more useful to us.

Broadly speaking, miners always mine mineral resources with a limited precision. That is to say, no matter how advanced the technique miner using, all hidden knowledge in data mineral resources cannot be completely obtained (Li, Ma, Le, Liu, & Liu, 2016).

There are various data types that can be used for mining. Furthermore, we have structured, semi-structured and unstructured data types (Taylor, 2018).

- Structured data type has a proper format and since it has a defined format extracting information from it is quite easy, such as data in excel spreadsheet, in databases or csv files.

- Semi structured data does not have a rigid pattern associated but it does have a noticeable pattern. An example of semi structured data would be an XML document, JavaScript Object Notation file, data in NoSQL databases or email.
- Unstructured data has no inherent data format at all and that has one extracting information from it can be quite a cumbersome task. Websites, mobile data, image files (.png), audio (.mp3) and video files (.mp4) are very common example of unstructured data type.

## **2.3 Data Mining Techniques**

Now let's look at the data mining techniques, in this session we'll be looking at detection of anomalies, association rules, classification, regression and clustering.

### ***2.3.1 Anomaly Detection***

The first data mining technique to be represented is anomaly detection. This technique is the process of finding the patterns whose behavior is significantly different from the others. These behaviors are also labeled as outliers, deviations, anomalies or unusual patterns (Agrawal & Agrawal, 2015). For instance, suppose we have a large dataset of information about thousands of people who have iPhones. Their average income is calculated as well. But, on the list also appears the name of Tim Cook. In this case, the income graph would shoot up. So, Mr. Tim Cook is considered as an outlier or an anomaly. Therefore, before processing further, an outlier dispensation has to be made, similar to the removal or a bit of adjustment to anomalies and then we can move on the foremost data analysis.

### ***2.3.2 Association Rule Mining***

The next technique is Association rule mining, known otherwise as “Market basket analysis”, as per its most common application. Also, it has a very interesting example and it goes by the name of the “Beer diaper syndrome”, usually in a supermarket it was found out that when a father buys a diaper on the store, there was very high possibility that a bottle of beer is preferred to be bought. It may seem rare, but this is what the supermarket was able to find out with the help of Association rule mining.

Anyway, a fact to mention is that not much attention has been paid on the quality of the discovered association rules, since much work has been focused on finding ever more efficient behaviors to determine all of the rules possible (Shaw, Xu, & Geva, 2008).

This method is also the basis for many of the recommender system let's take YouTube, suppose you watch Rita Ora songs, the recommender system also suggests you to watch Dua Lipa videos. So, the recommender system knows that usually clients who search for the first artist (this may be because of their descent), also search for the second one, they also both belong to the R&B genre. Anyway, a clear representation of Association rules and the algorithms used in this technique is provided in section 2.4.

### ***2.3.3 Classification***

Classification, as a problem of recognizing the category of new occurrences based on a training dataset that contains observations whose category or class membership is already known. This category is often called as a class label (Agrawal & Agrawal, 2015). It comes under the purview of supervised learning and they have predefined labels. Example: suppose we would like to distinguish the students into “males” and “females”. So, we already have an observation and we try to classify the students into one of the categories. The most popular techniques for classification are decision trees.

### ***2.3.4 Clustering***

Despite classification, clustering relies under the purview of unsupervised learning. Unsupervised learning belongs the area of machine learning, i.e., algorithms that learn on their own. No predefined labels are features of this technique, therefore, all the annotations based on the similarities are divided into clusters. It means that clustering does not assume any prior knowledge of clusters, anyway in practice; it is quite common that the class label of each object is not known; therefore, it represents a very important technique. Hence, the clusters overlap with each-other, permitting data points to belong to more created clusters simultaneously (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Search engines presents a good example how clustering can be used in the web. For a certain key word to be searched on google, for instance “Barcelona” the search engine will display all the documents related to this city, as it has previously grouped them based on their similarities.

### **2.3.5 Regression**

This technique is similar to classification, so it falls under the mode of supervised learning, the only difference is that regression analysis requires all the attributes to be of numeric type. The focal objective of this technique relies on finding out how the dependent output  $y$  change with respect to the independent one, in this case  $x$ . Among several types of regression it's worth mentioning: linear regression, Poisson regression, logistic regression and similar. However, the most used one implies simple linear regression, or as we are supposed to see it converted into formula of the straight-line graph  $y = mx + c$ .

The linear relationships in this method is measured between a dependent attribute  $y$  and one or more independent factors  $x_k$ ,  $k = \{1, 2, \dots, K\}$ , mathematically represented as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i$$

In the aforementioned equation, the subscript  $i$  specifies the  $i$ -th observation, whereas  $\beta_0$ , the regression coefficient, represents the point where  $Y$  axis is intercepted (Wang & Mueller, 2016).

## **2.4 Association rules**

ARM is recognized as the most important technique as related to the mining of data, utilized for discovering interesting rules among variables databases, especially in large ones. Their main usage relies on analyzing and predicting customer behaviors. Main objective of ARM is finding recurrent patterns, correlations, causal relationships or associations among every kind of data repositories (Kaur, 2013).

According to (Agrawal & Srikant, 1994) the association rules consist of a set of two-valued items. Therefore, let  $T$  be the set of data that contains several transactions  $T = \{t_1, t_2, \dots, t_n\}$ . Those transactions in  $T$  are a set of items, noted as  $I = \{i_1, i_2, \dots, i_n\}$  and has an exceptional ID of transaction and covers a subset of the items in  $I$ . A rule is defined as an implication of the form

$X \rightarrow Y$  where  $X, Y \subseteq I$  and there is no interception between them (Ziauddin, Kammal, Khan, & Khan, 2012).

Support and confidence are the metrics mostly used to traditionally define association rules. Support is nothing but the likelihood that both A and B are included into the transaction, although confidence can be noted as the likelihood one transaction that contains A, correspondingly contains B. Lift, on the other side describes the statistical correlation among A and B. If lift has value 1, correlation is detected and for value different than 1 a positive or negative correlation is indicated (Noh, Son, Park, & Chang, 2017). The parameters can be calculated using the formulas described as follows:

$$\text{support } (A \rightarrow B) = \frac{\text{Transactions containing both A and B}}{\text{Total number of Transactions}}$$

$$\text{confidence } (A \rightarrow B) = \frac{\text{Transactions containing both A and B}}{\text{Transactions containing A}}$$

$$\text{lift } (A \rightarrow B) = \frac{\text{Transactions containing both A and B}}{\text{Transactions containing A} * \text{Fraction of Transactions containing B}}$$

The items that have more possibilities to be purchased together are the object of study. This also represents a tricky approach of sellers to list the products that are frequently bought commonly. It's worth mentioning that AR mining involves the affairs amongst items in a set of data (Girotra, Nagpal, Minocha, & Sharma, 2013).

**Table 2.1:** PC Hardware Transaction

<b>TID</b>	<b>Items</b>
<b>T1</b>	Monitor, Printer, Mouse
<b>T2</b>	System Unit, Printer, Keyboard
<b>T3</b>	Monitor, System Unit, Printer, Keyboard
<b>T4</b>	System Unit, Keyboard
<b>T5</b>	Monitor, Printer, Keyboard

From the above table, we can calculate the association rule as follows:

- {Monitor, Printer → Keyboard}
- Confidence = 0.66
- Support = 0.20

Meaning that 66% of clienteles who buy monitor and printer, also tend to buy keyboard, as well as 20% of the total number of customers buy monitor and printer.

Consequently, the transaction contains the itemset {Monitor, Printer} noted as left-hand value (LHV) or otherwise an antecedent and the single item set {Keyboard} noted as right-hand value (RHV) or as consequent.

The case of buying PC hardware can also be represented as a matrix of binary form, where records signify the customers and columns signify items purchased. The value of the transaction is 1 if the item is bought, otherwise is noted as 0.

**Table 2.2:** Binary transaction matrix

<b>TID</b>	<b>Monitor</b>	<b>System Unit</b>	<b>Printer</b>	<b>Mouse</b>	<b>Keyboard</b>
<b>T1</b>	1	0	1	1	0
<b>T2</b>	0	1	1	0	1
<b>T3</b>	1	1	1	0	1
<b>T4</b>	0	1	0	0	1
<b>T5</b>	1	0	1	0	1

An open question, still remains the issue of how to set the minimum thresholds in a best manner. Usually, these thresholds are user established (Mazlack, 2001).

Nevertheless, for useful association rules, a high confidence is desired (Bhoopathi & Rama, 2017). Consider a collection of 10,000 items and if we are looking for rules containing 2 of the items in the LHS and only 1 item in the RHS, the outcome results in approximately 1,000,000,000,000 such rules (Girotra, Nagpal, Minocha, & Sharma, 2013).

Among the mostly used algorithms for mining association rules, it's worth mentioning:

- Apriori
- SETM
- AprioriTID
- AIS
- Apriori hybrid
- FP-Growth
- Eclat
- Recursive Elimination

However, in this dissertation we will explain only the most used one, i.e., Apriori algorithm.

### ***2.4.1 Apriori Algorithm***

Let us introduce the most important algorithm form mining frequent items. Apriori was firstly introduced from (Agrawal & Srikant, 1994) by making multiple steps through the dataset to generate 1-itemset, 2-itemset and so on. Items are stored in Hash table. This algorithm is constructed upon the idea that a subset of a recurrent itemset, must also be recurrent (Kaur, 2013). Those itemsets that frequently appear, are later utilized aiming the generation of association rules that please the minimum support (coverage) and conference (accuracy) constraint. The candidate sets that have less value than the min support and min confidence are removed from the table, which offers the pruning process.



```

1)  $L_1 = \{\text{large 1-itemsets}\}$ 
2) for ( $k = 2$ ;  $L_{k-1} \neq 0$ ;  $k++$ ) do begin
3)  $C_k = \text{apriori-gen}(L_{k-1})$ ;           // New candidates
4) for all transactions  $t \in D$  do begin
5)  $C_t = \text{subset}(C_k, t)$ ;           // Candidates contained in  $t$ 
6) for all candidates  $c \in C_t$  do
7)  $c.\text{count}++$ ;
8) end
9)  $L_k = \{c \in C_k \mid c.\text{count} \geq \text{minsup}\}$ 
10) end
11)  $\text{Answer} = \bigcup_k L_k$ ;

```

**Listing 2.1:** Apriori Algorithm

Main notations used in this algorithm are described below, whereas its pseudocode is provided in listing 2.1 (1994).

- $k$  itemset – An item containing  $k$  items
- $L_k$  – Set of great  $k$ -itemsets
- $C_k$  – Set of candidates  $k$ -itemsets

Firstly, the candidate generation process of one itemset is started. Every candidate item ( $C_k$ ) that does not fulfil the minimum support threshold is removed from the frequent itemset ( $F_k$ ). Generally, it represents an expensive step since multiple passes over database should be performed (Ait-Mlouk, Gharnati, & Agouti, 2017). The set is extended by one after every iteration extracted. If any of the subsets does not appear in the previous step frequent itemset, then that itemset can be freely deleted, therefore all infrequent itemsets can be pruned.

The exclusion of undesirable/un-useful factors is known as pruning. In associative classification, this method is utilized to eliminate item-sets as well as the weak association rules that are not frequent (Ram Pal, Pathak, Yadav, & Ora, 2019).

## 2.5 Causal Learning

When defining causality, a step back in time must be made. Based on philosophy, Aristotle is credited on his theories and writings regarding causality. According to him, there are four types of causes (Rammohan, 2010): material cause, indicating the substantial - out of which something is made, formal cause signifying the blueprint of the same, efficient cause demonstrates in most of cases the main character that works on the things to occur and the final cause, which intuitively focuses on the aim why the event happened.

Note however, rendering the counterfactual definition,  $X$  is causing  $Y$ , if and only if:

- $X$  has occurred,  $Y$  would have occurred;
- $X$  has not occurred,  $Y$  would not have occurred.

### 2.5.1 Directed Acyclic Graphs

In graph theory the mathematical structures that are utilized to present set relations between nodes are referred as graphs. It consists of (un)ordered pair  $G = (V, E)$ , where  $V$  is a set of vertices (habitually named as points or nodes), and  $E$  is a representation of edges (also recognized as arcs, lines or links). A simple graph can have two kinds of edges: directed and undirected ones. Thus, a graph containing undirected lines (—) is called undirected graph as illustrated in the figure 2.1 (a), and accordingly the graph containing arrowhead directed arcs ( $\rightarrow$ ) is called a directed graph, presented in figure 2.1 (b) (Spirtes, Glymour, & Scheines, 2000). A directed graph which allows no return back to the starting point is denoted as Directed Acyclic Graph (DAG).

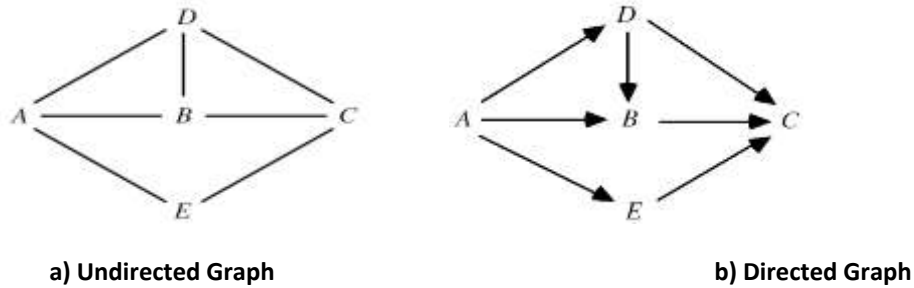


Figure 2.1: Graphs Types

### 2.5.2 Conditional Probability

The likelihood that an occurrence  $X$  will happen, taking in consideration the condition that an alternative event  $Y$  has already happened is known as a conditional probability, frequently symbolized as the probability of  $X$  given  $Y$ . Since the event  $Y$  should have already happened, that is why the formula presented below has the requirement that  $P(Y) \neq 0$ , consequently the entire sample space has shrunk to this already occurred event.

$$P(X | Y) = \frac{P(X \cap Y)}{P(Y)}$$

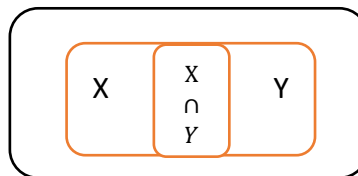


Figure 2.2: Dependent sets

In cases when events  $X$  and  $Y$  are not dependent, it means that they are not related, and they do not depend on the occurrence of the other event, resulting those two sets being mutually exclusive, i.e., cannot happen simultaneously as represented in figure 2.3 and consequently noted as  $X \cap Y = \emptyset$ , therefore  $P(X \cap Y) = P(X) * P(Y)$  the formula of conditional probability takes the form as follows:

$$P(X | Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{P(X) * P(Y)}{P(Y)} = P(X)$$

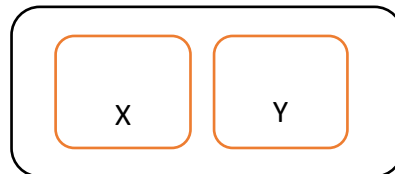


Figure 2.3: Independent sets

### 2.5.3 Conditional Independence

The variables independence can also be calculated in conditional model. Given knowledge that Z occurs, knowing that X happens doesn't provide information on the probability of happening Y, and vice versa. Moreover, X and Y are conditionally not dependent given Z, in a case that below expression holds.

$$P(X \cap Y | Z) = P(X|Z) * P(Y|Z)$$

### 2.5.4 Bayes Theorem

Conditional probability serves as a crux of the Bayesian Theorem. Moreover, it demonstrates the relation among a conditional probability and its inverse.

$$P(X | Y) = \frac{P(X \cap Y)}{P(Y)} \quad \text{and} \quad P(Y | X) = \frac{P(Y \cap X)}{P(X)}$$

However, based on the commutative property of set addition  $(X \cap Y) = (Y \cap X)$ , from the aforementioned formulas we gain:

$$P(X \cap Y) = P(X | Y) * P(Y) \quad \text{and} \quad P(X \cap Y) = (Y | X) * P(X)$$

If we equalize the left sides we have:  $P(X | Y) * P(Y) = (Y | X) * P(X)$  therefore mathematically expressed, the final equation presenting the Bayes Theorem will be:

$$P(X | Y) = \frac{P(Y | X) * P(X)}{P(Y)}$$

The aforementioned equation can still be generalized, thus allowing the event X to have as many mutually exclusive categories as is requested  $X_1, X_2, ..., X_n$ , in the entire sample space denoted by S.

$$\bigcup_{i=1}^n X_i = S$$

Therefore, the likelihood of event Y is  $P(Y) = P(X_1 \cap Y) + P(X_2 \cap Y) + \dots + P(X_n \cap Y)$ . On the other hand from the conditional probability we already know that  $P(X_1 \cap Y) = P(Y|X_1) * P(X_1)$  so the final equation will be presented as:

$$P(Y) = P(Y|X_1) * P(X_1) + P(Y|X_2) * P(X_2) + \dots + P(Y|X_n) * P(X_n) = \sum_{i=1}^n P(Y|X_i) * P(X_i)$$

If we replace this notation in the above Bayes Theorem, we gain the generalized form of the Bayes Theorem:

$$P(X_i | Y) = \frac{P(Y | X_i) * P(X_i)}{\sum_{i=1}^n P(Y|X_i) * P(X_i)}$$

### ***2.5.5 Naïve Bayes Classification***

Represents an efficient data mining as well as machine learning algorithm whose basis rely on the Bayes theorem with nondependent postulation (Shukla, Yadav, Ram Pal, & Pathak, 2019). It is named this way since it makes a “naive” assumption that attributes or features are conditionally independent given the label Y. In general, it is simple and easy to understand; convenient for implementation (Hassani, Huang, & Ghods, 2017) it’s used for building fast models and make quick predictions.

### ***2.5.6 Bayesian Networks***

Bayesian networks have been counted as a central contribution to the field of Artificial Intelligence (AI) in the last decade. Aiming to model the probability distribution of the conditional independence, the Bayesian networks uses the graphical demonstrations of the DAG’s along with conditional probability tables (CPTs). In this depiction each line represents the conditional dependency i.e., the direct influence of one variable on another, and each node represents a distinctive random variable.

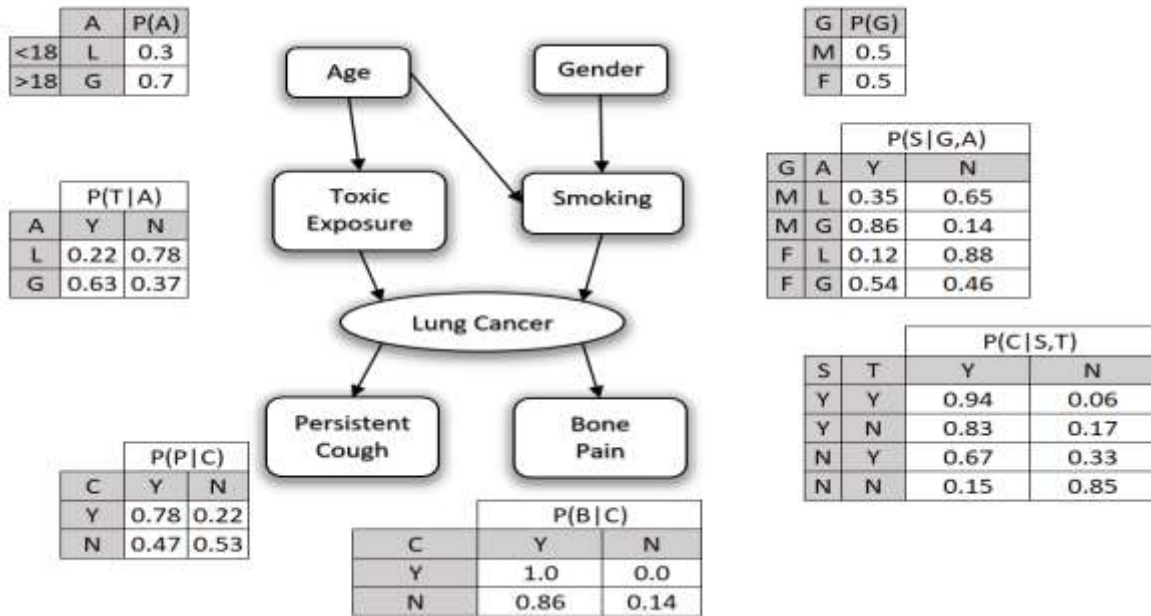
In order for the Bayesian network to model a probability it must satisfy the Markov Condition implicating that each inconstant is conditionally non-dependent of its non-descendants, given the parents (Pellet & Elisseeff, 2008), (Cooper G. F., 1997), (Spirtes, Glymour, & Scheines, 2000), (Mani & Cooper, 2001), (Bowes, Neufeld, Greer, & Cooke, 2000), (Singh, Gupta, Tewari, & Shroff, 2018). Mathematically we can say:

$$P(X_{n+1} = x | X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = P(X_{n+1} = x | X_n = x_n)$$

In the following example, we have modeled the Lung Cancer diagnosis using Bayesian Network. The conditional probability values are randomly assigned, taking in consideration the fact that  $0 \leq P(X) \leq 1$  and by fulfilling the criteria below:

$$\sum_{i=1}^n P(X_i) = 1$$

It contains 7 variables named: Age (A), Gender (G), Toxic Exposure (T), Smoking (S), Lung Cancer (C), Persistent Cough (P) and Bone Pain (B), as displayed in figure 2.4.



**Figure 2.4:** Bayesian Network

Hence, the joint probability distribution of those variables also known as the conditional probability of the variables is calculated through noted equation:

$$P(A, G, T, S, C, B, P) = P(A) * P(G) * P(T|A) * P(P|C) * P(B|C) * P(C|S, T) * P(S|A, G)$$

The pattern resulting from this graph leads to the idea that the probability of a certain variable, presented by a node, depends only on the probability of its parent(s) node. As a result, the Bayes Network can be formulated:

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n p[X_i | \text{Parents}(X_i)]$$

The algorithms for discovery of causal structures can be separated into two categories, score-based and constraint-based. The first one use statistical tests to discover the causal relationship among variables through in(dependence) constraints and the second one assigns a certain score to every generated causal graph and then selects the graph by taking into account those scores (Shen X. , 2020).

Based on the fact whether common consequence or common cause is observed, causal structures can be presented as colliders and confounders, illustrated in figure 2.5. Colliders form the so-called V-structure when variables  $X_1$  and  $X_2$  turn out to be dependent conditional on  $X_3$ . Confounder  $Y_1$  causes both variables  $Y_2$  and  $Y_3$ , anyway a change in one variable won't reflect a difference on another.



**Figure 2.5:** Possible causal structures

One of the main issues and problems of learning the structure of the Bayesian network from data is the enormously large search space (Rammohan, 2010). When the same is compared to the number of Directed Acyclic Graph nodes, an exponential growth line is noticed. The Robinson formula (Robinson, 1977) provides the insight.

$$G(n) = \begin{cases} 1, & \text{when } n = 0, \\ \sum_{i=1}^n (-1)^{i+1} \binom{n}{i} 2^{i(n-1)} G(n-i) & \text{otherwise} \end{cases}$$

Let us take into consideration the exponential growth for the first 10 nodes (Rammohan, 2010), as displayed in table below.

**Table 2.3:** Number of DAGs compared to the number of nodes

$n$	$G(n)$
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	$1.1388 \times 10^9$
8	$7.8370 \times 10^{11}$
9	$1.2134 \times 10^{15}$
10	$4.1751 \times 10^{18}$

## 2.6 Chapter Discussion

The chapter gives an overview of the process of discovering knowledge in databases, consisting of phases such as: selection of the data, it's pre-processing, transformation, mining of data, pattern interpretation or evaluation and therefore reaching knowledge discovery. Moreover, several types of data were explored and divided into structured, semi-structured and unstructured data. The chapter also discusses the data mining techniques like: anomaly detection, association rule mining, classification, clustering and regression. Since our topic is related to association rules, usually counted as one of the vital practices of data mining, more emphasis is given to Apriori algorithm.

In order to explain the causal association rules, an introduction to basic terminology is provided, where the main concepts were described in detail, consisting of directed acyclic graphs, conditional probability, conditional independence, Bayes theorem, naïve Bayes classification, and Bayesian networks.



# LITERATURE REVIEW

---

### 3.1 Causal Discovery

The idea of causality assumes a focal part in science and it is as yet set apart as one of the imperative examination inquiries in scientific investigation of different qualities (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020). Perceiving genuine causality is something troublesome and it addresses a difficult idea to characterize. On the off chance that we can't alter any factors in a framework under examination (as in the climate), understanding causal connections can assist us with determining or predicting the future by observing contemporary value of the variable. Considering that we can modify certain factors, that point thinking about causal relations will permit us to apply control over the framework's behavior (Karimi, 2010).

The causal rules, also referred as causal reasoning or causal association relationship, is represented by two events or two variables. The first event that does the causing is termed the cause or IV - independent variable and the second event that gets caused or is influenced by the cause is called the effect or the dependent variable (DV). Accordingly, the causal relationships are often called as cause-and-effects relationships. If converted to rule, the causal relationship often refers to IF... THEN rule. "Cause" is not the only term that typically expresses causality. There are also other tricky entitlements like: accordingly, then, thus, and so, as a result, therefore, because, hence, consequently, for this reason and so on.

Computer science, physics, statistics and philosophy are several of disciplines where causality has been considered. Time has played a significant role as a crucial and essential fragment of the intuitive understanding of causality, because it is usually measured compulsory that the cause  $X$  should have occurred previously from the effect  $Y$ , as related to the time factor (Karimi, 2010).

The graphical representation of causal relationship, is the same as for association relationship and is presented by a simple arrow between two variables. For learning causal reasoning, researcher's effort is on defining if there happens a causal relationship among two events, noting that the presence of  $A$  makes a difference of  $B$ , or simply noting, variable  $A$  seemed to cause  $B$ .

$A \rightarrow B$  indicates  $A$  causes  $B$

Example:

Air pollution in Polog region  $\rightarrow$  Patients with lung disease i.e., Air pollution causes the increasing number of patients with lung disease in this region.

In association rule mining  $A \rightarrow B$  is an association relationship means  $A$  is associated with  $B$ . However, this rule indicates only a statistical relationship. They do not indicate the nature of the relationship (Bhoopathi & Rama, 2016), (Silverstein, Brin, Motwani, & Ullman, 1998).

All the existing rules in the database that content min support and confidence constraints are found by association rule mining. In this manuscript, in place of mere associations, we focus on the issue of shaping rules of causal type among the variables and how they affect each other (Dehkharghani, Mercan, Javeed, & Saygn, 2014). For example, buying 2 items together e.g., mouse and headphones does not point to the conclusion that buying headphones comes as a consequence of buying a mouse.

Let us take in consideration the Traffic example (Mazlack, 2001) for understanding the difficulty of finding causality: A person calls his friend on the telephone and asks him to drive over and visit him. On his way, the driver ignores the Stop sign and drives through an intersection. He is hit by another car. He dies. Who caused his death?

- The driver?
- The other car driver?
- The friend?
- The Traffic engineer who designed the intersection?
- Fate?

Through use of Randomized controlled trials (RCT), which present an experimental technique, is the most reliable strategy to determine causal discoveries. With any case, the key issue in this system is the trouble in doing preliminaries, which is fundamentally because of ethical concerns or cost contemplations (Alharbi & Rajasekaran, 2015), (Mazlack, 2001), (Li, et al., 2013), (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020), (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

Under those same parameters, observational studies are regarded as the greatest alternatives compared to RCTs, and the same has demonstrated that observational studies that are well-designed can reach proportionate and certified findings in the same way as RCTs do (Li, et al., 2015), (Singh, Gupta, Tewari, & Shroff, 2018), (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020). Causal revelation with observed information presents an elective arrangement when the test methods are infeasible (Alharbi & Rajasekaran, 2015).

The research studies can be divided into two groupings: experimental studies or observational ones (Institute for Work&Health, 2016).

- Experimental studies are ones where investigators determine who is exposed and who is not. They introduce an intervention and directly control the conditions under which the experiment is being held. Thereafter they study the effects. In those studies, the participants generally are picked randomly.
  - Randomized controlled trial (RCT): The participants taken into consideration are randomly divided into certain groups, such as the first divided group receives the intervention some kind of drug, whereas the control group does not receive

anything or receives a placebo. These experiments are usually costly and sometimes impossible or irrational to perform.

- Observational studies typically are based on observed data in passive manner but can afford powerful results. This kind of study helps to examine relations that are unable to be tested under randomized controlled experiments. The observational studies moreover can be classified into case control and cohort studies.
  - Cohort study: Represents any group of participants who have any similarity to share or have common characteristics, as a basis for group assigning. They are divided into exposed subjects as well as non-exposed, otherwise known as the control group which are carefully chosen and then tracked in order to observe the incidence of the result (Li, et al., 2015).
  - Case control studies: As the name implies, this study has got two groups of people or study subjects, the one is the cases, and the other one is control. The two existing groups are compared based on some theoretical attribute. It's usually used in researches related to rare diseases.

Darwinian evolution and Newtonian gravity theory, on the other hand, were created nearly entirely without experimentation (Ramsey, et al., 2018).

### 3.2 Causal Rules Definition

A data set denoted by  $D$ , is elected as the set containing variables from binary types,  $D = (X_1, X_2, \dots, X_m, Z)$ , where:

- $X_1, X_2, \dots, X_m$  are cause variables and
- $Z$  is the effect variable

At the other hand, let  $P$  be a joint inconstant consisting of numerous variables noted by  $X_1, X_2, X_3, \dots, X_n$  where  $n \geq 2$

- If  $(X_1 = 1, \dots, X_n = 1)$  then  $P = 1$ , otherwise  $P = 0$ .

This rule is represented in the way of  $(P = 1) \rightarrow (Z = 1)$  or  $p \rightarrow z$ , where  $z$  presents the equation  $Z = 1$  and  $p$  presents the equation  $P = 1$ . The definitive goal is to find out if  $p \rightarrow z$  indicates a causal relationship.

In the approach presented in this dissertation, it is initially considered the association amid  $P$  and  $Z$  because to have a causal relationship, an association is mandatory (Li, et al., 2013), (Bhoopathi & Rama, 2017), (Li, et al., 2015).

### **3.3 Related Work**

This section examines the literature on association rules as well as causal relationship mining. Association rule mining inspects the foundation and advances in ARM, while causal mining looks at strategies as well as the futuristic and state-of-the-art in the issue domain.

#### **3.3.1 Association Rule Mining**

ARM - Association Rule Mining was foremostly announced in 1993 by Agrawal, R., Imielinski, T., and Swami, A., however, the pioneering work in this field was settled with the discovery of the Apriori algorithm one year later (Fast Algorithms for Mining Association Rules, 1994). The researchers opened up the practicability of mining significant association rules on large sets of data. Apriori as a standard algorithm for extracting rules of association, was considered to function on transaction-based datasets. In the same study afterwards is discussed how can Apriori and AprioriTid best features be collected into a hybrid approach named AprioriHybrid.

A general chronological survey of the research work of ARM since its beginning was presented by (Ziauddin, Kammal, Khan, & Khan, 2012). Nerveless, they argue that even though it has been developed as a novel technology, ARM still continues to be in a step of development and exploration. Same opinion supports (Kaur, 2013), in a survey related to state-of the-art research, noting that there still exist some important matters that need to be taken into account when detecting suitable rules of association.

A comparative approach based on AR algorithms is completed by (Girotra, Nagpal, Minocha, & Sharma). In the paper, researchers discuss the particular algorithms features and limitations. The comparison has been with various AR techniques, including: Apriori, Eclat, AIS, FP-Growth, AprioriTID, Recursive Elimination, SETM as well as AprioriHybrid concluding which one is best suitable for a certain case.

According to (Shaw, Xu, & Geva, 2008) the quality of the exposed association rules has taken less attention compared to finding much more efficient methods to determine all of possible rules. Thus, the aim of his survey relies on developing a novel method for extracting multi-level, non-redundant and cross level rules of association from sets of data with numerous concept levels and utilize the same in recommender systems.

A novel approach via AR algorithm to disclosure ordering relationships from ordinal data system is presented in (Liu, Gao, & Zhao, 2008). The main problem is how to transmute information tables to a newly created transaction database. Additional interesting ordering rules were found through this method. Mining rules of association in large databases can be done using online approach (Singh, Chaudhary, Rana, & Dubey, 2011). A weighted graph which was directed has been visualized, resulting in a novel and more enhanced algorithm for online rule generation, using depth first search.

Nevertheless, ARM produces huge quantity of rules, and it does not consider the duplicate rules (Ram Pal, Pathak, & Luma-Osmani, 2021). Potentially, the paper (Rameshkumar, Sambath, & Ravi, 2013) propose the n-cross validation procedure aiming the reduction of association rules which are not relevant to the set of transactions. The projected algorithm, PVARM used partition-based methods to support AR validation.

### **3.4 Systematic Literature Review Steps**

A variety of methods for doing a literature review have recently been presented recently. The study included in this research employs Creswell's 5 stages technique (Creswell, 2012) as represented in figure 3.1. Each phase has its own set of outcomes, which are represented by tables, figures, or charts (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).



**Figure 3.1:** Creswell's five steps to Literature Review

**Step 1:** The initial phase of the review is to define the major questions connected to causal association rules in order to get the necessary literature. (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

One of the fundamental issues is attempting to identify the most recent work on causal reasoning, accompanied by the tools, techniques, and algorithms used in this field, existing datasets, ethical issues related to their use, and finally focusing on causal reasoning in areas that have gone unnoticed, such as agriculture, computer crime, and cyberstalking.

There are therefore primary ideas and phrases retrieved in table 3.1, in order to create the search statement relating to each issue, where double quotes are employed to forcibly match by performing Boolean search (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

**Table 3.1:** Keyword Search Statement

Main Concepts	Search Statement
<b>Causal Association Rules</b>	"Causality" OR "Cause–effect Relationships" OR "Causal Rules" OR "Causal Association Rules" OR "Causal Reasoning"
<b>Causal Discovery Tools, Techniques and Algorithms</b>	("Causal Discovery" AND ("Tool" OR "Technique" OR "Algorithm" OR "Method"))

<b>Public Dataset</b>	<b>("Causal") AND ("Public" OR "Free") AND ("Data set" OR "Dataset" OR "Data-set" OR "Repository") NOT "Synthetic"</b>
<b>Research Ethics in Data Science</b>	<b>"Ethics" AND ("IT" OR "Data Science" OR "Causality" OR "Association Rules" OR "Research" OR "Experiment")</b>

**Step 2:** The next stage is the search for relevant literature after the key terms have been established. As per (Creswell) this progression incorporates area of writing about a subject by counseling a few sorts of resources and information databases, taking into account those accessible on the Internet, as well as academic libraries. On that basis, the searching method entails articles from journals, conferences, books or book chapters, reports and articles published in electronic sources (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020). Major computer sciences digital libraries were utilized, such as:

1. IEEE Xplore<sup>1</sup>
2. Google Scholar<sup>2</sup>
3. ResearchGate<sup>3</sup>
4. ACM Digital Library<sup>4</sup>
5. Springer Open<sup>5</sup>
6. DBLP<sup>6</sup>
7. Elsevier<sup>7</sup>

---

<sup>1</sup> <https://ieeexplore.ieee.org/Xplore/home.jsp>

<sup>2</sup> <https://scholar.google.com/>

<sup>3</sup> <https://www.researchgate.net/>

<sup>4</sup> <https://dl.acm.org>

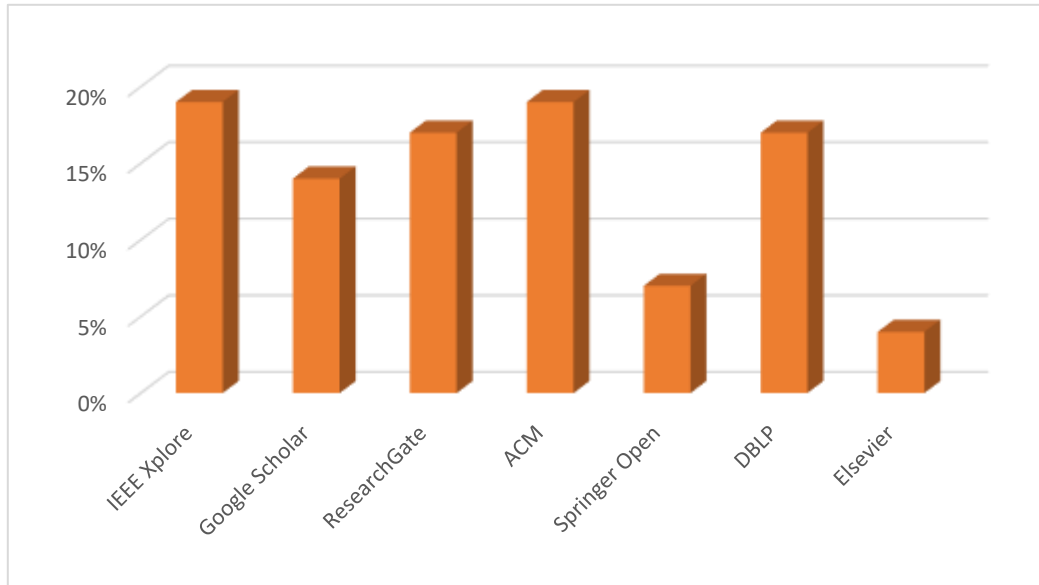
<sup>5</sup> <https://www.springeropen.com/>

<sup>6</sup> <http://dblp.uni-trier.de>

<sup>7</sup> <https://www.elsevier.com/>



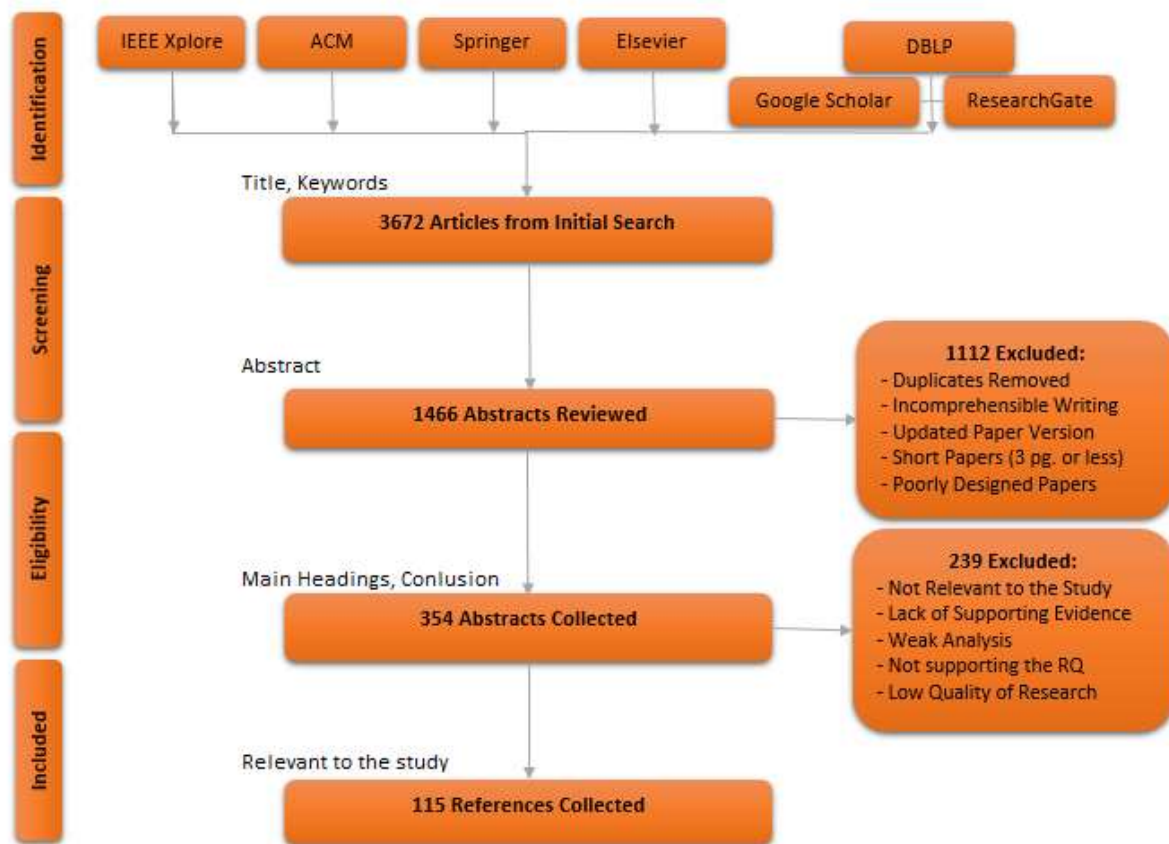
The percentage of databases utilized is depicted in the graph below (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020). Subsequently, as can be noticed in figure 3.2 the majority of the manuscripts are housed in ACM and IEEE electronic libraries (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).



**Figure 3.2:** Digital Libraries Used

**Step 3:** Critically analyze and choose the literature for the review is one of the most essential steps in the literature review when determining if it is relevant to the study being undertaken. An advanced search of the digital libraries listed above, using specified search statements, yielded 3672 publications. Duplicated articles, brief papers, and updated version papers were eliminated based on their titles and abstracts. Furthermore, based on introduction, main headings and conclusion, papers that did not support the core objective of the research were excluded as well. Only 115 of these publications were selected for further examination. The entire process of the selection and evaluation of manuscripts is introduced in figure 3.3 (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020), (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

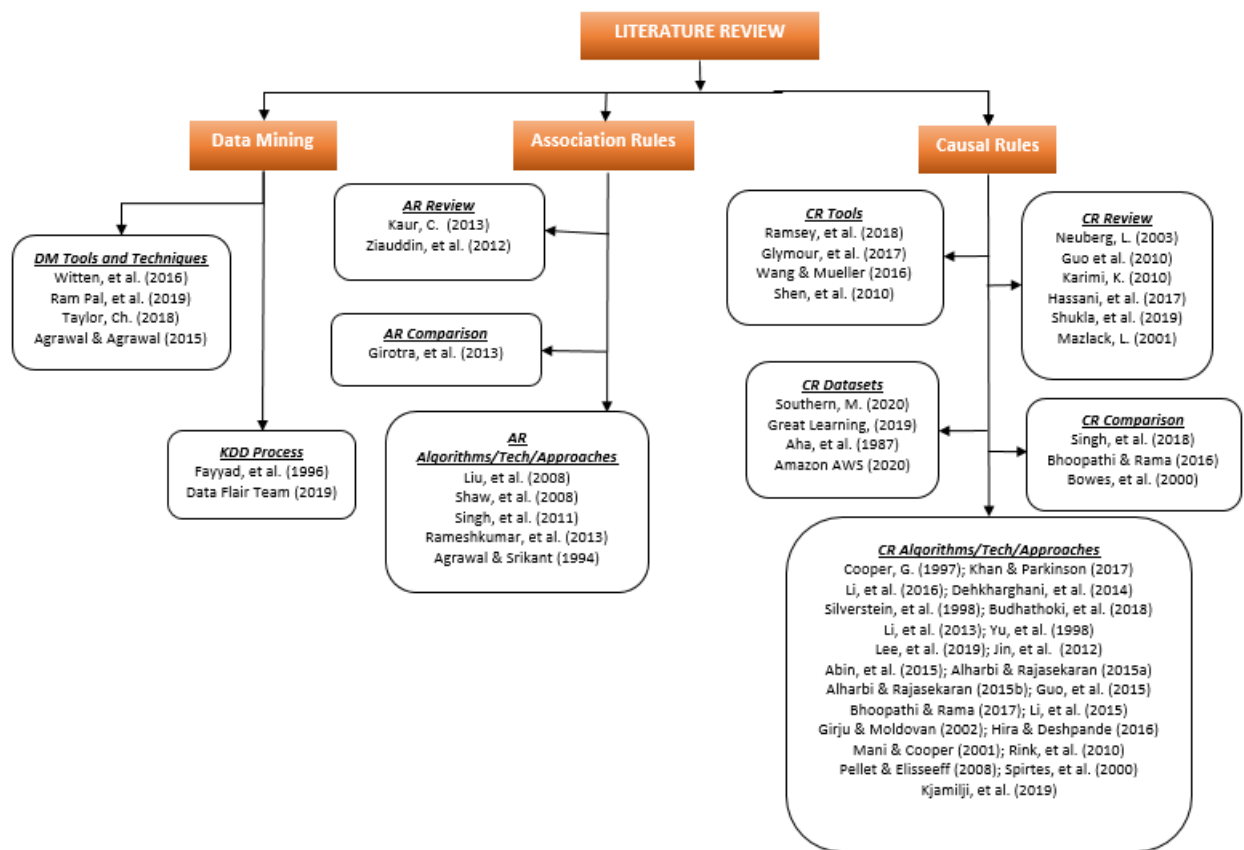
Frankly, the underlying inquiry came about a bigger number of papers, in some cases they were not identified with the primary goal, and thusly a manual filter based on the article titles or year was performed.



**Figure 3.3:** Literature Evaluation and Selection Process

**Step 4:** Organizing and coordinating the literature is the phase of storing and ordering the applicable distributions to perform further assessment. It's liked to be in a table arrangement, so various kinds of sort rules can be applied. For every distribution, reference type, article type, area, title, primary commitment, future work and so forth can be found online on <https://www.seeu.edu.mk/en/~f.ismaili> (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

The Literature Map of the most relevant articles to the research topic is shown in figure 3.4. The writings are assembled in those that expand on data mining methods overall to continue with association rules and causal rules (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).



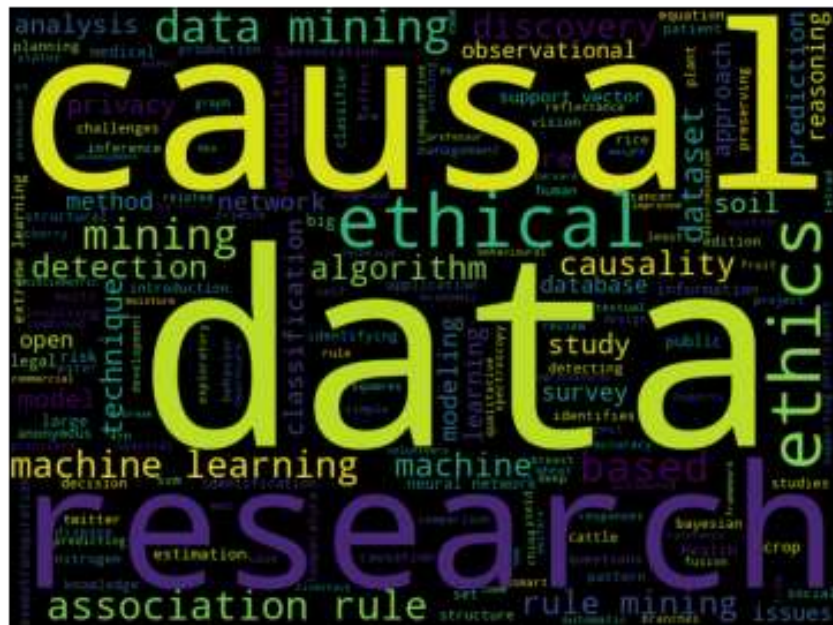
**Figure 3.4:** The Literature Map

**Step 5:** The final stage is to write a literature review that summarizes the reports and conclusions drawn from the literature study. The relevant articles are evaluated and classified using the above-mentioned categorization system (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

### 3.5 WordCloud Analysis

In order to check the most used terms during the literature review process, we utilized the area of NLP - Natural Language Processing, more accurately a WordCloud Analysis has been performed. A literature review dataset has been created and therefore, three analyses were made based on the paper titles, problem definition part and conclusion.

Firstly, there were analyzed the titles of the papers and a plot was consequently generated. As noticed, predominate the words “data”, “casual”, “research” along with the terms “data mining”, “association rule” and “ethics”, which also comprise a certain chapter within this dissertation. It means that mostly used title words on articles have to do with causal data research, which also represents our focal point.



### Figure 3.5: Title WordCloud Analysis

The upcoming analysis was completed based on the part of the paper where the authors define the research problem. As far as we can see, main problems related to Causality papers deal with “algorithms”, “proposed”, “method” or “technique”. Obviously, it is the part where the researchers note their own solution toward a certain problem on the manuscript, therefore they present a new approach for solving the problem.



**Figure 3.6:** Problem Definition WordCloud Analysis

The last breakdown directs us on the conclusions where text mining has been executed. “Causality”, “data”, “algorithm”, “rule” and “mining” are between the most prevailing words. Nevertheless, they are accompanied by the notion’s “future”, “discovery”, “machine” and “dataset”, as predictable on most concluding paper statements and presented on figure 3.7.



recognized as valuable and usable in our further research, thus helping to focus only on the focal points of our study.

The whole process of the systematic literature review resulted graphically mapped visualization. As per results obtained the main attention of our study was on elaborating the causal rules discovery algorithms, approaches and techniques.

To conclude, WordCloud analysis as an NLP technique has been realized. Main focus of the analysis was pointed to the titles of the literature review papers, as well as their problem definition and closing remarks. As expected, major articles were focused on the Causality and the algorithms used while discovering such relations among the data.



## RESEARCH METHODOLOGY

---

The research methods that will be used in the dissertation are noted as follows:

- **Fundamental Research:** will allow for the assessment of data results by examining existing methods and algorithms as well as the suggested arrangement.
- **Empirical Research:** statistical analyses of the data, revealing favorable possibilities and advantages.
- **Action Research:** the study approach of taking action while doing research and iteratively evaluating the practical solutions for specific areas.
- **Observational Research:** under certain conditions, observational studies are seen to be the best replacements of experiments, and the results show that well-made observational studies may provide comparable findings to RCT's (Singh, Gupta, Tewari, & Shroff, 2018), (Li, et al., 2015). The discovery of causal relationships based on observational data is an substitute solution when the experiments, due to several ethical and cost matters, are infeasible (Alharbi & Rajasekaran, 2015). Using observational data was the most suitable solution for major researchers working in the area such as: (Guo, Xing, & Lee, 2015), (Mani & Cooper, 2001), (Li, et al., 2013), (Hira & Deshpande, 2016), (Ram Pal, Pathak, Yadav, & Ora, 2019), (Bowes, Neufeld, Greer, & Cooke, 2000), (Li, Ma, Le, Liu, & Liu, 2016) (Jin, et al., 2012), (Bhoopathi & Rama, 2017), (Kjamilji, Idrizi, Luma-Osmani, & Zenuni-Kjamilji, 2020), (Budhathoki, Boley, & Vreeken, 2018), (Rameshkumar, Sambath, & Ravi, 2013),



(Guo, Cheng, Li, Hahn, & Liu, 2010), (Dehkharghani, Mercan, Javeed, & Saygn, 2014), (Singh, Gupta, Tewari, & Shroff, 2018), (Girju & Moldovan, 2002).

## 4.1 Objectives of the Study

The dissertation comprises of several objectives as listed below:

- The analysis of the statistical causal association between two or more events (variables).
- Evolve hybrid techniques in order to detect causal structures and to enable diverse of algorithm combinations including supervised or unsupervised learning.
- Develop a model that employs the utility of the proposed research idea among a set of measured cause and effect factors.
- Explore the areas where there is potential lack for mining databases in certain domains to discover causal relationships.

## 4.2 Research Questions

The corresponding work is carried out on four research questions (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020):

**RQ1: What are the most substantial issues that have to be taken into consideration when we identify causal relations and why it seen as an essential problematic in science?**

In answering this question, the most essential thing is to know the essence of the phenomena of causality. To answer it we should go through the fundamental or the basics of examination, and significate how important causal disclosures in science and innovation are with regards to forestalling destructive incidental effects.

**RQ2: Which are the causal mining algorithms, techniques and methods that focus in the problem of discovering causality?**

The response to this inquiry gives us a rundown of the literary review. We need to discuss the methodologies we have found to define the causal guidelines from datasets, which yield the best outcomes.

**RQ3: How are publicly available data utilized by the researchers and what are the possibly reimbursements from it?**

This methodology takes pleasure in the large number of open data which is freely accessible on the internet, additionally, successor use is less exorbitant, less tedious in the review part, and there's no danger for the participant's behalf.

**RQ4: What is the main challenge of conducting experiments because of ethical concerns?**

Fundamentally recollecting that the human existence and being moral towards it is a higher priority than the aftereffects of the examination results (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

The systematic literature review discoveries are presented in the form of responses to the research questions raised.

**RQ1: What are the most substantial issues that have to be taken into consideration when we identify causal relations and why it seen as an essential problematic in science?**

The marvel of causal phenomenon has been intensely bantered throughout the long term, yet at the same time is listed as a central subject in various scientific studies. Bayesian networks present a pioneering discovery in this area, since its application served as a base approach to a lot of studies that came later. They are arisen as a significant strategy for revelation of causal structures (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

Recuperating the causal relations from information which have been noticed is the focus on any information expert and analyst, since it is a primary issue in science. Following quite a while of attempt, causality research was featured when the researcher Judea Pearl was declared as a winner of the Turing grant in 2011 for research related to causal induction (Wang & Mueller, 2016), (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

Causality is perceived as a large assembly of the cause with the effect that fallouts. It is difficult to define since this is frequently due to our intuitive understanding of causes and consequences,

such as: the student failed on the exam since he was sluggish or the trees fall down because the meteorological conditions were windy (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

When analyzing causalities using data, we must grasp the distinctions between statistical correlations and triggers (Guo, Cheng, Li, Hahn, & Liu, 2010), (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020). Another major issue is if an interaction is causal. By and by, association rules just evaluate the general conjunction of qualities. It would be significant if a collaboration were to be viewed as causal (Mazlack, 2001). It would be vital.

Beside acting like an extraordinary philosophical discussion, causality is contemplated and utilized in different disciplines, such as: social networks (Dehkharghani, Mercan, Javeed, & Saygn, 2014), medicine (Abin, Mahajan, Bhoj, Bagde, & Rajeswari, 2015), (Bowes, Neufeld, Greer, & Cooke, 2000), text processing (Rink, Bejan, & Harabagiu, 2010), computer security (Khan & Parkinson, 2017), teaching purposes (Bhoopathi & Rama, 2017), cancer prediction (Shukla, Yadav, Ram Pal, & Pathak, 2019), (Kjamilji, Idrizi, Luma-Osmani, & Zenuni-Kjamilji, 2020), big data (Hassani, Huang, & Ghods, 2017), (Guo, Cheng, Li, Hahn, & Liu, 2010), construction (Lee, Cha, Han, & Hyun, 2019), (Girju & Moldovan, 2002), (Guo, Xing, & Lee, 2015), economics (Hira & Deshpande, 2016), decision making and predicting upcoming events. It provides an excellent chance for doing research, particularly in the field of KDD (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) as well as mining of data (Witten, Frank, Hall, & Pal, 2016).

Machine learning techniques are widely utilized in healthcare, particularly for cancer diagnosis and prediction. An obvious investigation was carried out in order to use several machine learning approaches to identify and forecast breast cancer by (Shukla, Yadav, Ram Pal, & Pathak, 2019). Those techniques comprise of Decision Trees, Naive Bayes Classifier, Support Vector Machines, Logistic Regression, K-Nearest Neighbor as well as Artificial Neural Networks. In any event, the application of causal induction in medical care would fall short of the standard AI approaches. An effective technique in detecting the Adverse Drug Reactions - ADR, has been explored by (Abin, Mahajan, Bhoj, Bagde, & Rajeswari, 2015). The study was carried out using patient data gathered from symptom data and medication. The combinations of drug indications are generated, the support is computed, and finally the causal structure is analyzed and linked with the causality

categories that are used to detect such kinds of ADRs. A limited secure and flexible building blocks constructed on previously trained models of ML, that can support the construction of different classification arrangements focused on privacy-preserving are announced by (Kjamilji, Idrizi, Luma-Osmani, & Zenuni-Kjamilji, 2020), (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

Scientists (Bowes, Neufeld, Greer, & Cooke, 2000) directed an experiment with contraceptive method choice data, by a comparison approach of the efficiency of causal algorithms along with association rules induction for determining discrete data patterns. The execution has been done in Tetrad II software, with DBMiner system focused on causal extrapolation algorithms with Bayes network, and the contraceptive method choice (CMC) dataset. In case information is causally adequate, the PC calculation is utilized, despite what might be expected, the Fast-Causal Inference (FCI) calculation is utilized, in this way the eventual outcomes drove many intriguing rules.

The reputation of the causality analysis in big data is offered in (Hassani, Huang, & Ghods, 2017), where the most current significant uses of data mining methods in causality analysis are given. Additionally, (Guo, Cheng, Li, Hahn, & Liu, 2010) discovered a gap amongst learning causality and big data, by taking into consideration the frontier and traditional methods all together. The same was supported by a discussion of several open learning causality concerns. Following that, learning causality approaches are classified and examined for common issues, as well as various forms of data connections amid machine learning and causal knowledge are noted (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

Discovery of causal effects might be also derived from text in order to provide causal relationships among the text data. A new concept of "sentimental causal rules" was introduced that incorporates causal rules between dissimilar characteristics extracted from written data and sensations relating to these aspects, coupled with the approaches to extract and demonstrate the effectiveness of sentimental rules of causal type, as data resource on Twitter, presented by (Dehkharghani, Mercan, Javeed, & Saygn, 2014), (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020). To determine the causal connection amid events encoded in text, and therefore constructing a graph representation of the sentence, then repeatedly extract multiple graph patterns (or sub

graphs) and rank them according to their significance, is provided in (Rink, Bejan, & Harabagiu, 2010). The process that semi-automatically finds lexico-syntactic examples in English messages identifying with the causal relationship is introduced by (Girju & Moldovan, 2002). On the other hand, (Guo, Xing, & Lee, 2015) present a way for detecting student mechanical explanations of scientific events using a semi-automated methodology that combines ARM and human intuition (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020). These experiences may be utilized to improve teaching and learning as well as to help develop future curricula.

Finding causal linkages between security event log entries without the use of human help was investigated in (Khan & Parkinson, 2017). Non-professionals might use the retrieved knowledge to develop measures to improve system security. A data mining way to deal with all the more advantageously accumulate imperfect causal data by distinguishing then measuring interrelated causal rules as related to the construction defects from accessible data sets was led by (Lee, Cha, Han, & Hyun, 2019), (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

Despite the fact that causal reasoning has been used in numerous fields, it is still underutilized when compared to conventional machine learning techniques. Furthermore, a variety of issues surrounding causality must be addressed (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

**RQ2: Which are the causal mining algorithms, techniques and methods that focus in the problem of discovering causality?**

The Pearl's book "Causality: Models, Reasoning, and Inference", which is otherwise denoted as "The Book of Why", seeks to combine causal and effect structures on the basis of psychology cognitive science, epidemiology, statistics and econometrics. The Causal Bayesian Net is amongst the most established and widely accepted theories for determining causality (CBN). DAG, Directed Acyclic Graph as a tool, was used to describe causal interactions in this paradigm. Pearl suggested an approach for discovering causal structures from linked conditional independence, on which various tools for identifying cause-and-effect linkages have been devised (Neuberg, 2003), (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

The habitual Local Causal Discovery - LCD algorithm formerly established by (Cooper G. F., 1997) indicates the way how observational data can compel the causal connection among the deliberate factors, now and then to where it tends to be surmised that one variable causes the other one (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020). The Local Causal Discovery method is likewise built upon seven premises, with the restrictions correspondingly outlined.

Acquiring rules by iteratively creating candidate rules and by including their events in the database remained licensed by Yu et al. (New York, USA Patent No. US005832482A, 1998), (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020). The next set of candidate rules for evaluation is utilized for the creation of newly discovered causality rules. To generate the causality rules for the consequence, the ideal embodiment employs an iterative technique. Set sizes, as well as the triggering Set sizes. LCDa, LCDb, and LCDc are three variants of the LCD algorithms for efficiently discovering potential causal associations from large observational databases using the public Alarm dataset (Mani & Cooper, 2001), (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020). Researchers (Silverstein, Brin, Motwani, & Ullman, 1998) investigated a number of algorithms built by the idea and concepts of LCD algorithm. LCD negates the notion that certain factors doesn't have triggers, as an alternative, the three substitute causal relationships are stated without selecting among them. CCC and CCU rules were indeed utilized. The outcomes prove that the technique provided here is computationally viable as well as efficient at detecting significant causal structures. By "manipulating" observable data, causal laws can be discovered (Institute for Work&Health, 2016) rather than altering populations, as the CAR algorithm may lead to enhancements (Li, et al., 2013), (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020). Rendering to the findings of the study, CAR establishes causal relations for mutual variables, therefore the number of generated rules is greater compared to algorithms CCU and CCC as represented below in table 4.1:

**Table 4.1:** CAR, CCC and CCU Causal Rules

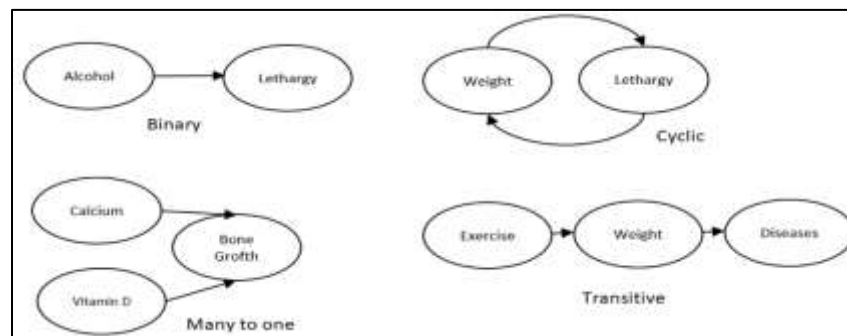
	<b>CAR</b>	<b>CCC</b>	<b>CCU</b>
<b>Adult</b>	49	53	46
<b>Sick</b>	18	13	3

The approach of CDT - Causal Decision Tree, where vertices have causal understandings was established by researchers (Li, Ma, Le, Liu, & Liu, 2016). The given approach applies a well-recognized platform of causal deference and uses a classical statistical test for partial associations, the Mantel-Haenszel Test, and it is determined that the predictor variable has a causal link with the result if the causal effect is substantial. Contrasting with standard choice trees, the CDT offers a succinct graphical causal portrayal. In causal diagram, the most significant factors for certain hub are its parents, children, and children's parents (or spouses), otherwise called the Markov blanket (Pellet & Elisseeff, 2008). The identification of the spouses results in the recognition and therefore in causal direction of V-structures (two distinct causes leading to the same outcome). Within the same architecture, an explicitly backwards feature selection heuristic Total Conditioning TC is developed and a version TCbw is developed for Gaussian data (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

The FCI, which makes no presupposition such as this, is the main algorithm for limit-based learning, compared to algorithm PC, which makes an assumption that no latent variables will be presented (Spirtes, Glymour, & Scheines, 2000). Both of the algorithms present a substitute to Bayesian techniques. PC starts with a full unguided diagram, assuming that each node is interlinked. After the conditional independence among the nodes is calculated, the nodes will then be implemented. Whenever a test shows that these variables are conditionally independent from the additional variable, the resultant charts would mean that a node along with all its descendants positioned on the left side, and all their descendants are independent of any other node. Whereas FCI can find causal regulations from conditional independence testing as a constraint-based approach and is feasible in big variable sets even if hidden variables are implicated (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020). However, as a detriment, an exponential run-time complexity might be addressed.

The researchers of (Budhathoki, Boley, & Vreeken, 2018) investigate the issue of determining causal rules based on previously observed data. A technique termed CR-CS (Li, et al., 2015) has been created to identify such causal relationships from observational data. It is done by integrating the mining of association rules, along with retrospective cohort studies for mining CRs in big datasets, which are accomplished to find a trigger that consists of several factors or

variables. A method for mining causal rules in large binary variable datasets was suggested by (Jin, et al., 2012). Their solution broadens the scope of causality discovery to the causal connection with multiple causative factors, allowing for the formulation of both single and combination causal rules (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020). Aiming to determine conjunctive correspondingly disjunctive causal rapport in observational data, scientists suggested two algorithms, CCCRUD (Alharbi & Rajasekaran, 2015), and DCCRUD (Alharbi & Rajasekaran, 2015), respectively. Their main focus was set on single and multiple factors by employing an association and partial association approach (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020), (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022). Fixated to discover the present causal structure effectively in large time series datasets, (Hira & Deshpande, 2016) as shown in Figure 4.1, used a strategy that broadens the possibility of discovering causal rules by grouping diverse causal rules into: binary, cyclic, multiple to one and transitive. Temporal odds ratio and temporal association were utilized by the researchers to rule out no causal interaction and to verify the high reliability of the found causal rules (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).



**Figure 4.1:** Types of Causal Relationships

To discourse the problem of techniques for rule discovery, (Karimi, 2010) outlined a number of tactics to articulating causality and temporality, seen from philosophical, technical and physical perspectives. In addition, the study briefly presents techniques utilized in discovering causal structures, that employ sequential observations, Bayesian nets, and the Minimum Message Length Model. Causality presents an especially difficult phenomenon, since datasets tend to be huge (Bhoopathi & Rama, 2016), (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).



The article compares several algorithms and their performance, such as CC-Path, LCD, CU-Path, according on three criteria's: number of database passes required, memory use and the execution time. Same authors recommended one year later a cause-finding algorithm dubbed CRM. Such causal rules are mined based on the Pre-Order Coding - POC tree (Bhoopathi & Rama, 2017), (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020). Therefore, causal discovery algorithms are divided into two categories: (1) those that assume acyclic and no latent factors, like: PC, MMHC, GES, MCMC, GIES, and CPC, and (2) those that accept both latent factors and cycles well, like: CCD, FCI, RFCI and FCI+ approaches. outcomes of experiments may be compared from three points of view: structural accuracy, standard prediction and counterfeit accuracy (Singh, Gupta, Tewari, & Shroff, 2018), (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

Mazlack in his manuscript (Considering Causality in Data Mining, 2001) presented a comprehensive summary of causal structures, using applied instances and categorizing causality into Conjunction, Network and Chaining. A distinction has also been drawn between causality and statistical dependency, emphasizing that two occurrences may be reliant on each other, but this does not imply that one event causes the other (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

The experts in (Rule Discovery for Exploratory Causal Reasoning, 2018) examine the challenge of generating trustworthy causal rules from observational data (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020). For this purpose, a new descriptive rule discoveries approach is being introduced based on the reliable estimation of the conditions in the face of potential confusion, and by implementing the branch and bound search algorithm, a powerful optimized algorithm is being derived which successfully detects valuable rules in many real datasets.

### **How are publicly available data utilized by the researchers and what are the possibly reimbursements from it?**

Data are obviously omnipresent, and there are numerous ways to leverage it to generate new and engaging content. While the data generation process is vital for science, it is becoming a contentious subject as an alternative discovery and utilization of public data sets (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020). Fortunately, there are various publicly accessible data waiting to

be acquired and analyzed but the important part is to locate relevant usage-specific data sets, by searching keywords. At instance, the use of an existing collection of data may enable a researcher get findings much faster, at lesser costs and without putting many possible disadvantages connected with involvement in research into new research respondents and demonstrate excellent practice more rapidly (Doolan, Winters, & Nouredini, 2017), (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

Google's Datasets, Kaggle, UCI, Amazon and .gov Datasets are among the top 5 data repositories based on the 2019's survey (Great Learning). They consist of numerous datasets, spanning government websites, satellite pictures, ecological resources, and so on. Diverse observational examination of AI calculations was finished utilizing them by AI people group, with each challenge having a special dataset (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020). The text roar we have summed up the articles that help public informational collections and use it on their investigation for getting critical conclusions.

The next text highlights manuscripts supporting public datasets where important findings are drawn. As the eldest dataset on the Internet, UCI ML Repository, has been often performed in application approaches and been in focus of data miners (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

According to our review, the mostly utilized datasets include: Adult, Breast Cancer, Mushroom, Chess, Wine, Tic-tac-toe, Glass, Nursery, Zoo, Heart, Lymphography, Wisconsin, and Automobile (Budhathoki, Boley, & Vreeken, 2018), (Kjamilji, Idrizi, Luma-Osmani, & Zenuni-Kjamilji, 2020), (Ram Pal, Pathak, Yadav, & Ora, 2019), (Rameshkumar, Sambath, & Ravi, 2013), (Li, et al., 2015), (Bhoopathi & Rama, 2017), (Li, et al., Mining Causal Association Rules, 2013), (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

(Li, et al., 2013) used four UCI and Harvard Medical School datasets. Sick, along with Hypothyroid specify two kinds of medical data, part of Thyroid Disease folder, in the UCI repository. Moreover, Adult data set presents an extraction from the United States Census dataset in 1994, and they also utilized a big set of data such as Census Income, to evaluate the scalability of their technique

according to data of several sizes. And finally, the Lung Cancer public data set retrieved from Dana-Farber Cancer Institute<sup>8</sup>.

Other authors (Ram Pal, Pathak, Yadav, & Ora, 2019) have examined the impact of three trimming strategies on the amount of rules generated. The study consisted of 10 public datasets from UCI repository: Breast, Labor, Heart, Prima, Glass, Tic-tac-toe, Zoo, Lymph, Iris and Wine.

(Budhathoki, Boley, & Vreeken, 2018) used 22 available data sets in their analysis. Among them, the most used ones are: Adult, Automobile, Zoo, Nursery, Tic-tac-toe, Lymphography, Breast Cancer, Chess, Mushroom etc. Likewise, the Titanic data set from Kaggle<sup>9</sup> repository was considered. According to the survey, one of the causes for such sad loss of life was a paucity of lifeboats, as well as the fact that during the evacuation, certain passengers were held to a different standard than the others.

Publicly available datasets such as: Jobs<sup>10</sup>, PROMO<sup>11</sup>, Weblogs<sup>12</sup>, CPS Extract<sup>13</sup>, SIDO<sup>14</sup>, Twins<sup>15</sup> and Absciscic Acid Signaling Network<sup>16</sup> were used by (Guo, Cheng, Li, Hahn, & Liu, 2010) aiming the study of different types methods and data, that are used to explore causal effects and to recognize causal rules between the attributes on the aforementioned datasets (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

Two years later, (Li, et al., 2015) proved their proposed algorithm CR-CS in open set of data, including German, Mushroom, Chess, also known as King-Rook vs. King-Pawn and Tic-tac-toe (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020). These real-world datasets demonstrated viability and efficacy of mining causality.

---

<sup>8</sup> <http://leo.ugr.es/elvira/DBCRepository/LungCancer/LungCancer-Harvard1.html>

<sup>9</sup> <https://www.kaggle.com/c/titanic/data>

<sup>10</sup> <http://users.nber.org/rdehejia/data/nswdata2.html>

<sup>11</sup> <http://clopin.net.com/causality/data/promo/>

<sup>12</sup> <http://www.causality.inf.ethz.ch/repository.php?id=13>

<sup>13</sup> <https://economics.mit.edu/faculty/angrist/data1/data/angkru95>

<sup>14</sup> <http://www.causality.inf.ethz.ch/data/SIDO.html>

<sup>15</sup> <https://github.com/AMLab-Amsterdam/CEVAE/tree/master/datasets/IHDP>

<sup>16</sup> <https://archive.ics.uci.edu/ml/datasets/Abciscic+Acid+Signaling+Network>

The equivalent dataset (Adult) in blend with Ultra Short Stay Unit - USSU information from the emergency department of a common clinical center in Australia were utilized by (Li, Ma, Le, Liu, & Liu, 2016) to show that causal decision trees can discover more interpretable connections when compared with ordinary decision trees (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

Efficient algorithms are evaluated and refined to discover causal principles from observational data in (Mani & Cooper, 2001). The data have been simulated in a medical domain using a recognized causal network – the Bayesian alarm causal network. It consists of 37 vertices and 46 causes. Alarm was designed to model probable interactions with the patient's anesthesia in the operating room by Beinlich while providing anaesthesia to the patient (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020). The Alarm network was developed using his skills as an anesthesiologist and medical academic articles. Alarm was broadly utilized in Bayesian network induction testing and is seen to be a helpful standard benchmark. In this pattern, the estimated number pairs are 666.

Far ahead, it was extended for different operations like the Alarm dataset, which was originally designed to help understand monitoring data to alert anesthesiologists to different operating room situations, Insurance has been developed to determine car insurance risks, Hailfinder a standard program forecasting Northeastern Colorado severe summer hail, Carpo is meant to support the diagnosing of carpal tunnel syndrome, as well as dataset Diabetes containing insulin users data for the dose alteration model (Pellet & Elisseeff, 2008), (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

The effort of (Girju & Moldovan, 2002) examines causal development depictions in openly accessible English compositions, involving the transmission of several morphology, linguistic structure, and semantic - based phonetic classifications (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

Causal revelation from public information texts' assembled from interpersonal organizations was directed by (Dehkharghani, Mercan, Javeed, & Saygn, 2014) where causal standards were removed ward on the catchphrases recently procured from Twitter. Hence, different 5000 tweets were collected and analyzed using data mining and opinion examination methods. In the first place, from the tweets which were utilized in causal standard mining, the watchwords were removed. Researchers next analyzed tweets with feelings that classify them as good, bad or neutral (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

For the upcoming analysis was employed the 1990 Census data gathered in Washington, comprising of two public data collections, Reuters Newswires Written Information and UPI (Silverstein, Brin, Motwani, & Ullman, 1998). The research included 3056 news stories that were extracted from the news chain termed clari.world<sup>17</sup>, .txt file of 18 GB (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

University of South Australia's free dataset acquired from the Data Analytics Group was used to mine disjunctive merged causal rules (Alharbi & Rajasekaran, 2015). This then incorporates 8 quality criteria and one variable as objective, divided into 100,000 records. Socioeconomic process evaluation is done out, as well as the establishment of a detailed cause-effect link between various economic indicators (Hira & Deshpande, 2016). Data from the International Monetary Fund - IMF <sup>18</sup> , World Bank data<sup>19</sup> and World Trade Organization - WTO <sup>20</sup> , are used to generate real-world economic statistics. The World Commerce Organization (WTO) collects information on global trade in goods and marketable services. Data from 189 countries related to the economic indicators are available from the IMF. Time series data from 250 nations are included in the World Bank regarding issues like health, employment, climate and agriculture, among others (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

---

<sup>17</sup> <http://www.ibiblio.org/usenet-i/hier-s/clari.world.html>

<sup>18</sup> <http://www.imf.org>

<sup>19</sup> <http://www.worldbank.org>

<sup>20</sup> <http://www.wto.org>

DREAM4 dataset, involving data from the regulatory grid of in-silico genetic factors, was utilized for the algorithm's optimization for learning structures. Dataset, made up of in silicon regulatory network data used to optimize learning structure algorithms. Five networks for each of 10 and 100 nodes with differing topologies that incorporate feedback mechanism are built from DREAM4. Sachs data set comprises of the empirical data gathered following general disturbance that was based on the simultaneous assessment of single-cell expression patterns of eleven proteins of the pyrophosphate implicated in initial T cell signaling. Sachs' compendium has 1756 different types of insights. Both are used for causal computations and correlation in. Both are used in (Singh, Gupta, Tewari, & Shroff, 2018) for causal algorithms comparison (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

(Wang & Mueller, 2016) present the Visual Causality Analyst, a novel visual blueprint that makes possible to edit and review causal links. It collaborates with the causal detection algorithms to define a relevant causal net. This is accomplished through the use of two publicly available datasets: the Auto-MPG dataset and Heart, a genuine dataset on Cardiopathy Diagnosis. It involves 270 records and 7 category factors or variables, as following: gender, sickness, angina, quickBloodSugar, chestPainType, restECD and thalassemia (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).

The CMCs (Contraceptive Method Choice) in the UCI Machine Learning Report, including 1427 full entries, consisting of eight subtle and two consecutive parameters, were utilized in their poll in Indonesia (Bowes, Neufeld, Greer, & Cooke, 2000). The end results showed several fascinating correlations which indicated that a real cause of media exposure, contraceptive use, and education for her spouse is the woman's level of education. The first two results are supposed to be less obvious, namely that the education level of a woman is the cause of the education level of her partner.

The suggested, (Jin, et al., 2012) method uses the Arrhythmia dataset (UCI) that makes a distinction and categorizes it into several categories between the occurrence and lack of heart arrhythmia (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020). The data collection comprises 452

records, and 279 attributes and one attribute of one class are provided with each entry. All of the data characteristics were converted into binary values, 1 specifying yes and 0 no.

(Bhoopathi & Rama, 2017) demonstrated that the causal relationships mined from huge databases, can provide the necessary corporate intelligence to make expert judgments. Regarding investigations, adult datasets are utilized frequently for the purposes of data mining. Also referred to as the data set for Census revenue. It is a heterogeneous dataset comprising categorical and continuous data. This has 48842 occurrences and 14 attributes. The results showed that education, work and the working class are causally related to employees' salaries.

Utilized by (Kjamilji, Idrizi, Luma-Osmani, & Zenuni-Kjamilji, 2020) Wisconsin Breast Cancer original data of the UCI repository has been utilized to test the performance of several reliable and powerful construction blocks to build different systems of data protection classification built on previously trained ML models (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020). Then the blocks are employed and practically used as a use scenario in order to facilitate a private life-saving classification of Naïve Bayes in the semi-honest model for use in breast cancer diagnosis.

Freeware samples of varied sizes, item sizes, and other behaviors, including mushroom, chess and heart illness, have been used for assessing the suggested PVARM algorithms on mining medical relevant (Rameshkumar, Sambath, & Ravi, 2013). These data sets investigate the influence of limitations with affirmation on the testing dataset and eliminate faulty rules.

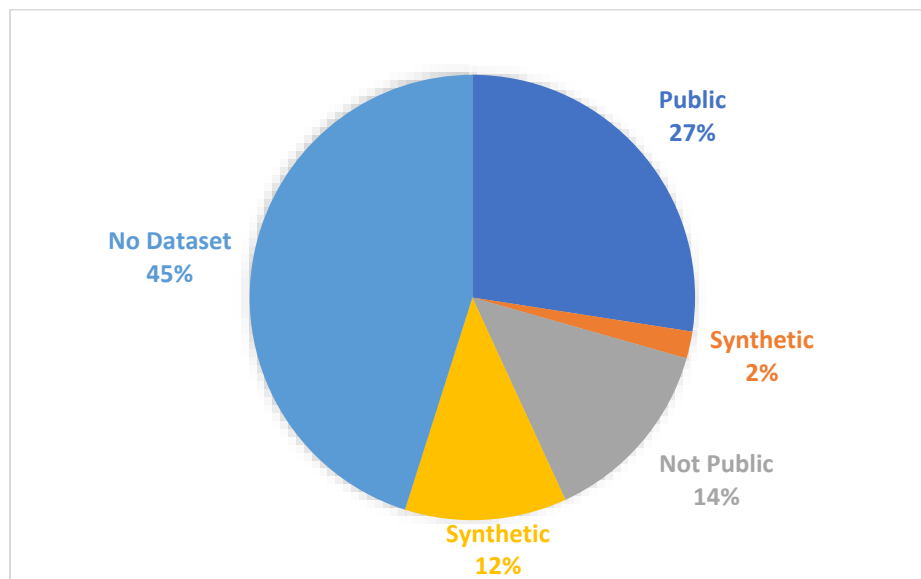
(Guo, Xing, & Lee, 2015) provide an example for the use of this technique in the written replies for students in public data that involve issues related to climate change. As part of the planning examinations for students participating in the Concord Consortium<sup>21</sup> online program for climate change, the questions were posed.

Taking into consideration the literature used so far, public datasets were utilized by the 27% of the publications, synthetic dataset was utilized by one paper only, presented as 2% (Alharbi & Rajasekaran, 2015) and not public data was used by 14% of the studies. Six publications or represented in percentage 12%, utilized the both approaches (Li, et al., 2015), (Alharbi &

---

<sup>21</sup> <http://authoring.concord.org/sequences/47>

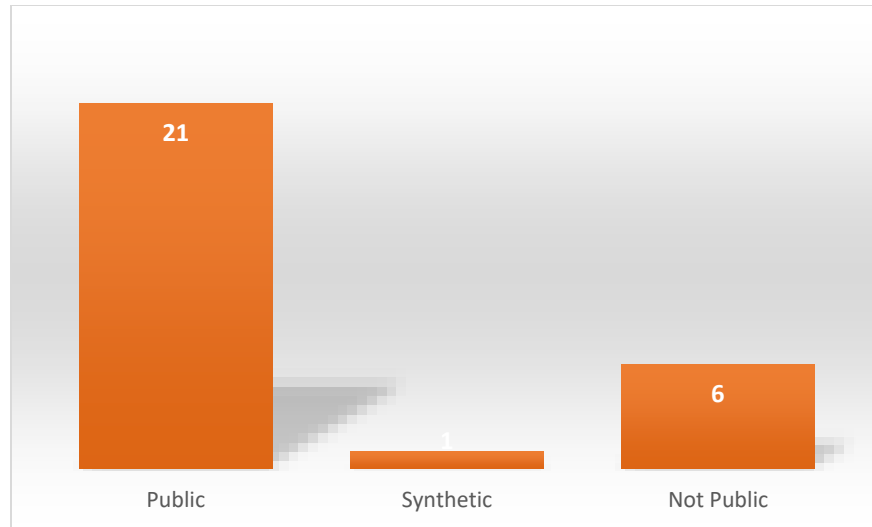
Rajasekaran, 2015), (Li, Ma, Le, Liu, & Liu, 2016) , (Hira & Deshpande, 2016), (Rameshkumar, Sambath, & Ravi, 2013) and (Singh, Gupta, Tewari, & Shroff, 2018). The generated synthetic sets of data, mostly present Tetrad applications (2017), (2018). Ultimately, in 45 percent of the articles, no database was employed. Most of these publications provide reviews or present a new method or technique but the same is tested on a particular data set. A comprehensive explanation and visualization of aforementioned issue is presented the following chart (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020).



**Figure 4.2:** Datasets used in Literature

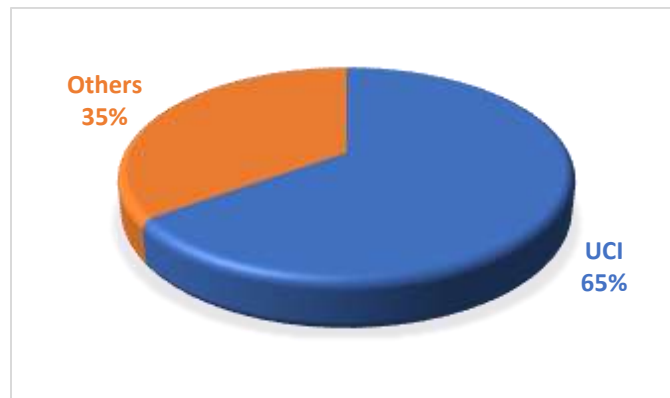
As regards to the papers that applied databases in the surveys, the type of datasets, is presented on the figure 4.3.





**Figure 4.3:** Types of Datasets

Whereas most of the papers when using data repositories, they have chosen the UCI ML data warehouse (UCI Machine Learning Repository, 1987), as presented below.



**Figure 4.4:** Data Repositories

**RQ4: What is the main challenge of conducting experiments because of ethical concerns?**

In the process of data mining, the data source is a crucial part. Ethical issues are continually confronted with the evolution and further development of data mining. Until recently, protection of privacy and ethical warnings were largely uninterested in mainstream KDD research. In the KDD (Shapiro, et al., 2001), (Fule & Roddick, 2004), (Reynolds G. W., 2015), (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

(Fule & Roddick, 2004) claim that the approach of creating unique mining rules is an ethical problem, particularly when the results are utilized in decision-making processes that affect individuals, or when mining customer data in an innocent manner jeopardizes those customers' privacy. Researchers provide a procedure for assessing a regulation in terms of perceived privacy and ethical sensitivity in the poll (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

In addition, there are numerous ethical problems in the realm of web mining. It is connected to data mining and similar approaches in order to automatically identify and extract information and uncover relevant patterns from web documents and services. It is vital to recognize that certain fundamental ethical principles, such as privacy and autonomy, are under threat (van Wel & Royakkers, 2004). Web mining makes it impossible for a person to independently supervise the availability of data pertaining to the private life. To investigate the mentioned risks, the article divides them into main kinds: "web structure and content mining" and "usage mining" (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020). When mining data published online, and combining it with supplementary data, even if it is used a completely new perspective, web content and structure mining are a source of worry. When tracking and analyzing online users' behavior without their permission, web usage mining poses privacy concerns (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

Cook recognizes in his book chapter (Ethics of Data Mining, 2009) that data miners and decision-makers are obviously expected to follow the law, but ethical concerns are frequently more rigorous than what is legally necessary. Thus, according him, it is regrettable that some IS experts either lack a clear grasp of the outcomes of data and data mining in their companies or realize that this is not their business (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

In the domains of medical and health sciences a considerable number of ethically accountable databases exist (Fule & Roddick, 2004).

In the article published in 2006 (Yadav) discusses a victim case report following a month's incident. The report is then a medical case. Article considers all the ethical and legal involved with the case accident victims. The chapter also briefly interacts with the constitutional obligation of the government to safeguard its people's "right to life," one of their individual liberties.

During 2012, a research program was performed on the ethical and legal problems of 35 nurses in Iran (Aliakbari, Hammad, Bahrami, & Aein, 2015). The manuscript deals with many ethical concerns, such as: keeping legal health data by keeping records to prevent loss, damage or illegitimate use. Exceptional moral issues related to data protection might potentially lead to calamities. Caregivers should be aware and respect the ethical standards of data protection legal concerns.

### 4.3 Hypothesis

The hypotheses of the theses are listed as follows:

**NULL:** The data relationship discovery can be broadly empowered and improved through the use of causal reasoning approaches.

1. The developed system will improve the efficiency and effectiveness of the data relationship discovery algorithms in public datasets.
2. The potential misuse of data and causal reasoning findings creates additional ethical and legal challenge that requires the definition and regulation by normative framework.

### 4.4 Unexplored Causality Areas

From the perspective mentioned so far, we see the potential research paths of causal reasoning applied specifically in areas where a more targeted decision making can be done if a more methodical causal analysis would be conducted.

Moreover, in all the review done, as displayed in figure 4.5 below, we can conclude that causal reasoning is partially or not analyzed at all, in areas like: Smart Agriculture, Data Ethics and Computer Crime.

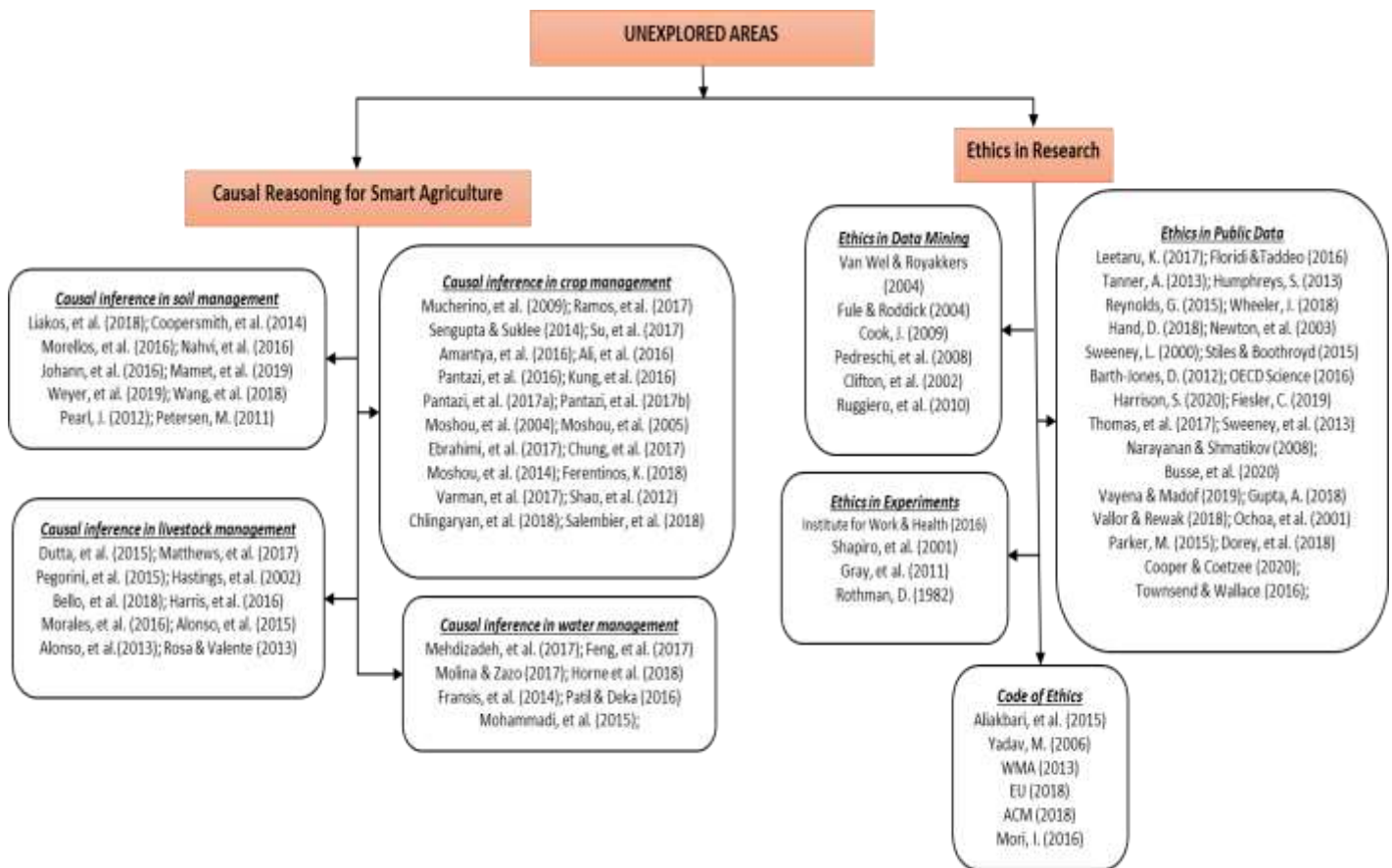


Figure 4.5: Unexplored Areas

#### 4.4.1 Smart Agriculture

The objective of this subsection is to offer a comprehensive overview of important works on causal thinking focusing on intelligent agriculture. This study examines numerous research topics that allow the causal reasoning problem in smart agriculture to be fully summarized and analyzed from several points of view: applications as well as technical. The overview of connected work in using causal thinking in intelligent agriculture covers four main components of intelligent agriculture, as categorized in (Liakos, Busato, Moshou, Pearson, & Bochtis, 2018).

The segments in question include crops, livestock, water management and soil management (Luma-Osmami, Ismaili, Raufi, & Zenuni, 2020).

## Causal Reasoning in crop management

Crop management in smart farming is one of the most essential jobs. This is an essential niche in the production of agricultural food for the supply of high quality, demand and sickness-free crops. Crop managers must consider very key aspects such as yield estimates which address the process of forecasting the right crop volume to be expected for a given season (yield estimate) and the correspondence between crop demand and supply (yield matching) and crop management in order to increase yield production (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

In comparison to causal reasoning techniques, there exists a plethora of Machine Learning algorithms linked to crop management (Mucherino, Papajorgji, & Pardalos, 2009). Features such as counting coffee fruits on coffee trees utilizing coffee pictures extract with the help of Support Vector Machines (SVM) (Ramos, Prieto, Montoya, & Oliveros, 2017) number of immature fruits identified under natural settings such as in (Sengupta & SukLee, 2014) or in rice growth and projection of yield (Su, Xu, & Yan, 2017) detecting cherry branches with full foliage using Bayesian model with Gaussian Naïve Bayes (GNB) algorithm (Amatya, Karkee, Gongal, Zhang, & Whiting, 2016); using ANNs for estimation of grassland biomasses for managed grassland farms by featurizing vegetation indices, spectral band of red and near infrared (NIR) (Ali, Cawkwell, Dwyer, & Green, 2016); wheat output in a given field variation prediction by including normalized values of the projected online soil characteristics and the Standardized Difference Vegetation Index (NDVI) (Pantazi X. , Moshou, Alexandridis, Whetton, & Mouazen, 2016).

The use of agricultural datasets comprising historical information from weather, environment, economic and harvest registers is also observed on ANNs (Kung, Kuo, Chen, & Tsai, 2016) which introduces a method for exact analysis of agricultural productivity forecasting. It should be noted that there are many publications in crop management focusing on issues including subcategories like disease detection and yield forecasts. During the disease detection process, we discovered documents dealing with the detection of nitrogen from stressed, yellow rust infections in wheat, and differentiation between fungal infected and healthy cultures in the case of *Silybum marianum* (Pantazi, et al., 2017), (Pantazi X. E., et al., 2017), (Moshou, et al., 2004), (Moshou, et al., 2005) using and ANNs and Kohonen Self Organizing Maps (SOMs); parasite detection and

classification (Ebrahimi, Khoshtaghaza, Minaee, & Jamshidi, 2017), (Chung, et al., 2016) in explicit crops including both rice and strawberries, disease detection, water stressed parasite detection in wheat (Moshou, Pantazi, Kateris, & Gravalos, 2014) using Support Vector Machines and some approaches even encompass deep learning in diagnosis and detection of generic plants (Ferentinos, 2018), (Varman, Baskaran, Aravindh, & Prabhu, 2017). In yield forecast there have been acknowledged papers using ANNs and SVM for guesstimating yield revenue, (Chlingaryan, Sukkarieh, & Whelan, 2018), (Shao, Zhao, Bao, & He, 2012), (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

The utilization of remote detecting procedures Chingarayan (2018) emphasis in assessing the yield utilizing nitrogen status assessment in soil, while in Shao (2012), is presented a strategy for nitrogen status assessing the in soil for rice, by utilizing the Least Squares SVM model contrasted and Partial Least Square and back proliferation neural organization strategies (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

Several techniques that tend to add something to causal reasoning focus on agricultural design reasoning (Salembier, Segrestin, Berthet, Weil, & Meynard, 2018) and aim at improving the general reasoning of agriculture design in a holistic manner. The study provides a better understanding of modern agricultural techniques and provides a clearer view on the evolution of crops (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020). They also examine the impact and commitment of the design reasoning in order to better understand various reasoning patterns in agriculture and draw attention to certain future research routes in the sense that this design reasoning can fit in with the rest of the process, in the enrichment "designing toolbox" by agronomists and their co-design in agriculture. In agriculture, there are also semantive techniques to causal reasoning as shown (Lagos-Ortiz, Salas-Zárate, Paredes-Valverde, García-Díaz, & Valencia-García, 2020) a retro-reflective knowledge base architecture for smart agricultural decision support for AgriNET. The frame is a rule-based SWRL-based inference mechanism. All of the following techniques do not address the causal inference (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

## Causal inference in livestock management

Livestock management is undoubtedly very important aspect of animal welfare in the area of smart agriculture. Predicting livestock productivity, illnesses and correct decisions is vital to provide the highest possible production of food (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

Although significant livestock management work in connection to machine learning is carried out, there are still lack of causal models in this sector.

Categorization of cattle behavior using bagging with tree learners (Dutta, et al., 2015), calves classification and detection of chewing habits by using decision trees are examples of work in this area (Pegorini, et al., 2015), behavior annotation and animal tracking of the pigs, in order to measure interactive changes in pigs for health and welfare monitoring with the assistance of Gaussian mixture model (GMMs) (Matthews, Miller, Plötz, & Kyriazakis, 2017), estimation of the model of rumen fermentation in cattle with the aid of ANNs from milk fatty acid (Craninx, Fievez, Vlaeminck, & De Baets, 2008), early detection and warning of commercial hen eggs production as well as weight trajectories in cattle using Support Vector Machines (Alonso, Castañón, & Bahamonde, 2013), (Morales, Cebrián, Blanco, & Sierra, 2016), (Alonso, Villa, & Bahamonde, 2015), (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

One study that discusses causal inference is (Rosa & Valente, 2013) which presents an effort at inferring causal connections from empirical evidence in livestock. The article emphasizes the difficulty of confounding in such contexts, where data mining methods were utilized in particular specific circumstances. Further publications dealing with decision-making elements in cattle management that are connected to causal reasoning include (Hastings, Branting, & Lockwood, 2002), (Bello, Ferreira, Gianola, & Rosa, 2018), (Harris, et al., 2016) to name a few (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

### **Causal Inference in water management**

Water management in smart agriculture necessitates substantial effort and plays an important part in climatological, hydrological and agronomic balance.

Data mining studies concentrate mostly on evapotranspiration by estimating the mean monthly reference of arid and semi-arid evapotranspiration using regression techniques (Feng, Peng, Cui, Gong, & Zhang, 2017), (Mehdizadeh, Behmanesh, & Khalili, 2017) weekly evapotranspiration estimates using data gathered from 2 weather stations as well as daily dew point temperature predictions using Artificial Neural Networks (Mohammadi, Shamshirband, Motamedi, Petković, & Gocic, 2015), (Patil & Deka, 2016), (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

(Molina & Zazo, 2017) demonstrate causal reasoning in water management by attempting to analyze the temporal dynamics of river runoff. The examination of causal reason is conducted using Bayesian Networks, which solely consider statistical characteristic dependence without considering confounding and counterfactuals. Another similar technique has been found in (Horne, et al., 2018), (Francis, Guikema, & Henneman, 2014) which focuses on planning the information required for water management systems utilizing conditional probability networks (CPNs), which represent Bayesian based Belief Networks in nature (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

### **Causal Inference in soil management**

The last category of smart agriculture review articles, as defined by (Liakos, Busato, Moshou, Pearson, & Bochtis, 2018) is causal inference in soil management. Contributions to this area address machine learning techniques, especially software packages on predicting and identifying the agricultural soil features, like: soil dryness, condition, temperature, and moisture content estimate Throughout this regard, we can distinguish approaches such as: evaluation of agricultural soil drying, using k-nearest neighbor and Artificial Neural Networks (Coopersmith, Minsker, & Wenzel, 2014), forecasting the organic carbon (OC) of soil, total nitrogen - TN and moisture content - MC, in about 140 soil samples using SVM and regression techniques (Morellos, et al., 2016), estimation of soil temperatures taken from various depths using ANNs (Nahvi,



Habibi, Mohammadi, Shamshirband, & Razgan, 2016) also soil moisture estimation (Johann, Araújo, Delalibera, & Hirakawa, 2016) (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

We see contributions to causal reasoning in winnowed soils for microbiome detection of plant invasiveness in microbial networks (Mamet, et al., 2019) utilizing structural equation modelling (SEM) in soil management (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020). On the other side, (Weyer, et al., 2019) seek to determine risk variables in soils using Bayesian Networks. The conclusion is based entirely on the interdependence of parameters that measure soil restoration in coal areas and their interdependence with Bayesian belief Nets. A causal approach for assessing the danger of heavy metals in soil is described in (Wang, Liu, Chen, Li, & Yu, 2018). The latter utilizes the correlation from data that can be used for causation. The latter makes use of data correlation to determine causality. The study does not explain how causal reasoning is done with only data, and bear in mind that correlation as perceived via causal reasoning does not always imply causality, (Petersen, 2011), (Li, et al., 2013), (Pearl, 2012), (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

#### **4.5 Limitations of the Study**

The methodology extracted from the aforementioned literature review has some limitations:

Specific keywords, such as “causality”, “causal reasoning”, “smart farming”, etc., were taken as a bases in the analyzes of the extracted articles. It indicates that papers not comprising keywords within our search, might have been bypassed in the retrieval procedure (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

Discoveries are based on data composed only from eminent online digital libraries, therefore other web sources which might cover additional case studies on causal reasoning and causal reasoning in smart farming may have been left out of the study (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

## 4.6 Chapter Discussion

The survey cycle is led following the five stages recognized by (Creswell, 2012). The summed-up aftereffects of gathered and investigated related work are assembled by:

- Computer sciences digital library where the study is published - IEEE Xplore, Elsevier, Google Scholar, ACM Digital Library, ResearchGate, Springer Open, DBLP.
- Reference Type categorizing the publication as journal article, conference proceeding, book chapter etc.
- The publication year as well as the country where the research was conducted with the intention of identifying the actuality of the problem definitions.
- Data repositories utilized in manuscripts of causal inference application, include public, synthetic, public-synthetic and theoretical datasets. The intention is to prove the concept that further research in this direction can be done using data sets available online such as Kaggle, Data.gov, Google, Amazon, UCI etc.
- Problem definition in order to identify the domain of causal reasoning application, whereas is in health, smart farming or any other research area.
- Categorization of the publication in harmony with its significance to research questions which are evaluated and answered in the section above (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

# ANALYSIS OF RESULTS

---

### 5.1 Tools and programming language used

According to PYPL index, the statistics done on March 2021, Python is leading the market as far as related to the programming languages searched in Google, followed by Java and JavaScript. Based on those calculations, the more a programming language is searched on the engines, the more populating it results to be (Carbonnelle, 2021). Top ten programming languages sorted by popularity are displayed in the figure 5.1 below.

Rank	Change	Language	Share	Trend
1		Python	30.17 %	-0.2 %
2		Java	17.18 %	-1.2 %
3		JavaScript	8.21 %	+0.2 %
4		C#	6.76 %	-0.6 %
5	↑	C/C++	6.71 %	+0.8 %
6	↓	PHP	6.13 %	+0.0 %
7		R	3.81 %	+0.0 %
8		Objective-C	3.56 %	+1.1 %
9		Swift	1.82 %	-0.4 %
10	↑	Matlab	1.8 %	-0.0 %

Figure 5.1: Top 10 Programming Languages

Python, released initially in Netherlands in 1991, presents a powerful programming language of a high-level, offering a general-purpose approach. It is an interpreted, object-oriented, scripting and scalable programming language that supports packages and standard modules that can be taken for free from the certified web site <https://www.python.org> (Python, 2021). Among the companies that use python on their daily routines are Google, Facebook, Netflix, Instagram, Dropbox, Spotify etc. (Reynolds J. , n.d.). A great feature of this programming language is the Indentation, where the brackets are not used at all and whitespaces become meaningful. Also, the case sensitive variables don't need declaration, since Python figures the appropriate types on its own.



Anaconda Navigator is an open data science platform for programming languages such as R and Python which supports operating systems macOS, Linux and Windows. It provides a huge variety of packages like Jupyter Notebook, PyCharm, Glueviz, JupyterLab, Spyder, Orange 3, RStudio, VSCode, Qt Console (Anaconda Navigator, 2021). Its main advantage is that creates a simply manageable environment to gather data from several sources that support wide assortment of AI, ML as well as DL algorithms.

Jupyter Notebook, primarily was launched as part of IPython Project in 2014, representing an open-source web application that supports collaborative development of data science projects. This browser-based tool powerfully contributes in plotting different visualizations, mathematical equations, text or any other type of media in dozens of programming languages (Jupyter Notebook, 2021). Jupyter notebook also integrates its own native file format .ipynb.



Firstly, let us print the important data related to the Python version, system properties, hardware details as well as timestamps information, by using the `%load_ext watermark`, `%watermark` and `%watermark -v -m -p numpy, scipy, sklearn` magic extension.

```
Python implementation: CPython
Python version       : 3.8.5
IPython version      : 7.18.1

numpy : 1.19.1
scipy : 1.5.0
sklearn: 0.23.2

Compiler      : MSC v.1916 64 bit (AMD64)
OS            : Windows
Release       : 10
Machine       : AMD64
Processor     : Intel64 Family 6 Model 78 Stepping 3, GenuineIntel
CPU cores    : 4
Architecture: 64bit
```

**Listing 5.1:** System Properties

Surveys, as a form of observational studies, often used to gather data from a sample group and consequently have an overview of the entire population. Different kind of surveys include distribution of questionnaires through mail, phone interviews, observations taken from house visits, censuses and similar (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022). In all of the above-mentioned forms, the researcher has no impact on the result obtained. Therefore, a study can focus on factual thoughts or information, depending on the objective of the study (Collecting Data: Surveys, Experiments, & Observational Studies, n.d.), (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).



**Figure 5.2:** Digitalization around the world

Precisely cyberstalking implies using different technologies and the Internet to harass or stalk someone, whether they are one or several individuals. Their identity may be robbed, monitored, falsely charged, threatened, vandalized, scolded, and so on.

We can note that in current culture everybody is associated with the internet in their day-to-day routine exercises and even as a singular it is hard to get by without being associated with innovation. Recently, more cultural consideration has been drawn towards cyberstalking victimization. Existing examination has endeavored to investigate cyberstalking exploitation alongside recognizing factors which increment the danger of being cyberstalked. Consequently, extra exploration is needed to more investigate cyberstalking exploitation.

The topic elaborated in this dissertation deals with current scenario in regard to cybersecurity in general and cyberstalking in particular. Several manuscripts tackle the issue of causality in this cybersecurity, whereas as related to cyberstalking, little or no literature has not been explored in this field.

Web-based and Social Media Platforms (Van Der Walt & Eloff, 2019) simplify the communication worldwide. Yet, on the opposite side, through these online media stages individuals can be focused on for pernicious purposes and one of them is character duplicity. The current research has been done to recognize deception detection. Conduct proof examination might help in exploring advanced (Al Mutawa, Bryce, Franqueira, Marrington, & Read, 2019) unfortunate activities which includes human collaboration among wrongdoer and the victim. It additionally helps in better understanding the elements of explicit offense. Malicious information (Sobhani & Straccia, 2019) may be examined using ontology representation and categorization of complicated semantic events. An embedded method (Cao & Kevin Wang, 2020) of social support from criminology and comorbidities has been used to examine the association of stalking victimization. The study finds that both individual life choices and societal circumstances may be to blame for maltreatment. Cyberstalking is a socially unacceptable practice that occurs on a large scale (Abu-Ulbeh, et al., 2021) and takes various dimensions.

A qualitative approach concluded several interviews with cyberstalking victims was conducted by (Haron & Bt. Mohd Yusof, 2010). The outcomes implied that the main reasons that lead to this phenomenon are love obsession, different love affairs, revenge, and flirting. The study also engages the psychological consequences that come after undergoing a situation like this (Luma-Osmani, Ismaili, & Ram Pal, Building a Model in Discovering Multivariate Causal Rules for Exploratory Analyses, 2021). SLR concerning the causes of cybercrime victimization was elaborated in (Abdullah & Jahan, 2020). A total of 111 papers from e-repositories like Scopus and ASSIA were thematically studied targeting the detection of the causal features that lead to cybercrime. The investigation was carried out from the perspective of cyberbullying, online fraud and harassment, software piracy as well as computer hacking. It elaborates the element that low self-control, lack of awareness, excessive and unrestrained computer usage, exposure of personal data over social networks, psychopathic behaviors are among the issues that can be noted as core causes of cybercrime (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022), (Luma-Osmani, Ismaili, & Ram Pal, 2021).

Scientists in (Abela, Tang, Singh, & Paek, 2020) gave their effort in determining the causes that initiate data breach. The outcomes lead to the fact that malware and hacking are amid the most protruding causes. The Microsoft's framework DoWhy for causal inference (Sharma & Kiciman, 2019) was used for analyses. They proposed an algorithm for running nonstop in the system and perform monitoring on real-time. The agent foremost tasks included recognizing new attacks, keeping attitude towards them, and predicting the future attacks (Mugan, 2013). A network security traffic analysis focusing on the detection of stealthy malware has been performed in (Zhang, Yao, Ramakrishnan, & Zhang, 2016). The researchers calculated the triggering causal rapport based on the network requests. Frequent such activities have been apparent, considering unauthorized transfer of data, spyware and DNS bots on numerous hosts (Luma-Osmani, Ismaili, & Ram Pal, 2021).

Diverse attack discovery on sensor-based information, built upon the causal relations between the current measurement signals the has been elaborated in (Shi, Guo, Johansson, & Shi, 2018). Four kinds of occurrences have been explored, including: replay attacks, innovation-focused deception, denial-of-service, and data injections (Luma-Osmani, Ismaili, & Ram Pal, 2021).

Previous systematic literature (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020), (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020) emphasizes the fact that less attention has provided to the causal rules compared to association ones, albeit knowledge turn out to be more meaningful when is denoted by causal models, instead of associative models (Rammohan, 2010), (Luma-Osmani, Ismaili, & Ram Pal, 2021).

The research aims to find the causes and effects of cyberstalking in high school's adolescents in the city of Tetovo.



## 5.2 Method

### 5.2.1 Ethics

As an essential fragment of each survey, the ethical declaration assures partakers that research purposes are the only aim of data utilization, simultaneously all the credentials will be saved confidentially (Luma-Osmani, Ismaili, & Ram Pal, 2021).

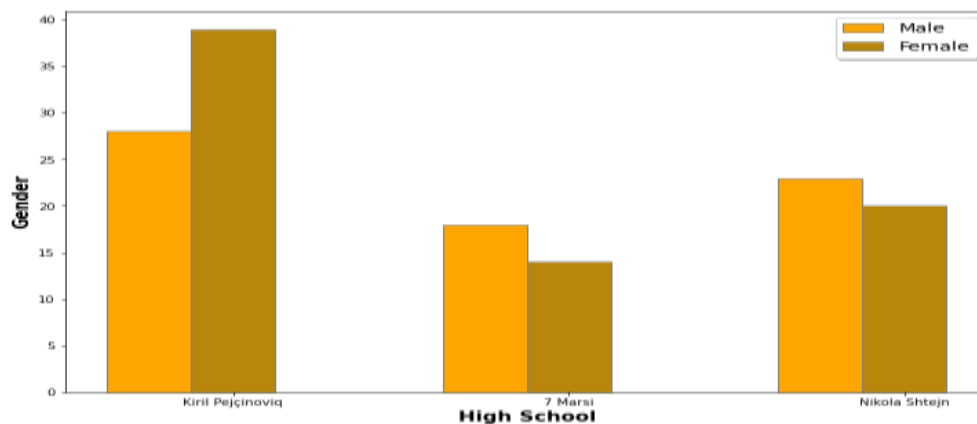
### 5.2.2 Participants

“Cyberstalking” dataset covers 142 data records collected from 3 high schools, Gymnasium “7 Marsi”, “Kiril Pejčinović” and the Medical High School “Nikola Shtejn” located in the city of Tetova in North Macedonia. Data distribution has been visually acquiesced in table 5.1 and in figure 5.3 as well (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022), (Luma-Osmani, Ismaili, & Ram Pal, 2021).

**Table 5.1:** Cyberstalking Dataset Information’s

School	Male	Female	Total
Gymnasium “Kiril Pejčinović”	28	39	67
Gymnasium “7 Marsi”	18	14	32
Medical High School “Nikola Shtejn”	23	20	43
	69	73	142

It can be noticed that the majority of participants, 67 in total were from high school “Kiril Pejčinović”, 32 from “7 Marsi” and 43 members from the Medical High School “Nikola Shtejn” (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022), (Luma-Osmani, Ismaili, & Ram Pal, 2021).



**Figure 5.3:** Participant’s data

It should be considered that the graphs were plotted in Jupyter Notebook, using the matplotlib Python library. The same is widely used for creating different kind of visualizations, whether static, interactive or animated ones (Luma-Osmani, Ismaili, & Ram Pal, 2021). The extract of the code used for plotting the participants gender based on the institutions they belong to is clearly showed in listing 5.2.

```
barWidth = 0.24

fig = plt.subplots(figsize =(12, 8))

M = [28, 18, 23]
F = [39, 14, 20]

br1 = np.arange(len(M))
br2 = [y + barWidth for y in br1]

plt.bar(br1, M, color ='orange', width = barWidth, edgecolor ='grey',
label ='Male')

plt.bar(br2, F, color ='darkgoldenrod', width = barWidth, edgecolor
='grey', label ='Female')

plt.xlabel('High School', fontweight ='bold', fontsize = 15)
plt.ylabel('Gender', fontweight ='bold', fontsize = 15)

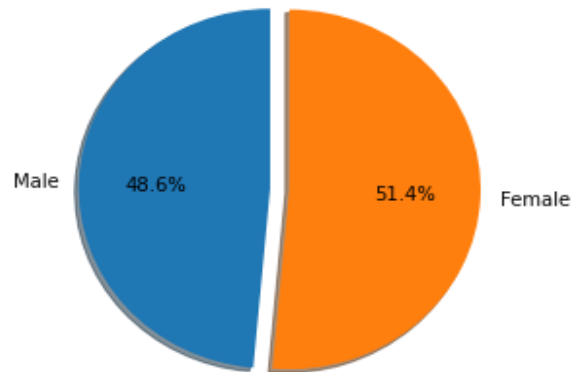
plt.xticks([r + barWidth for r in range(len(M))], ['Kiril Pejčinović',
'7 Marsi', 'Nikola Shtejn'])

plt.legend(loc=0, shadow=True, fontsize='x-large')plt.xticks([r +
barWidth for r in range(len(M))], ['Kiril Pejčinović', '7 Marsi',
'Nikola Shtejn'])

plt.legend(loc=0, shadow=True, fontsize='x-large')
```

**Listing 5.2:** Participant's data plotting code

Figure 5.4 shows the overall number of subjects by gender, out of a total of 142 subjects 73 (51.40 %) were females and 69 (48.59 %) males (Luma-Osmani, Ismaili, & Ram Pal, 2021). It's noticed that the sample is more or less symmetrical.



**Figure 5.4:** Gender

**Listing 5.3:** Gender plotting code

```
import matplotlib.pyplot as plt
%matplotlib inline
alpha_color=0.3

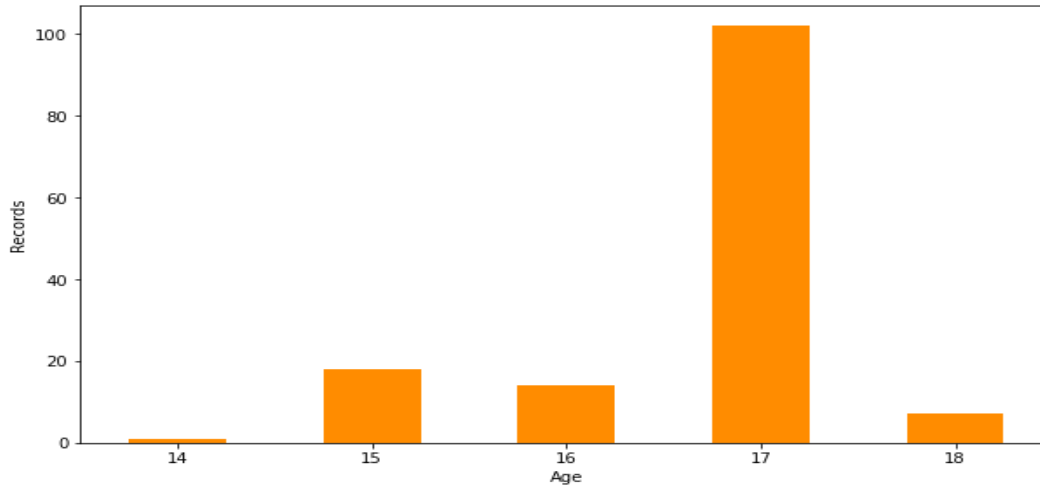
labels = 'Male', 'Female'
sizes = [48.591549, 51.408451]
explode = (0, 0.1)

fig1, ax1 = plt.subplots()

ax1.pie(sizes, explode=explode, labels=labels,
autopct='%1.1f%%', shadow=True, startangle=90)

ax1.axis('equal')
```

Figure 5.5 below shows the inclusion of participants by age in our research, based on the situation data it can be concluded that the majority consists of adolescents 17-year-olds, represented in percentage of 71.83 %, the other group is followed by 15-year-olds with 12.67%, 16-year-olds with 9.85%, 18-year-olds 4.92 % and a smaller number of subjects included in the research have been 14-year-olds with 0.7%. It illustrates that the mean age is ( $M = 16.67$ ) whereas the minimum age ( $Min = 14$ ) and the maximum age ( $Max = 18$ ) (Luma-Osmani, Ismaili, & Ram Pal, 2021).



**Figure 5.5:** Age

**Listing 5.4:** Participants Age plotting code

```
plt.figure(figsize=(10,6))  
  
ds['Age'].value_counts().sort_index(axis=0).plot(kind='bar', color='darkorange')  
  
plt.xlabel('Age')  
plt.ylabel('Records')  
plt.xticks(rotation=360)
```

### **5.2.3 Procedure**

The data in the survey we are going to present in the dissertation, are taken from a questionnaire conducted physically in period January-February 2020. The classes were randomly selected, and the questionnaire was fulfilled by the students, with the prior permission of their teachers. After that, the data were transferred to electronic devices for further analysis and processing (Luma-Osmani, Ismaili, & Ram Pal, 2021).

### **5.2.4 Limitations**

It ought to be borne as a top priority that the acquired study results won't give data to the general demeanor of the teenagers, since the exploration was led distinctly with Albanian understudies, in three secondary schools in the region of Tetova (Luma-Osmani, Ismaili, Pathak, & Zenuni,

2022). The data provided here are exclusively for teens (limited age), Albanians and in the area of Tetovo, and although the sample is chosen randomly, the fact should be obvious (region centered). If we take the following factors into account, the findings should not be biased (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

### 5.2.5 Data Preprocessing

Prior analyzing the dataset and proceeding further, data preprocessing has been performed. The unnecessary raw data was dropped off from the set of data. Data cleaning included modifying, correcting and formatting relevant parts consisting of columns or rows within the database (Luma-Osmani, Ismaili, & Ram Pal, 2021).

## 5.3 Cyberstalking Dataset

Since the dissertation's main focus is on data that can be reached online, the Cyberstalking Dataset is publicly published and retrieved in Kaggle repository on the link that follows: <https://www.kaggle.com/shkurtelumaosmani/cyberstalking-in-tetovo> (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022), (Luma-Osmani, Ismaili, & Ram Pal, 2021).

The same is explained in detail in the table 5.2 below. It comprises of 15 variables, 7 of them are textual (quality variables or nonmetric) data type and 8 numeric (quantity variables) (Luma-Osmani, Ismaili, & Ram Pal, 2021).

**Name of the dataset:** Cyberstalking

**Type of the dataset:** Observational

**Size of the dataset:** 142 records, 15 variables (Luma-Osmani, Ismaili, & Ram Pal, 2021)

**Table 5.2:** Variable's description

#	Variable Name	Type	Description
1	School	String	The name of the high school. It may contain three values: "Kiril Pejčinović", "7 Marsi" and "Nikola Shtejn".
2	Gender	String	Gender

<b>3</b>	<b>Age</b>	<b>int64</b>	<b>Age</b>
<b>4</b>	Victim_of_cyberstalking	int64	Information whether the respondent has ever been a victim of cyberstalking (Y/N).
<b>5</b>	Social_media_cyberstalking	String	If the stalking took any electronic form, it is an indication of which form the stalker used most often. Options are a) Facebook b) Instagram c) Snapchat d) Twitter e) Other.
<b>6</b>	Form_of_harassment	String	The form in which the harassment was done. Options vary from a) Threatening or abusive emails b) Threatened in chat rooms or comments c) Posted false information d) encouraged others to harass or insult you e) Ordered goods online in your name f) any other behaviour founded distressing in any way.
<b>7</b>	Know_the_stalker	String	Whether the stalker is known or not. Assuming the categories: a) No b) ex-partner c) friend d) other.
<b>8</b>	Social_media_communication	float64	Did the stalker communicated via the social media platforms (Y/N)
<b>9</b>	Cyberstalking_frequency	String	On average, how often was received some sort of contact from the stalker. Optional values are a) Hourly b) Daily c) Weekly d) Monthly e) 2-3 months.
<b>10</b>	Get_rid_of_the_cyberstalker	float64	Did the respondent managed to get rid of the stalker so he/she stopped harassing (Y/N).
<b>11</b>	You_harassed_someone	int64	Knowing whether the respondent has ever harassed someone (Y/N).

<b>12</b>	<b>Cyberstalking_purpose</b>	<b>String</b>	<b>If the 11<sup>th</sup> question is yes, what was the purpose that made him/her cyber stalk someone. A choice between a) Personal issues b) They harassed me earlier c) For gossip in group chat d) Other.</b>
<b>13</b>	<b>Cyberstalking_achieving_goal</b>	<b>float64</b>	<b>Reaching the goal while cyberstalking (Y/N).</b>
<b>14</b>	<b>Cyberstalking_pleasure</b>	<b>float64</b>	<b>Finding pleasure whilst stalking someone (Y/N).</b>
<b>15</b>	<b>Criminal_offense</b>	<b>float64</b>	<b>Knowing that cyber stalking is in fact a criminal offense (Y/N).</b>

In the count plot beneath, we have depicted different types of harassment for persons who have or have not been victims of cyberstalking. We can see that just two persons were not victims, and both of them claimed that someone purchased products online using their information. More than twenty individuals who have been harassed, stated that the kind of harassment they experienced was publishing false information about themselves; this is also the most common form of harassment (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022). The second most common kind of abusive behavior (fewer than 20, more than 15 persons) was defined as any other conduct that was disturbing in any manner. The third form is harassed in chat rooms or remarks by approximately 15 persons. Five individuals have reported receiving threatening or abusive e-mails. There have been less than 5 people who have been harassed in both ways, such as publishing fake information and any other conduct that has been considered disturbing in any manner. Various combinations of other forms have been recorded by fewer than three people. The count plot may show this in greater depth (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

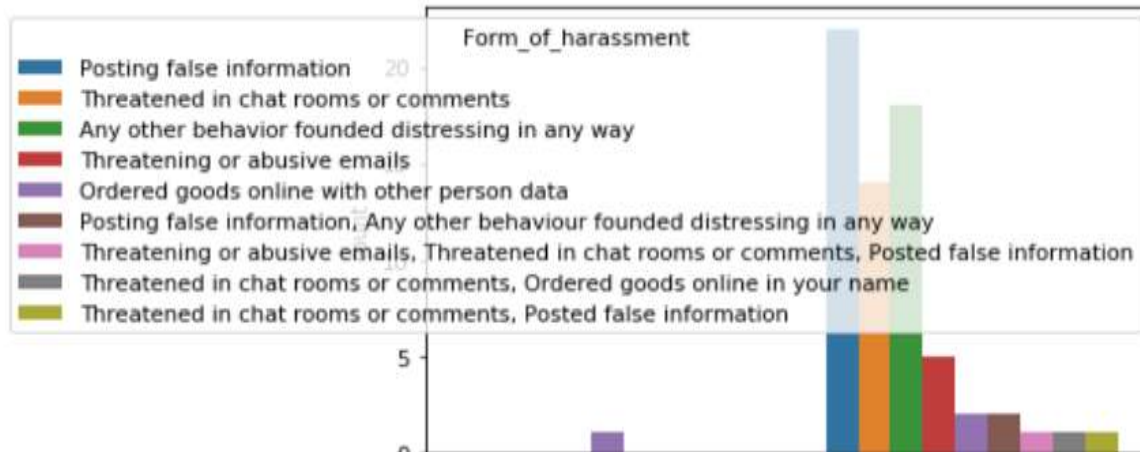


Figure 5.6: Form of harassment

Intensity of cyberstalking in victims of cyberstalking differs. Initially let us clarify the blue bar at the 0-left side of the subplot. It implies that one individual that isn't a victim of cyberstalking has been harassed one time monthly (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022). This might be a blunder on the grounds that the victims of cyberstalking have been requested to answer this part while the others no, or it is possible that the individual who addressed was not felt as a victim of cyberstalking, even though in any case has been bothered one time each month (or perhaps only once in a lifetime). A greater number of participants have been harassed twice or three times every month (more than 12). Every week, about ten people have been harassed oncely. Every day, 9 individuals are harassed, and 7 people are harassed every hour, which is quite alarming (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

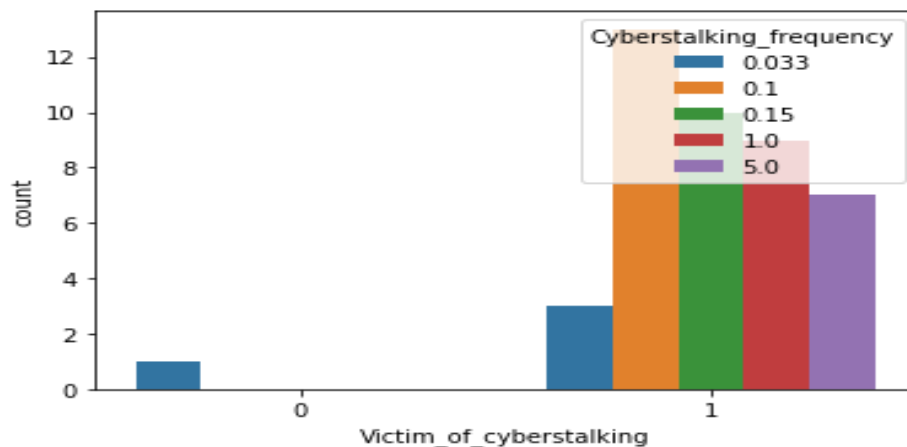
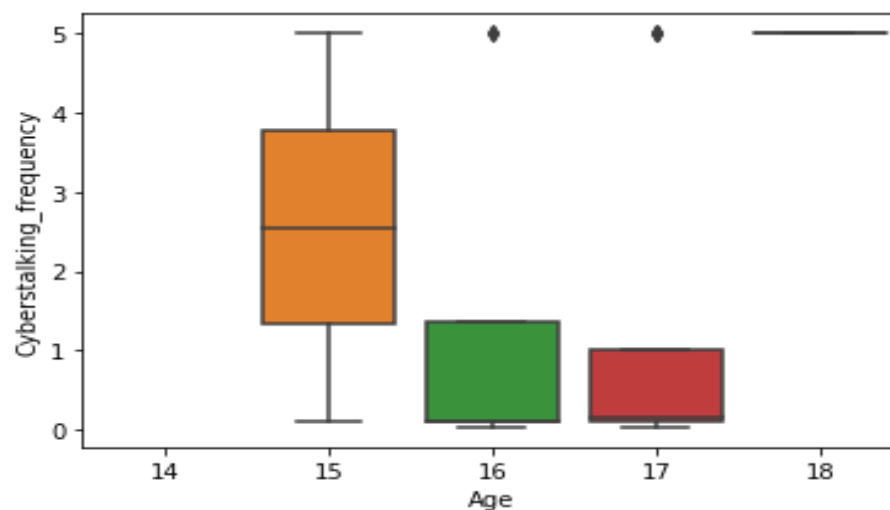


Figure 5.7: Cyberstalking frequency



Boxplots, as we know, provides five kinds of summaries: the minimum, maximum, first quartile, median, and third quartile. The size of the box (which reflects the interquartile range) indicates that the provided data has a wide range of dispersion around its median value (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022). The 15-year-old kids have a typical box plot with no outliers; the greatest frequency of cyberstalking in this age group is 5 times, and 0 is the minimum. Since the total observations number is even, consequently the median is 2.5. (142 in total). The other two groups have outliers, and their highest and minimum values (number of times they have been cyberstalked) are both 5 and 0 (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).



**Figure 5.8:** Age based cyberstalking frequency

## 5.4 Correlation

We can see a substantial connection in a single plot called heatmap from the Seaborn library to examine the correlation between the variables in the dataset and the direction in which their linear relationship is (positive or negative). Correlations, on the other hand, may or may not indicate a causal link (Wang & Mueller, 2016), (Cooper G. F., 1997) Correlation indicates the degree to which two factors are linearly associated. Linearity is observed when specified correlation coefficients are observed. The Pearson's coefficient of correlation is applied down here (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

Python data visualizations using the Matplotlib and Seaborn libraries are an excellent method to rapidly evaluate the relationships amid attributes in a set of data. As a result, the stronger the color's shade, the greater the magnitude of the association. Of course, the correlation of a column with itself always yields a value of one, which is known as the perfect positive correlation. The values that emerge closer to 0 indicate that there is no linear trend spotted between the two columns, whereas the values that appear closer to the value 1, point to the fact that variables are more significantly linked to each and every one, it means that both of them will increase or decrease synchronously, whereas the values that appear closer to -1 indicate that variables are inversely associated, i.e., one increases the other (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

According to Karl Pearson correlation (ElegantJ BI, 2018), the factor can signpost the level of the correlation amid the variables, and can be calculated as noted below (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).



Figure 5.9: Pearson correlation coefficient

The coefficient, indicating how strong this linear relationship is, can be calculated using the below described equation:

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 - \sum (y_i - \bar{y})^2}}$$

- *r* - correlation coefficient
- *x<sub>i</sub>* - values of the variable *x*
- *y<sub>i</sub>* - values of the variable *y*

- $\bar{x}$  - mean of the variable  $x$
- $\bar{y}$  - mean of the variable  $y$

There are some notable connections between the provided variables (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022). It would provide a more comprehensive examination of the model developed for the dependency of variable "Victim\_of\_cyberstalking" Consequently, the matrix yields the following results:

Having started with the dependent variable "Victim\_of\_cyberstalking" we may conclude that, while there is some evidence of linear connection to other independent variables, there are no significant coefficients of association, even positive or negative. For example, there is a modest connection between "Age" and "School" valued at 0.36, as well as a minor positive linear correlation of 33% among victims of cyberstalking and getting rid of the stalker. The positive linear connection between victims of cyberstalking and if you have tormented someone is 27%. Addressing the other variables, which are assumed to be independent to some degree in a linear analysis, there are notable coefficients across schools and a 25% positive connection if you harassed somebody (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

It's worth noting the -38 percent negative correlation coefficient across "Victim\_of\_cyberstalking" and "Cyberstalking\_achieving\_goal" The victims of cyber stalking had lower correlation coefficients with gender and other factors. "School" is also inversely related to "Cyberstalking\_pleasure" with a negative 26% correlation. Also here are two weak correlations involving "Age" and "Pleasure," as well as "School" and "Social\_media\_communication" both of which equaled -0.23, indicating that the high school "Kiril Pejčinović" communicates more through social media than "Nikola Shtejn" and "7 Marsi". "Cyberstalking\_pleasure" and "Cyberstalking\_achieving\_goal" have a higher coefficient of correlation, with a positive of 47 percent. Such finding is predicted since there is a theoretical link between these two sociopathy assessed factors, as shown in further detail in the heatmap 5.10 below (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

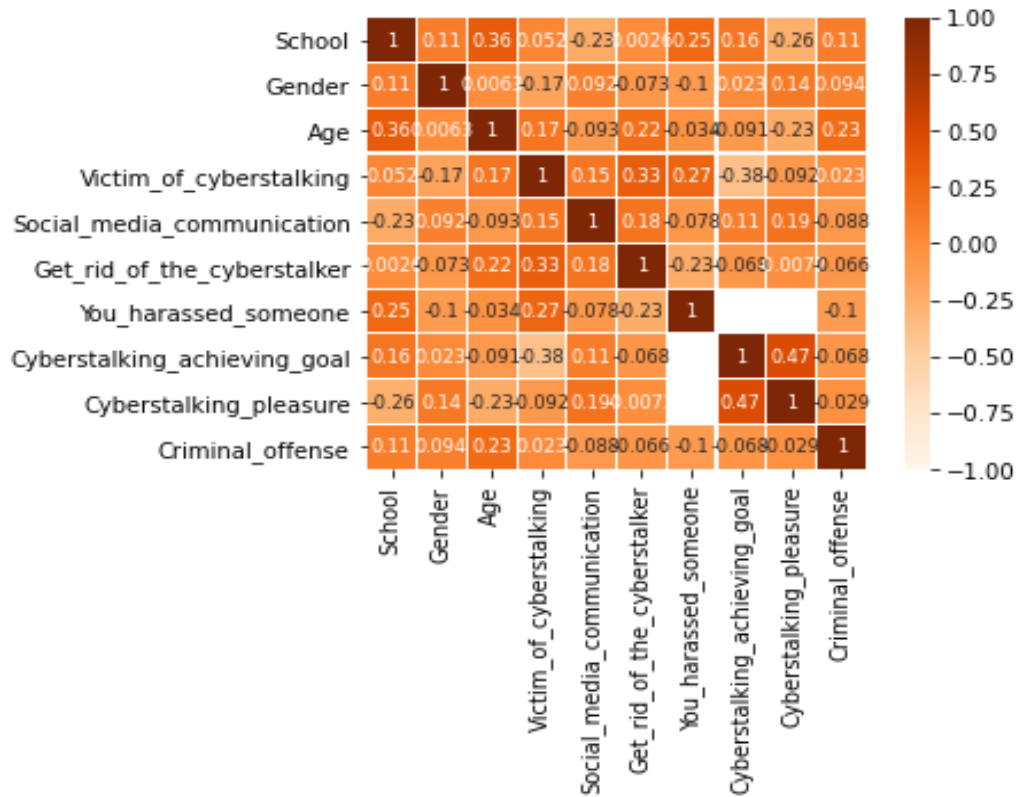


Figure 5.10: Heatmap Correlation

## 5.5 Discovering causality

Keeping in mind that causal rules indicate associations but that the opposite is not true always, our technique began by first generating association rules, which were then discovered and evaluated. The Apriori method was utilized to generate the AR's, and the total number of association rules was 446. The metric sort was Confidence with a value of 0.7, the lower bound for minimal support was valued 0.1, whereas the upper bound for minimum support is numbered 1.0, and the delta factor value for progressively decreasing support was 0.05 (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

Consequently, the top 10 acquired association rules from the "Cyberstalking" dataset are itemized below, although the full list is available in the Appendix B (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

1. Gender=Female Get\_rid\_of\_the\_cyberstalker=Yes 36 ==>  
Victim\_of\_cyberstalking=Yes 36 <conf:(1)> lift:(2.15) lev:(0.14) [19] conv:(19.27)
2. Social\_media\_communication=Yes 41 ==> Victim\_of\_cyberstalking=Yes 41  
<conf:(1)> lift:(2.15) lev:(0.15) [21] conv:(21.94)
3. Get\_rid\_of\_the\_cyberstalker=Yes 59 ==> Victim\_of\_cyberstalking=Yes 59  
<conf:(1)> lift:(2.15) lev:(0.22) [31] conv:(31.58)
4. Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes 38 ==>  
Victim\_of\_cyberstalking=Yes 38 <conf:(1)> lift:(2.15) lev:(0.14) [20] conv:(20.34)
5. Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=Yes 38 ==>  
Victim\_of\_cyberstalking=Yes 38 <conf:(1)> lift:(2.15) lev:(0.14) [20] conv:(20.34)
6. Cyberstalking\_achieving\_goal=Yes 37 ==> You\_harassed\_someone=Yes 37  
<conf:(1)> lift:(2.41) lev:(0.15) [21] conv:(21.63)
7. Cyberstalking\_pleasure=No 33 ==> You\_harassed\_someone=Yes 33 <conf:(1)>  
lift:(2.41) lev:(0.14) [19] conv:(19.29)
8. Social\_media\_cyberstalking=Instagram 31 ==> Victim\_of\_cyberstalking=Yes 31  
<conf:(1)> lift:(2.15) lev:(0.12) [16] conv:(16.59)
9. Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=Yes 31 ==>  
Victim\_of\_cyberstalking=Yes 31 <conf:(1)> lift:(2.15) lev:(0.12) [16] conv:(16.59)
10. Social\_media\_cyberstalking=Instagram Get\_rid\_of\_the\_cyberstalker=Yes 30 ==>  
Victim\_of\_cyberstalking=Yes 30 <conf:(1)> lift:(2.15) lev:(0.11) [16] conv:(16.06)

Multiple investigations (Spirtes, Glymour, & Scheines, 2000), (Shimizu, Hoyer, Hyvärinen, & J., 2006), (Budhathoki, Boley, & Vreeken, 2018), (Li, et al., 2013), (Pearl, 2010), have used causality principles with probabilistic distributions generated with the help of acyclic networks.

In the book (Causation, Prediction and Search 2nd Edition, 2000) Spirtes et al. stress that research based on diverse experiments and observations do not necessarily provide the same deductions and conclusions. Furthermore, these agreements demonstrate that there is a deep bond between causality and probability, and that this correlation can help many more topics in statistics, such as comparative power of observation amidst experiment, or as the Simpson's

paradox is recognized, errors in regression models, sampling, and inconstant selection (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

Simultaneously, many of the causal rules that we are interested in do not display perfect correlations, and man-made measures are not perfect, thus scientific causal models are often probabilistic in character (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

The idea of cause-and-effect must be contextualized into independent and dependent variables in the study; this is the first and most important step in establishing causal links in scientific inquiry (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

According to (Li, et al., 2013) the causal rule is given as a combined cause of two or more binary parameters,  $(X_1, X_2, \dots, X_n, Y)$ , in which the subsets of X represent the causative variable (LHS) and Y represents the effect (RHS). The attributes of the approach of employing combined variables in determining causes are based on the fact that such factors alone do not suggest causation, but their combination may (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

The results of the Apriori association rules are the starting point for the hybrid method given in the thesis. As a result, there seems to be a correlation between the variables Gender and Victim of cyberstalking, and the condition asserts that: **Gender=Female  $\rightarrow$  Victim\_of\_cyberstalking=Yes** (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

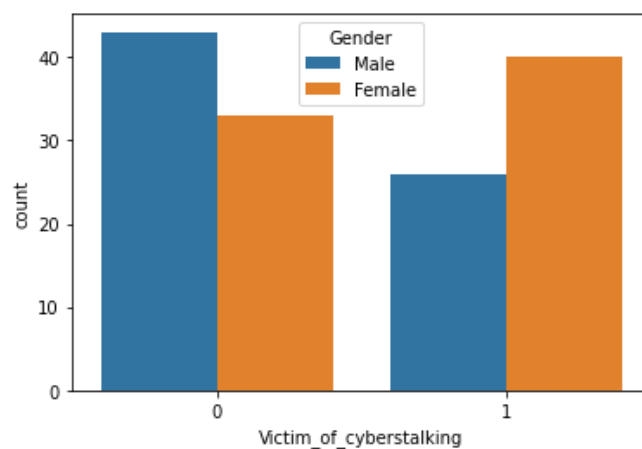
**Table 5.3:** Two variables ratio

	<b>Victim of Cyberstalking = Yes</b>	<b>Victim of Cyberstalking = No</b>
<b>Gender = Female</b>	40/73 = 0.548	33/73 = 0.452
<b>Gender = Male</b>	26/69 = 0.377	43/69 = 0.623

In the female group, the ratio of being a victim of cyberstalking to not being a victim of cyberstalking is 1.21:1, but in the male group, it is 0.60:1. A female's chance of becoming a victim of cyberstalking is 1.21, whereas a male's chance of being a victim of cyberstalking is 0.60. This implies that girls are twice as likely as guys to be cyberstalked. After all, when the odd ratio is 1,

it indicates that a variable has an equal chance of appearing in both gender groups, which may not be the case in our instance (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

In the countplot presented bellow we can notice intently the count of casualties of cyberstalking by gender identity. The irregularity in the length of the bar is obvious, and it has a mirrored structure in certain ways. Those who have not been cyberstalked outnumber those who have been, and vice versa, those who have been cyberstalked outnumber those who have been. Figure 5.11 depicts the created rule graphically as well (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).



**Figure 5.11:** Victim of Cyberstalking variable

If we go further into the investigation and add another trait, “**You\_harassed\_someone**”, the resultant 5.4 table is as follows (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022):

	Victim of Cyberstalking = Yes	Victim of Cyberstalking = No
<b>Gender = Female &amp; You_harassed_someone = Yes</b>	23/34=0.676	11/34=0.324
<b>Gender = Female &amp; You_harassed_someone = No</b>	17/39=0.436	22/39=0.564
<b>Gender = Male &amp; You_harassed_someone = Yes</b>	14/25=0.56	11/25=0.44
<b>Gender = Male &amp; You_harassed_someone = No</b>	12/44=0.273	32/44=0.727

**Table 5.4:** Three variables ratio

As a result, when we look at the third variable, “**Victim\_of\_cyberstalking**” we see that 68 percent of women who have harassed someone have been victims of such occurrences, compared to 32 percent who have not. On the other hand, 44 percent of females who have never harassed anybody have been a victim of cyberstalking, and 56 percent have never bothered someone or been a victim (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

In terms of males, 56 percent have indeed harassed someone and have been a victim of cyberstalking, and 44% have done such badgering, yet not been a casualty. Lastly, 27% haven't irritated anybody, yet anyway encountered a following and 73% haven't done provocation or had such an outcome (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

Let us recently add another variable "Age" to observe how the same impacts on the variable “Victim\_of\_cyberstalking”. The relevant rules show that the possibility of greater victimization increases with age (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

**Table 5.5:** Four variables ratio

<b>Victim of Cyberstalking</b>	<b>Yes</b>	<b>No</b>
Gender = Female, Age = 14 & You_harassed_someone = Yes	0	0
Gender = Female, Age = 15 & You_harassed_someone = Yes	1/5=0.2	4/5=0.8
Gender = Female, Age = 16 & You_harassed_someone = Yes	3/3=1	0
Gender = Female, Age = 17 & You_harassed_someone = Yes	19/26=0.73	7/26=0.27
Gender = Female, Age = 18 & You_harassed_someone = Yes	0	0
Gender = Female, Age = 14 & You_harassed_someone = No	0	0
Gender = Female, Age = 15 & You_harassed_someone = No	1/5=0.2	4/5=0.8
Gender = Female, Age = 16 & You_harassed_someone = No	3/4=0.75	1/4=0.25
Gender = Female, Age = 17 & You_harassed_someone = No	11/27=0.407	16/27=0.593
Gender = Female, Age = 18 & You_harassed_someone = No	2/3=0.667	1/3=0.333
Gender = Male, Age = 14 & You_harassed_someone = Yes	0	0
Gender = Male, Age = 15 & You_harassed_someone = Yes	0	3/3=1
Gender = Male, Age = 16 & You_harassed_someone = Yes	2/2=1	0
Gender = Male, Age = 17 & You_harassed_someone = Yes	12/20=0.6	8/20=0.4
Gender = Male, Age = 18 & You_harassed_someone = Yes	0	0
Gender = Male, Age = 14 & You_harassed_someone = No	0	1/1=1
Gender = Male, Age = 15 & You_harassed_someone = No	0	5/5=1
Gender = Male, Age = 16 & You_harassed_someone = No	3/5=0.6	2/5=0.4
Gender = Male, Age = 17 & You_harassed_someone = No	8/29=0.276	21/29=0.724
Gender = Male, Age = 18 & You_harassed_someone = No	1/4=0.25	3/4=0.75



(Budhathoki, Boley, & Vreeken, 2018) in the following, indicate the deviation value of conditional Y on causal rules, whether they are true or not (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022):

$$e(\sigma) = E[Y | \sigma = \mathbf{T}] - E[Y | \sigma = \mathbf{F}] \quad \dots (1)$$

Many writers have popularized and discovered notions related to causality by employing probability distributions based on directed acyclic networks (Pearl, 2010), (Spirtes, Glymour, & Scheines, 2000), (Shimizu, Hoyer, Hyvärinen, & J., 2006). We seem to be more focused on the explanation provided by (Budhathoki, Boley, & Vreeken, 2018) it takes into account the explanations of the other writers as well (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

The triangle used in this study for causal analysis is constructed using the following principles and connections:

- Y is the binary target variable and, in our case, it is our victims of cyberstalking variable.
- $X_1$ ,  $X_2$  and  $X_3$  as a subset of X are description variables and, in our case, we have three of them: gender, age and the fact if you have harassed someone or not (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

Along these lines, we can compose the accompanying dependent on the earlier data:  $Y = \{0, 1\}$ , and we realize that it numerically addresses the space of Y. Then again, the space of  $X_i$  is either real or categorical. In our case it is both. (Budhathoki, Boley, & Vreeken, 2018) say that the domain of X is an M dimensional external item space  $X = X_1 \times X_2 \times \dots \times X_m$  (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

For the sigma causal principles referenced in the situation (1) we can characterize "an unbiased intervention  $do(\sigma)$  as the randomized operation of satisfying  $\sigma$  by setting  $X_\sigma$  to some  $x$  such that  $\sigma(x) = true$  according to the probabilities  $p(X_\sigma = x | \sigma = True)$ " and one needs to find rules of causal reasoning sigma that maximize the causal effect defined as the difference of expected value of Y which would be under two "interventions"  $do(\sigma)$  and  $do(\neg \sigma)$  (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022):

$$e(do(\sigma)) = E[Y | do(\sigma)] - E[Y | do(\neg \sigma)] = p(Y = 1 | do(\sigma)) - p(Y = 1 | do(\neg \sigma))$$

The proceed approach is built on a Bayesian network, a methodology that represented a major leap in cause discovery (Luma-Osmani, Ismaili, Zenuni, & Raufi, 2020) with each new node on the problem based on the conditional probability distribution table, as seen in the graph below (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

- $X_1$  – Gender (0: Female, 1: Male)
- $X_2$  - Knowing whether the participant harassed someone (0: No, 1: Yes)
- $X_3$  - Age (0: Doesn't have 17 years, 1: Has 17 years)
- $Y$  - Victim of cyberstalking (0: No, 1: Yes) (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022)

**Table 5.6:** Cyberstalking Bayesian Network

			$p(X_1=0)$	$p(X_1=1)$
			0.51	0.49

$X_1$	$p(Y=1 X_1)$	$p(Y=0 X_1)$
0	0.55	0.45
1	0.38	0.62

$X_1$	$X_2$	$p(Y=1 X_1, X_2)$	$p(Y=0 X_1, X_2)$
0	1	0.68	0.32
0	0	0.44	0.56
1	1	0.56	0.44
1	0	0.27	0.73

$X_1$	$X_2$	$X_3$	$p(Y=1 X_1, X_2, X_3)$	$p(Y=0 X_1, X_2, X_3)$
0	1	0	0	0
0	1	0	0.2	0.8
0	1	0	1	0
0	1	1	0.73	0.27
0	1	1	0	0
0	0	0	0	0
0	0	0	0.2	0.8
0	0	0	0.75	0.25
0	0	1	0.41	0.59
0	0	1	0.67	0.33
1	1	0	0	0
1	1	0	0	1
1	1	0	1	0
1	1	1	0.6	0.4
1	1	1	0	0
1	0	0	0	1
1	0	0	0	1
1	0	0	0.6	0.4
1	0	1	0.28	0.72
1	0	1	0.25	0.75

## 5.6 Multiple logistic regression model

Knowledge becomes more meaningful when the same is represented by causal models, rather than associative models (Rammohan, 2010). There are several models that one can build in order to see if there is any relation and a way of prediction of binary variables. Logit models, which have a binary explained or dependent variable and other independent or explanatory variables could be metric or non-metric (nominal or ordinal). The idea behind these models is the same as those of linear regression models, but just few things regarding the estimation of the coefficients are different (Luma-Osmani, Ismaili, & Ram Pal, 2021). The coefficient estimation of logit models is done using maximum likelihood method. Let us briefly see the difference between the three mentioned models and then discuss why we chose the logit model as a form of presentation (Luma-Osmani, Ismaili, & Ram Pal, 2021).

As far as is known, the regression parameters can be estimated using two methods: OLS (Ordinary Least Square) - Calculus based and MLE (Maximum Likelihood Estimation) - Probability based. However, the results in most of cases lead to the same coefficients, since minimizing sum square distances is the same as trying to maximize the probability of occurrence (Luma-Osmani, Ismaili, & Ram Pal, 2021).

The main aim through the study was to notice and establish the causes that lead to cyberstalking. In the survey this concern is labeled with the question: Have you ever been a victim of cyberstalking? Other questions, which could be divided into two categories, whether they are in the controlled or uncontrolled group, are the ones that give us the idea of what could cause the cyberstalking in the first place (Luma-Osmani, Ismaili, & Ram Pal, 2021).

An OLS method, as the linear or one factorial model for example, could be written as (Luma-Osmani, Ismaili, & Ram Pal, 2021):

$$y = x'\beta + e$$

According to our prior mathematics knowledge, a straight forward access to the coefficients of this model using OLS could be done. On the other hand, the binary models are ranked as second most used models in multivariate statistical study, next to the linear regression which takes the

first place. Binary models that used regression and therefore estimated with OLS cannot be interpreted, because the forecast likelihoods would not be restricted among 0 and 1. Additionally, this can be noted as (Luma-Osmani, Ismaili, & Ram Pal, 2021):

$$P = pr [y = 1 | x] = x' \beta, \text{ where } x' = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ X_n \end{bmatrix}$$

Noted mathematically, the dependent  $Y$  variable and the descriptive  $X_1, X_2, \dots, X_k$  independent variables, can be written in the formula of multiple linear regression (Luma-Osmani, Ismaili, & Ram Pal, 2021):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

Alternatively, if we transcribe the logit model and its probability density function - pdf, as noted (Luma-Osmani, Ismaili, & Ram Pal, 2021):

$$P_i = E (Y = 1 | X_i) = \beta_1 + \beta_2 X_i$$

However, in practice there is no infinite exponential growth, since firstly, we have low growth, then rapid growth and lastly slow growth plotted in “S” shaped curve by the Sigmoid function, comprising the Euler’s number  $e=2.71728$ , as follows (Luma-Osmani, Ismaili, & Ram Pal, 2021):

$$P_i = \frac{1}{1+e^{-y}} = \frac{1}{1+e^{-(\beta_1+\beta_2 X_i)}} = \frac{1}{1+\frac{1}{e^{(\beta_1+\beta_2 X_i)}}} = \frac{1}{\frac{e^{(\beta_1+\beta_2 X_i)}+1}{e^{(\beta_1+\beta_2 X_i)}}} = \frac{e^{(\beta_1+\beta_2 X_i)}}{1+e^{(\beta_1+\beta_2 X_i)}}$$

If we replace  $y_i = \beta_1 + \beta_2 X_i$  the final result should take the form  $P_i = \frac{e^{y_i}}{1+e^{y_i}}$  which also represents the logistic distribution function (Gujarati & Porter, 2009), (Luma-Osmani, Ismaili, & Ram Pal, 2021).

The probability of not happening event, i.e.,  $(1 - P_i) = 1 - \frac{e^{y_i}}{1+e^{y_i}} = \frac{1+e^{y_i}-e^{y_i}}{1+e^{y_i}} = \frac{1}{1+e^{y_i}}$

To end, the odds ratio between  $P_i$  and  $1 - P_i$  is (Luma-Osmani, Ismaili, & Ram Pal, 2021):

$$\frac{P_i}{1 - P_i} = \frac{\frac{e^{y_i}}{1 + e^{y_i}}}{\frac{1}{1 + e^{y_i}}} = \frac{e^{y_i} (1 + e^{y_i})}{1 + e^{y_i}} = e^{y_i}$$

The natural logarithm of above-mentioned equation leads to interesting outcome in the variable L - logit, henceforth the model's name (Luma-Osmani, Ismaili, & Ram Pal, 2021):

$$L_i = \ln \left( \frac{P_i}{1 - P_i} \right) = y_i = \alpha_i + \beta_1 X_{i1} + \dots + \beta_k X_{ik}$$

The below listing shows the libraries used in Python for discovering causality.

**Listing 5.5:** Python Libraries

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import math
%matplotlib inline
from sklearn.linear_model import LinearRegression, LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
from sklearn.cluster import KMeans
from sklearn.svm import SVC, SVR
from sklearn.neighbors import KNeighborsClassifier, KNeighborsRegressor
from sklearn.metrics import confusion_matrix
```

Our main data frame, which represents a csv file, is accessed through below path:

```
df = pd.read_csv(r"C:\Users\user\Dropbox\Casual Rules
\Shkurte\CybDataset.csv")
```

So, we first define the dependent variables as displayed in the listing 5.6:

**Listing 5.6:** Variable Definition

```
#Independent Variables
X = df[['Gender', 'Age', 'You_harassed_someone']]
# Dependent Variable
y = df['Victim_of_cyberstalking']
C = df[['Gender', 'Age', 'You_harassed_someone', 'Victim_of_cyberstalking']]
```

A variable named log will store the logistic regression function, and another function named train\_test\_split is the command that trains the model and tests it, and of course the last command fit in model is filling the model with data.

**Listing 5.7:** Logistic Regression

```
log = LogisticRegression()
X_train, X_test, y_train, y_test = train_test_split(X, y)
log.fit(X_train, y_train)
```

## 5.7 Model Evaluation using Confusion Matrix

The forecast accuracy of the model exceeds 61% of our forecasts. Practically we see how our forecast has predicted and what's the true value from our dataset (Luma-Osmani, Ismaili, & Ram Pal, 2021). Confusion\_matrix predicts the matrix by means of true as well as false positive, along with true as well as false negative, the values are diagonally from top to bottom. In the code below we have 12 and 10 true positive and true negative respectively. Therefore, the output in this case shall be (Luma-Osmani, Ismaili, & Ram Pal, 2021):

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + T_N + F_N} = \frac{12 + 10}{12 + 4 + 10 + 10} = \frac{22}{36} = 0.6111$$

**Listing 5.8:** Prediction and Confusion matrix

```
log.score(X_test, y_test) →
0.6111111111111112

y_pred = log.predict(X_test)
y_pred →

array ([1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1,
0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0], dtype=int64)

np.array(y_test) →

array ([1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 1,
0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0], dtype=int64)

confusion_matrix(y_test, y_pred) →

array([[12,  4],
       [10, 10]], dtype=int64)
```

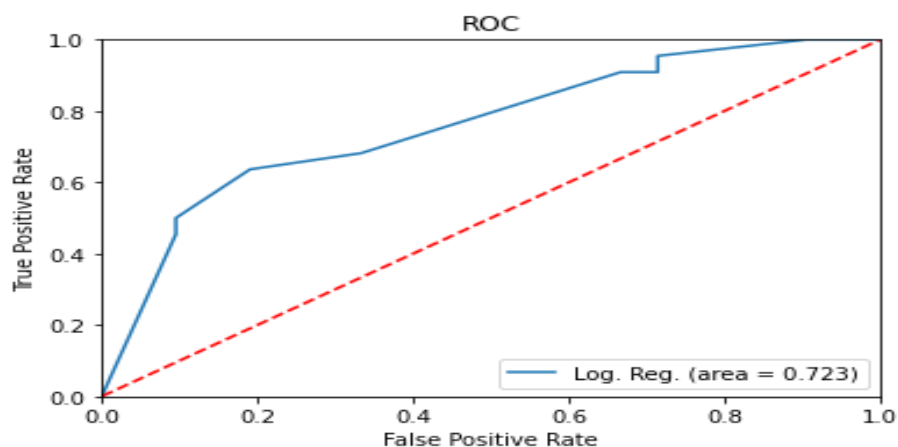
Aiming to plot the rate amid the true positive and false positive, the Receiver Operating Characteristic - ROC curve is plotted. The code for generating such plot is provided in listing 5.9, whereas the plot itself is presented in figure 5.12.

**Listing 5.9: ROC**

```
from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve
logit_roc_auc = roc_auc_score(y_test, logreg.predict(X_test))

r_fp, r_tp, thresholds = roc_curve(y_test, logreg.predict_proba(X_test)[:,1])
plt.figure()
plt.plot(r_fp, r_tp, label='Log. Reg. (area = %0.3f)' % logit_roc_auc)
plt.plot([0, 1], [0, 1], 'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.0])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC')
plt.legend(loc="lower right")
plt.show()
```

The red colored diagonal line divides the space of ROC plot. The points noted on the upper part of this line specify good classification results and the ones on the lower part identify bad outcomes. Therefore, as long as our outcomes rely away from this stroke, they can be measured as good outcomes (Luma-Osmani, Ismaili, & Ram Pal, 2021).



**Figure 5.12: Receiver operating characteristics**

**Listing 5.10: Model Coefficients**

```
regr = linear_model.LogisticRegression()
regr.fit(X, y)
print(regr.coef_) →

[[- 0.659772    0.551546    1.177470]]
```

The coefficients, the p-value and t-stat for the “Cyberstalking” dataset are calculated on table 5.7, as presented below.

**Table 5.7:** Coefficients and p-value of the main variables

Dependent Variable: VICTIM_OF_CYBERSTALKING		
142 observations		
	<b>coefficients</b>	<b>t-stat. (p-value)</b>
<b>C</b>	- 9.534084	- 2.302509 (0.0234)
<b>GENDER</b>	- 0.659772	- 1.834959 (0.0685)
<b>AGE</b>	0.551546	2.236363 (0.0274)
<b>YOU_HARASSED_SOMEONE</b>	1.177470	3.221538 (0.0014)

If we replace the results obtained in the equation, we gain the following (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022), (Luma-Osmani, Ismaili, & Ram Pal, 2021):

$$f(x_1, x_2, x_3) = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$
$$\widehat{Victim\_of\_cyberstalking} = -9.534 - 0.6598 (Gender) + 0.552 (Age) + 1.177 (You\_harassed\_someone)$$

**Listing 5.11: Logit Model**

```
import statsmodels.api as sm
model_logit=sm.Logit(y,X)
result=model_logit.fit()
print(result.summary2())
```



**Table 5.8: Logit Regression Results**

Results: Logit

Model:	Logit	Pseudo R-squared:	0.09820			
Dependent Variable:	Victim_of_cyberstalking	AIC:	188.5250			
Date:	2021-05-29 12:02	BIC:	197.3925			
No. Observations:	142	Log-Likelihood:	-91.262			
Df Model:	2	LL-Null:	-98.074			
Df Residuals:	139	LLR p-value:	0.0011005			
Converged:	1.0000	Scale:	1.0000			
No. Iterations:	5.0000					
-----						
	Coef.	Std.Err.	z	P> z	[0.025	0.975]
-----						
Gender	-0.6598	0.3536	-1.8785	0.0603	-1.3574	0.0288
Age	0.552	0.2500	2.2058	0.0274	0.0489	0.0186
You_harassed_someone	1.177	0.3688	3.1926	0.0014	0.3659	1.7669
C	-9.5340	4.2054	-2.2670	0.0234	0.3204	0.0288
=====						

Considering the way that we need to manage a twofold reliant variable, we picked the logit model as our objective model to check whether there is any critical factual coefficient of the free factors that could demonstrate to us that we have sufficient data to say that the casualties of cyberstalking could be anticipated in the event that we can handle some different factors. On the off chance that the quantitative variable has  $m$  categories, in regression we set  $m-1$  classifications (Luma-Osmani, Ismaili, & Ram Pal, 2021).

As per the table we get that the male gender is more averse to be casualties of cyberstalking, since the indication of the coefficient is negative and we made the assessment with maximum likelihood. To add that since the code for gender male is 1 and for female is 0, subsequently the end is that young men are more averse to be casualties, *ceteris paribus*. The gender coefficient is statistically significant for  $\alpha=10\%$  ( $p<0.07$ ) (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022). Age has a statistically significant coefficient at 5% ( $p<0.05$ ), except that the sign of the coefficient informs us that, taking into account the age group of the survey, the older you are, the more likely you are to be a victim of cyberstalking, *ceteris paribus* (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

The correct answer yes to the question of whether you have irritated someone appears to have an intriguing and significant consequence. Individuals who have irritated someone are sure to have been badgering; this means they are bound to be victims of cyberstalking. This variable's coefficient is also rather significant, at 0.05, (Luma-Osmani, Ismaili, & Ram Pal, 2021).

The model's quasi-R square coefficient is about 0.0984, which implies that using the supplied variables, we were only able to predict 9.84 percent of the variability in the dependent variable. To be more explicit, 9.84 percent of the variation in cyberstalking victims is explained by differences in the variables such as gender, age, and whether or not you harassed someone (Luma-Osmani, Ismaili, & Ram Pal, 2021). The model is useful for forecasting; however, it is clear that there are many additional characteristics that might help us anticipate the problem of cyberstalking in the future, but such variables are not included in this model. As a result, this model may be classified as exogenous (Luma-Osmani, Ismaili, & Ram Pal, 2021). This indicates that it is also affected by other variables that were not included in the study (Luma-Osmani, Ismaili, & Ram Pal, 2021).

Meanwhile we cannot understand the logit model coefficients, we may instead interpret the marginal effects of the logit model coefficients. The marginal effect merely demonstrates how many units the dependent variable will change if the impartial variable is changed by one unit. The marginal effect of the coefficients in the preceding model is as follows (Luma-Osmani, Ismaili, & Ram Pal, 2021):

**Table 5.9:** Marginal effect of the main variables

<b>Variable</b>	<b>Marginal effect</b>
<b>Gender</b>	-0.1444
<b>Age</b>	0.114
<b>You harassed someone</b>	0.2686

Males are 14.4 percent less likely to be victims of cyberstalking. Adolescents who are a year older are 11.4 percent more likely to be a victim of cyberstalking, and if you have previously harassed someone, you are 26.86 percent more likely to be a victim of cyberstalking. As there is practically no difference between OLS regression coefficients and marginal effect logit models, all marginal

effects are estimated using OLS (Gujarati & Porter, 2009), (Luma-Osmani, Ismaili, & Ram Pal, 2021).

## 5.8 Proposed Algorithm

When considering the causality, several theoretical studies are developed aiming the exposure of cause-and-effect rules. Those published algorithms can be arranged into three approaches: modifying the already existing algorithms, proposing a new algorithm, or hybridizing numerous ones (Abualigah, Diabat, Mirjalili, Abd Elaziz, & Gandomih, 2021).

Let us firstly define all the requested equations in order to obtain the joint entropy for two random variables A and B.

The joint probability and conditional probability formulas are related through the below equation:

$$\begin{aligned} p(A, B) &= p(A | B) * p(B) \\ &= p(B | A) * p(A) \end{aligned}$$

The general form of the entropy denoted by H, can be written:

$$H = \sum_{i=1}^m p_i * \log \frac{1}{p_i} = - \sum_{i=1}^m p_i * \log p_i$$

As a result, the joint entropy represents the entropy of the joint probability distribution, therefore,

$$H(A, B) = \sum_{i=1}^m \sum_{j=1}^n p(A_i, B_j) * \log \frac{1}{p(A_i, B_j)}$$

At the other hand, the conditional entropy presents the information amount that is requested to describe the dependent output variable B given that the value of another independent variable A is already known.

$$H(B | A) = \sum_{i=1}^m \sum_{j=1}^n p(A_i, B_j) * \log \frac{1}{p(B_i | A_j)} = p(A_i B_j) * \log \frac{p(A_j)}{p(A_i, B_j)}$$

If this formula is needed to be rewritten on another form, then it would be:

$$H(B | A) = - \sum_{i,j}^{m,n} p(A_i, B_j) * \log \frac{p(A_i, B_j)}{p(A_j)}$$

The algorithm we are presenting on this dissertation is conditional entropy based, hence the name COJEC (Conditional Joint Entropy based Causal rule discovery).

The multivariate input ( $X_m$ ) and output ( $Y_n$ ) variables on the algorithm are noted as follows:

- $X = \{X_1, X_2, \dots, X_m\}$
- $Y = \{Y_1, Y_2, \dots, Y_n\}$

---

**ALGORITHM 1: COJEC - Conditional Joint Entropy based Causal rule discovery**

---

INPUT:  $F, \tau$

$F$ : Feature's set of  $f$  discrete features  $F = \{F_1, F_2, \dots, F_f\}$ , where  $F_i = \{V_{1,F_i}, V_{2,F_i}, \dots, V_{|F_i|,F_i}\}$ .  $F_i$ , st.  $1 \leq i \leq f$

$\tau$ : a small threshold value

---

OUTPUT:  $W$

$W$ : the set of casual rules  $X \rightarrow Y$  for which the conditional joint entropy is smaller than the threshold  $\tau$ , i.e.  $H(Y|X) \leq \tau$

---

**Phase I – User:**

```

1 for  $m = 1$  to  $f - 1$ 
2    $F^{(m)} \leftarrow \text{allSubsetsOf } F(F, m)$ 
3   foreach subset  $F_S^{(m)}$  of  $F^{(m)}$ 
4      $F_C^{(m)} \leftarrow \text{cartesianProductOfSubsetOfFeatures}(F_S^{(m)})$ 
5     foreach  $X = \{X_1, X_2, \dots, X_m\} \in F_C^{(m)}$ 
6       for  $n = 1$  to  $f - 1$ 
7          $F^{(n)} \leftarrow \text{allSubsetsOf } F(F, n)$ 
8         foreach subset  $F_S^{(n)}$  of  $F^{(n)}$ 
9            $F_C^{(n)} \leftarrow \text{cartesianProductOfSubsetOfFeatures}(F_S^{(n)})$ 
10          foreach  $Y = \{Y_1, Y_2, \dots, Y_n\} \in F_C^{(n)}$ 
11            if  $X \cap Y = \emptyset$  and  $H(Y|X) \leq \tau$ 
12              add the casual rule  $X \rightarrow Y$  to  $W$  //  $X = \{X_1, X_2, \dots, X_m\} \rightarrow Y = \{Y_1, Y_2, \dots, Y_n\}$ 

```

**13 return  $W$**  //the set of casual rules  $X \rightarrow Y$  for which the conditional joint entropy is smaller than the threshold  $\tau$ , eq. (1)

---

**Figure 5.13: COJEC Algorithm Pseudocode**

In this algorithm, we have two inputs:

- **F** - which represents a set of  $f$  features, where each of the  $F_i$  properties can take certain values. In our case, the input data will be a dataset in csv format, which will consist of some features such as  $F = \{\text{Gender, age, school ...}\}$  and each of these elements will have certain values as  $F_{\text{age}} = \{14,15,16,17,18\}$ , and
- $\tau$  - a threshold, user-defined threshold, which will have a very small value.

as well as an output variable,

- **W** where will be stored the causal rules of the format  $X$  causes  $Y$ .

Now, we pass this whole set  $F$ , in order to find all the sub-sets of this set (given the fact that for a set with  $n$  elements, the number of sub-sets that can be obtained is  $2^n - 1$ ), and thus we form the set  $F_m$ , at the moment we find them, we calculate their Cartesian output, so in a way we multiply every element of a subset, with every element of another subset.

Within the  $F_m$  set, an independent variable  $X$  is declared which could take several values, i.e., be multi-valued, and will be a subset of the Cartesian output set.

In the same way, again we pass the set  $F$  and form another subset  $F_n$ , for which we also find all its possible subset, and the moment we find them, we again calculate their Cartesian output and store these products in a dependent variable  $Y$ , which will also be multi-valued, which is also the novelty of this algorithm, since most of the papers we have researched are based on outputs that can only take on a single value.

Here we set two conditions, as follows:

1. The independent-cause variables, and the dependent-consequence variables, should be disjoint sets, i.e. they should not have any common element and the elements that are in  $X$  should not be presented in  $Y$  and vice versa. This comes as a result of:  $\{\text{Gender, Age}\}$  cannot cause  $\{\text{Gender}\}$ .
2. And the second condition, the earned Entropy value which is calculated by the following formula, to not exceed this user-defined limits.

$$\begin{aligned}
 H(Y|X) &= P(X, Y) \log \frac{P(X, Y)}{P(X)} \\
 &= P(X = \{X_1, X_2, \dots, X_m\}, Y = \{Y_1, Y_2, \dots, Y_n\}) \log \frac{P(X = \{X_1, X_2, \dots, X_m\}, Y = \{Y_1, Y_2, \dots, Y_n\})}{P(X = \{X_1, X_2, \dots, X_m\})} \quad (1)
 \end{aligned}$$

Once these conditions are met, we can emphasize that we have enough information to say that the X variables cause the Y variables.

The output, as presented in figure 5.14 contains the obtained causal rules from the Cyberstalking dataset.

```
NOW FINDING THE CASUAL RULES
-----
HYX=0.0774648
inputsMatchOnly=11
outputsAndInputsMatchy=11

Feature 1=SCHOOL, value=1
Feature 2=CYBERSTALKING_ACHIEVING_GOAL, value=1
=> Feature=YOU_HARASSED_SOMEONE, value=1

-----
HYX=0.0730465
inputsMatchOnly=11
outputsAndInputsMatchy=10

Feature 1=SCHOOL, value=1
Feature 2=CYBERSTALKING_ACHIEVING_GOAL, value=1
=> Feature=CYBERSTALKING_PLEASURE, value=1

-----
HYX=0.0756188
inputsMatchOnly=12
outputsAndInputsMatchy=10

Feature 1=SCHOOL, value=1
Feature 2=CYBERSTALKING_PLEASURE, value=1
=> Feature=CYBERSTALKING_ACHIEVING_GOAL, value=1
-----
```

**Figure 5.14:** Cyberstalking causal rules

## 5.9 Chapter Discussion

Overall, logistic regression helps in how to build respective models. The same applied to this chapter as well. Through logistic regression a causal model that analyzing our research data as well as the results and research findings for defining Cyberstalking causes has been build (Luma-Osmani, Ismaili, & Ram Pal, 2021).

Three high schools in Tetovo, Republic of North Macedonia have been surveyed and the data has been analyzed in Python programming language. The participants include 48.6% male and 51.4% female aged from 14 to 18 years old. No strong correlation coefficients among variables have been noticed (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022). Several studies claim that there is a strong link between probability and causality. Association rules were firstly generated through Apriori algorithm and a ratio between three independent variables: “Gender”, “Age”, “You\_harassed\_someone” and the dependent variable “Victim of Cyberstalking” has been considered (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022). If you are a woman according to this scheme the chances are 54.8% approximately to be harassed online. Boys are more likely not to be harassed, even with 62.3% approximately, i.e., the probability of being bullied and being boy is 37.7% respectively (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022). Lastly, the likelihood of being cyberstalked increases, if you have previously harassed someone. The same consequence comes with increasing the age of the respondents (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022). All the coefficients are statistically significant for  $\alpha=5\%$  ( $p<0.05$ ) except for the gender coefficient which is statistically significant for  $\alpha=10\%$  ( $p<0.07$ ) (Luma-Osmani, Ismaili, Pathak, & Zenuni, 2022).

The Confusion Matrix was used to evaluate the model, and the same shows accuracy of 61.11% in the predictions. Anyway, as an exogenous model, the results lead to the fact that other causes should be considered in predicting the outcome (Luma-Osmani, Ismaili, & Ram Pal, 2021). In the last sub chapter, the joint entropy algorithm, COJEC was presented. It helps in discovering causal rules from the Cyberstalking dataset.

# ETHICAL ISSUES IN PUBLICLY AVAILABLE DATA

---

### 6.1 Ethics in Data Science

Because disclosure is based on considerations that are directed under investigations, ethics is a significant problem in this domain. The Tuskegee Experiment and the Willowbrook Study are two studies that included humans who violated any moral or ethical norm (Rothman, 1982) (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

- Tuskegee Experiment (1932-1972): Aiming the examination the long-standing effects of the illness, American researchers purposefully delayed treatment for 399 African-American people infected with syphilis. Even when a penicillin treatment was available in Tuskegee, Alabama, individuals had purposefully left to suffer from syphilis (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).
- Willowbrook Study (1963-1966): Hepatitis was purposefully instilled in youngsters with cognitive impairments. The study's goal was to look into the disease's progression and to evaluate a possible vaccine (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).



In such a case, data ethics is a field that should be studied. This chapter aims to offer a clear outline of the key characteristics of the ethical problems connected to public data on a bibliometric study. As a result, 1483 articles from the IEEE digital library were created between 1970 and 2020. It moreover shows the research partnership among co-authors, keywords, titles, and the H-index at the country level, and crucial conclusions are reached. VOSviewer software was used to visualize the data, and PyCharm in the Python programming language was used to generate the charts (Luma-Osmani, Ismaili, & Raufi, 2020).

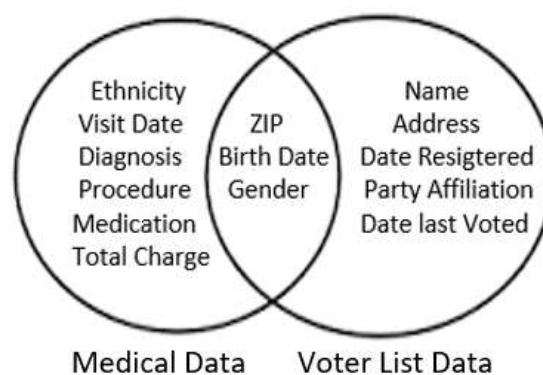
The digital age we live in is drastically altering the course of our lives. Yet, the priority of academics and the academic community is sometimes more on the reality of collecting knowledge, without considering how they do so, even though facing different ethical problems may be the price. According to conventional projections, 2.5 quintillion bytes ( $10^{18}$ ) of data are produced every single day (Mori, 2016), (Luma-Osmani, Ismaili, & Raufi, 2020).

As a concept, publicly-available data refers to information that is easily accessible, generally over the Internet, and may be obtained quickly and for free (Cooper & Coetzee, 2020). Data ethics is a new field of ethics that examines and evaluates moral concerns related to data with the goal of developing and promoting ethically sound solutions (Floridi & Taddeo, 2016). Ethical issues about data tend to be more exciting, compared to ethical concerns regarding other breakthrough technologies. It happens, since partly data along with data science are presented pervasively, therefore potentially might affect many fragments of society, and partly because of their inherent complexity (Hand, 2018), (Luma-Osmani, Ismaili, & Raufi, 2020).

A bibliometric study is commonly used to determine the evolution of a discipline situation and its current condition. Allen Richard, a prominent British researcher and scientist, created the term "bibliometrics" in 1969 as a substitute for "statistical bibliography." As a result, it represents the formal start of bibliometrics (Liao, et al., 2018). It is also distinguished as quantitative analysis, including mathematical and statistical approaches of publications (Guo, Huang, Guo, Li, & Guo, 2019).

## 6.2 Publicly available data ethics

All this began with Latanya Sweeney's low-tech experiment at the end of the 1990s. That alone merged the free medical records and a list of voters by merely adding three attributes {5-digit ZIP code, Birth Date, Gender}, as was shown in Figure 6.1. The target was Governor William Weld of Massachusetts. As a result, 87 percent of the confidential details of the anonymized patient became revealed (Barth-Jones, 2012) (Sweeney, 2000). Sweeney discovered that, with the birth date alone, 12 percent of the voting population can be remodeled. With the date of birth and gender, the number increases to 29% and with the date of birth and indeed the postal codes, increases to 69% (Privacy Lives, 2013).



**Figure 6.1:** Linking anonymized data

Sweeney post-identified more than 42 percent of the unidentified participants in a DNA survey in 2013, lecturing as a scholar for Harvard (Sweeney, Abu, & Winn, 2013) (Tanner, 2013). Likewise, (Narayanan & Shmatikov, 2008) using the Netflix Prize data collection featuring 500,000 subscribers' anonymity film reviews, and IMDb movie repository available to the public, thus some anonymous users have been detected and confidential information relating to customers, such as the political affiliation, collected. She stresses that face recognition software is a major threat to privacy (Newton, Sweeney, & Malin, 2003).

Study in Chicago (Ochoa, Rasmussen, Robson, & Salib, 2001), 35 percent of suicide victims were found by comparing the Social Security Death Index and the city data collection. The survey covered a time span of three decades. As shown in the preceding cases, linking or combining data

sets with asymmetric encryption records with other information sets, some of which are freely available, can facilitate the process of cross-identifying individuals and collecting sensitive data (ACM, 2018).

Although the amount of information acquired grows rapidly, ethical concerns about privacy and protection are increasingly turning into a topic of discussion among scientists around the world. Such a conundrum has been left to data mining (Ruggiero, Pedreschi, & Turini, 2010), (Agrawal & Srikant, Privacy-Preserving Data Mining, 2000), (Pedreschi, Ruggieri, & Turini, 2008), (Clifton, Kantarcioglu, & Vaidya, 2002), healthcare (Humphreys, 2013) and social media (Maddock, Mason, & Starbird) (Wheeler, 2018), (Townsend & Wallace, 2016), (Luma-Osmani, Ismaili, & Raufi, 2020).

The versatile information utilized to acquire driving and travel patterns are freely accessible (Cooper & Coetzee, 2020). Doubtlessly, there are moral and ethical issues with the utilization of such information, particularly data protection. Onlookers (Pedreschi, Ruggieri, & Turini, 2008) define discrimination as the consequence of several mining rules such as association rules and categorization rules. Information Mining for Discrimination Discovery has likewise been tended to by (Ruggiero, Pedreschi, & Turini, 2010) which handles such an issue by utilizing repositories derived from past judgment records produced by people or software. It being stated, it ought to be feasible to grasp privacy protection in any data mining strategy (Clifton, Kantarcioglu, & Vaidya, 2002) (Agrawal & Srikant, 2000).

(Thomas, Pastrana, Hutchings, Clayton, & Beresford, 2017) have highlighted ethical issues in research employing reports from questionable datasets, such as illegal sharing of hacked databases or sensitive information. Ethical concerns arising from the use of public data in research initiatives might inadvertently reveal human activities, behaviors, and relationships, necessitating the development of new forms of ethical support (Boyd, Keller, & Tijerina, 2016), (Luma-Osmani, Ismaili, & Raufi, 2020). Again, (Mori, 2016) lists all the objectives for data scientists, it is clear and important that participants feel that responsibility, protection, and safeguard should be addressed at all stages of a data research study.

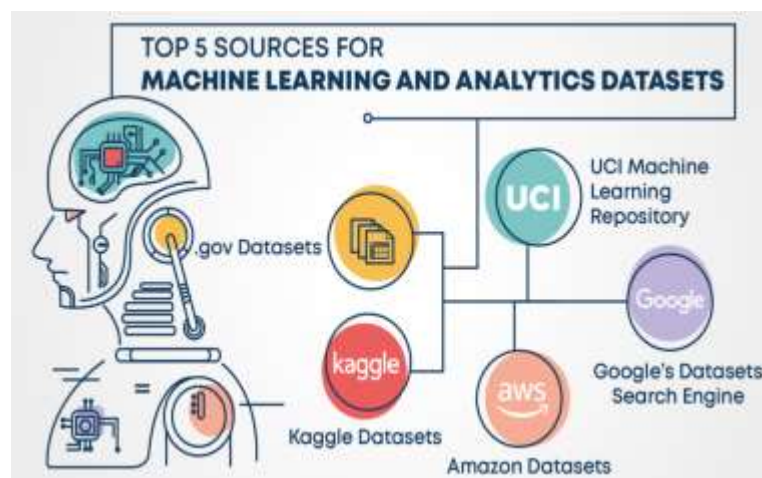
Inevitably, plenty of the ethical problems relate to information on individuals (Hand, 2018) and health care, marked as one of the central areas where data integrity is gaining in popularity. Although all healthcare professionals should adhere to the Helsinki Declaration (2013) established by the World Medical Association (WMA) as an assertion of ethical ideologies to be considered, when a medicinal study includes human partakers. As per the scientific community, numerous patients will probably be able to agree to the use of their records for study, but not to let them fully understand about these issues is wrong (Humphreys, 2013). The vast array of sources, the widely diverse nature of these data and the various reasons for their collection, both which challenge the public health community in ethical mining (Vayena & Madof, 2019). (Dorey, Baumann, & Andorno, 2018) aimed to achieve an in-depth awareness of the possible ethical risks and related responsibilities of those involved in projects using patient records. In addition, academics have to determine if the open data policy only applies to the final outputs (Leetaru, 2017). (Busse, Kernecker, & Siebert, 2020) proved that sharing survey information helps explain the responses of consumers to ethical questions.

Data security is both a primary concern for European research ethics and a fundamental human right. Established researchers should apply information security thoroughly (EU, 2018). Surely, the major electronic libraries such as IEEE (Vallor & Rewak, 2018) and ACM (2018) have fostered a morals code that can rouse and coordinate every single proficient professional computer (Luma-Osmani, Ismaili, & Raufi, 2020). In emerging nations, they also focus on new ethical principles. In the context of how it is used, researchers need to evaluate the evidence on which their results are based. NASA Metrics Database Program data sets were also extensively used in experiments with software defect prediction, but such data sets would need substantial pre-processing, so that they can be properly prepared for defect prediction (Gray, Bowes, Davey, Sun, & Christianson, 2011). (Hand, 2018) describes data as fresh coal or oil that has been refined to obtain energy in the same manner.

To adjust the morals of the outcomes, the exploration local area should decide how to figure forward-looking strategies, moral procedures and innovation (Barth-Jones, 2012), (Fiesler, 2019). We need lawful just as innovative systems to work with and manage the trading of these datasets while guaranteeing that people's security is regarded (Stiles & Boothroyd, 2015), (Gupta, 2018) (OECD, 2016) to ensure clients ' protection. In combination with (Parker, 2015) the development of a code of expert morality is required, bringing the full circle in terms of both damage and profit to individuals involved in data sharing. It is crucial because the measurements for Internet research cannot be standardized, thus the use of innovation and innovation develops on a continuous basis (Townsend & Wallace, 2016), (Luma-Osmani, Ismaili, & Raufi, 2020).

Affording to 2019 survey (Great Learning), top five 5 digital data warehouses include:

- I. Google's Datasets
- II. UCI ML Repository
- III. .gov Datasets
- IV. Kaggle Datasets
- V. Amazon Datasets



**Figure 6.2:** Mostly used datasets



Google Dataset Search<sup>22</sup> which was firstly launched in September 2018, but on January 24, 2020 is officially out of beta version, with nearly 25 million datasets included (Luma-Osmani, Ismaili, & Raufi, 2020). The company has added new features to Dataset Search based on feedback received from its users (Southern, 2020). A great feature is that it displays various sets of data based on the keywords from numerous repositories, involving government web pages, Kaggle, etc.

UCI ML Repository<sup>23</sup> as the oldest data source on the internet, represents a great initial stop, especially when observing for motivating



datasets. A number of databases consists of 488 (Luma-Osmani, Ismaili, & Raufi, 2020), utilized widely from research community, as well as data miners and ML society. It was initially created by David Aha in 1987 (UCI Machine Learning Repository, 1987). No registration is compulsory when downloading data from this repository, anyway a proper citation is required. The datasets are divided based on the default task they perform, attribute, format and data type, area and number of attributes & instances.



.gov Datasets are typically used in developed countries which are becoming superpowers in artificial intelligence domain. The rules and regulations correlated to these datasets containing government, federal, state or city data (Luma-Osmani, Ismaili, & Raufi, 2020), present a strict attention since are real data gathered from many segments of a nation. Hence, is recommended a cautious use of it (Great Learning, 2019).

Kaggle Datasets are known for hosting challenges related to deep and machine learning. As an electronic repository, Kaggle beside providing datasets, simultaneously afford a forum for the machine learning community (Luma-Osmani, Ismaili, & Raufi, 2020), (Great Learning, 2019).



---

<sup>22</sup> <https://datasetsearch.research.google.com/>

<sup>23</sup> <http://archive.ics.uci.edu/ml/index.php>



Amazon has listed some of the data repositories and databases available for open data. It can be publicly retrieved on their servers. Consequently, when utilizing AWS resources for modeling, these locally accessible sets will fasten the process of data loading. A characteristic is that it contains numerous datasets classified on the basis of application area, such as: ecological resources, satellite imageries and similar (Great Learning, 2019). Cloud data sharing allows the user to focus more data analysis afore data acquisition (Amazon AWS, 2020), (Luma-Osmani, Ismaili, & Raufi, 2020).

### 6.3 Methodology

This review focused on writing data recovered from the IEEE Xplore advanced library (IEEE Xplore, n.d.). In terms of the researcher community, it is a valuable resource for further in-depth discovery of technical and scientific information issued by IEEE and its partners (Luma-Osmani, Ismaili, & Raufi, 2020).

The primary phrase we are seeking for is "Data Ethics," and the time period chosen is the most recent fifty years, i.e., period 1970-2020, consisting of "all types" of writing that surfaced after aiming for virtually the same referred term. The very last inquiry culminates in seven different types of reports, totaling 1483 distributions. Conference Proceedings (1167) were the most frequently presented document type, accounting for 78.69% of all publications. Magazines (n=134, 9.04%) and Journals (n=121, 8.16%) follow. Books (37), Courses (8), Early Access Articles (8) and Standards (8) are among the other works with fewer than a hundred pages. Table 6.1 displays the data given previously. All the papers in .RIS format were recovered on August 4, 2020 (Luma-Osmani, Ismaili, & Raufi, 2020).

**Table 6.1:** Retrieved documents types

Type of Document	Frequency	Proportion
Conferences	1167	78.69
Magazines	134	9.04
Journals	121	8.16
Books	37	2.49

Courses	8	0.54
Early Access Articles	8	0.54
Standards	8	0.54
<b>Total</b>	<b>1483</b>	<b>100</b>

For bibliometric analysis, a variety of applications is utilized. VOSviewer (Eck & Waltman, 2020) a freely accessed software developed by Waltman and Eck in Netherlands, was used for map generation, and PyCharm (PyCharm, n.d.), developed by Czech company JetBrains, in Python programming language, was used for graph plotting and representations. Since the data was recovered from the IEEE Xplore library, the organizations created include the creators and watchwords examination, which was accomplished by creating a guide based on bibliographic data, as well as the title investigation, which was accomplished by creating a guide based on text data (Luma-Osmani, Ismaili, & Raufi, 2020).

The steps of bibliometric examination of ethical problems linked to public datasets were based on research conducted by (Guo, Huang, Guo, Li, & Guo, 2019).

**Table 6.2:** Stages of bibliometric analysis on data ethics research

<b>Stage 1:</b> Data Collection	<b>Retrieval Database:</b> IEEE Xplore
	<b>Retrieval Mode:</b> Advanced Search
	<b>Boolean Model:</b> (Ethics     Ethical Issues     Ethical Concerns) && [(Public     Free) && (Data Set     Dataset     Data-set)]
	<b>Time Span:</b> 1970 - 2020
	<b>Doc. No.:</b> 1483
<b>Stage 2:</b> Bibliometric Analysis and Information Visualization	Publications Number
	Citations Sum
	H-Index Calculations
	Co-Authors Analysis
	Keywords co-occurrence
<b>Stage 3:</b> Conclusions, Remarks, Limitations & Future work	Title terms co-occurrence
	Conclusions related to the latest trends in data science in general and data ethics in particular.
	Limitations of the study will be addressed and possible paths for future work.



## 6.4 Interpretation of Results

We attempt to represent the results of the bibliometric study in detail in the following subchapter, covering the key aspects of the investigation. As a result, subsection 6.4.1 discusses the present state of Data Ethics papers, while part 6.4.2 introduces the authors, keywords, and title analysis (Luma-Osmani, Ismaili, & Raufi, 2020).

### 6.4.1 The current status of the Data Ethics publications

Current yearly patterns of publications linked to data ethics will be covered in this section. In addition, a deeper look at the discussion's spread across nations, as well as their citation and H-index score.

Figure 6.3 depicts the yearly publications published between 1970 and 2020. In any case, the graph line is correctly aligned and begins in 1972, the year of the first publication. It demonstrates that a relatively modest rise was observed in the first twenty years (1970-1990). In the next decade (1990-2000), there was a minor growth, and after 2000, there was a considerable rise, and more academic researchers commenced to explore in this domain. The rapid blossoming of internet technology, as well as the rise of ethical concerns about data privacy and security, may be attributed to this. It is important to emphasize that a drop is presented in 2020 papers, but anyway, because the study was conducted in the third quarter of the year, yet there is no clear breakdown of journals published during this period (Luma-Osmani, Ismaili, & Raufi, 2020).

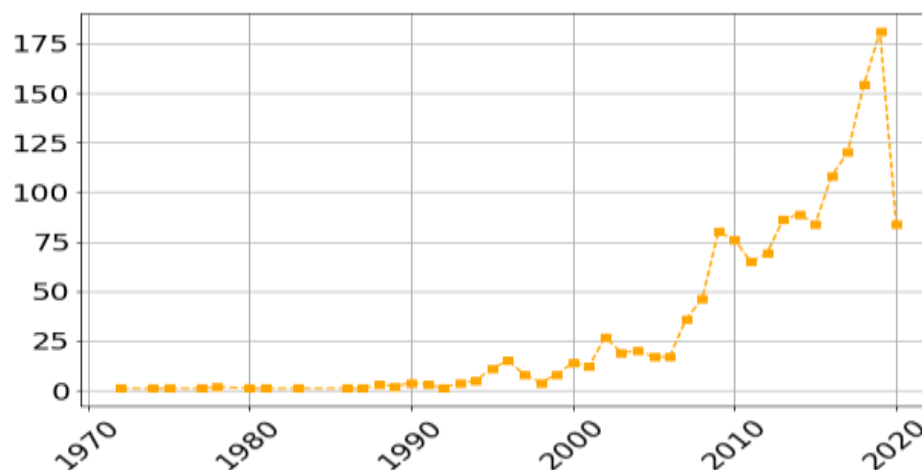


Figure 6.3: Publications per year

In terms of nations, the United States has the most publications, with 398 papers, accounting for 26.84 percent of all publications. The second nation in terms of number of publications is China (128 or 8.63 percent), and the third is the United Kingdom, with 118 publications accounting for 7.96 percent (Luma-Osmani, Ismaili, & Raufi, 2020). Figure 6.4 depicts a comprehensive overview.

In terms of the sum of citations, the United States once again leads with 2356 (46.76 percent) citations, followed by the United Kingdom with 726 (14.41 percent) citations and Canada with 216 (4.29 percent) (Luma-Osmani, Ismaili, & Raufi, 2020).

The quality of publications seems to be another essential metric that indicates the growth patterns of scientific research. Thus, this criterion (publication quality) is derived from how frequently papers are cited as the source by others (Guo, Huang, Guo, Li, & Guo, 2019).

Now that we've defined the quantity and quality of data for the top ten nations, we can calculate the H-Index score of the articles. The H-Index indicates that a researcher must have at least H publications that have been referenced more than or equal to H times (Guo, Huang, Guo, Li, & Guo, 2019), (Luma-Osmani, Ismaili, & Raufi, 2020). It acts as a criterion for demonstrating a researcher's scientific accomplishment. As a result, as shown in Figure 6, the United States is the bellwether, with the highest H-index score of 23. The UK is ranked second (13) while Australia (7) and Canada (7) are ranked third and fourth, respectively. Furthermore, Germany, India, and Japan appear to be in the same range, with an H-factor value of 6. China and the United Arab Emirates are ranked eighth and ninth, correspondingly, with a H value of 5. Such suggests that, in comparison to those top-tier nations, the quality of publishing in China has to increase. Nonetheless, we must notify that there were 818 citations missing from the downloaded data. (Luma-Osmani, Ismaili, & Raufi, 2020).

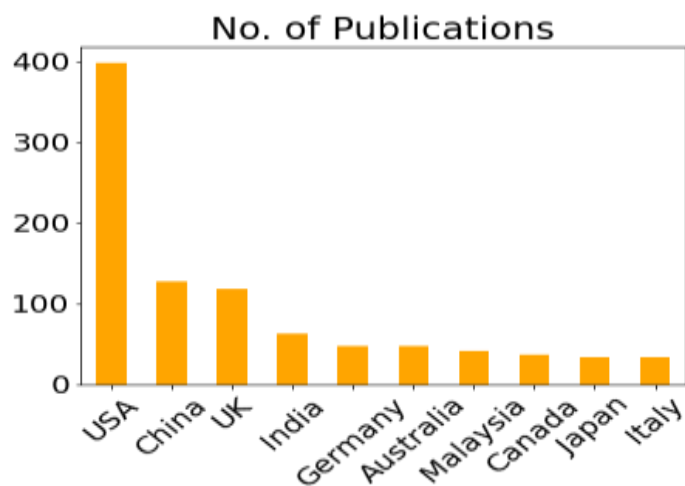


Figure 6.4: Number of Publications

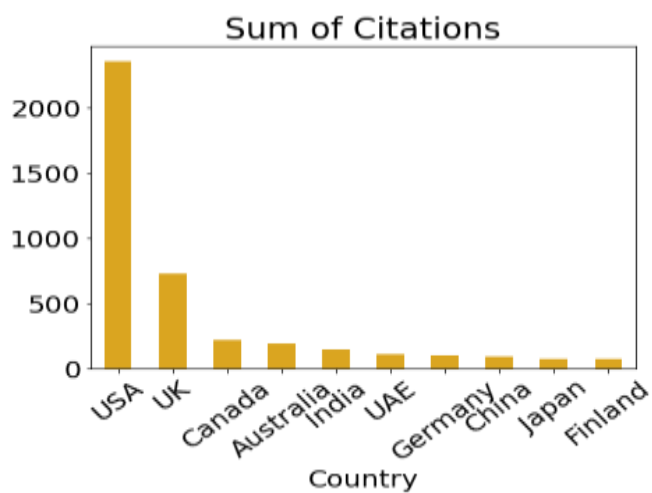


Figure 6.5: Sum of Citations

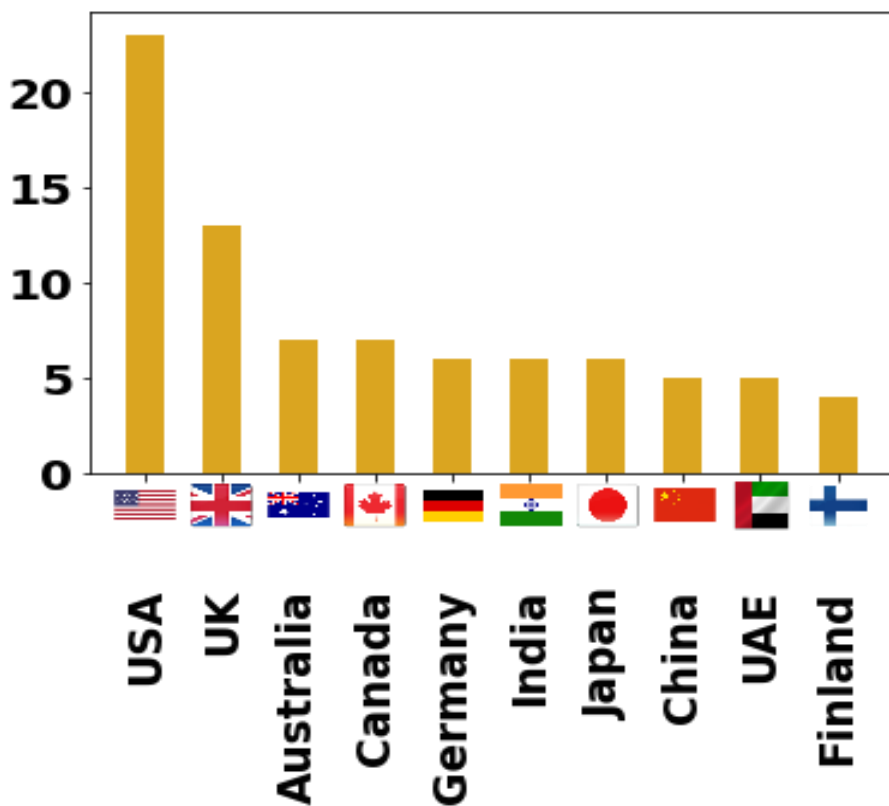


Figure 6.6: H-Index of Data Ethics Publications

### ***6.4.2 Co-authorship, Keywords and Title Analysis***

In the co-authorship analysis, the unit of analysis is the sum of 3831 authors. The above-mentioned indication looks like this when presented on a statistical scale: 7.60 percent ( $n=3831/291$ ) were scientific collaborators in two articles on data ethics, 1.64 percent ( $n=3831/63$ ) were credited in three, 0.52 percent ( $n=3831/20$ ) were co-authors in four papers, and just 0.29 percent ( $n=3831/11$ ) were collaborators in five or more publications. Throughout this study, it was required that the co-authors fulfill the two-paper criterion, which meant that at least one author had to appear in two articles. As a conclusion, 291 writers fulfill the requirements. Regrettably, if an author used two different names in publishing, they just cannot be combined, thus they appear on the map as two different authors (Luma-Osmani, Ismaili, & Raufi, 2020).

Because all of those writers are members of 143 collaboration clusters of various colors, the line connecting them identifies the collaboration ties between them. Like a consequence, the red cluster had the most members and was the most cooperative, with 14 authors including Y. Zhang, D. Zhang, Z. Han, L. Wang, Y. Gu, M. Pan, Z. Dawy, Z. Li, L. Liu, J. Xu, R. Smith, Z. Yang, Y. Yang and H. Zhang. Figure 6.8 shows the density map with the same cluster highlighted. At the same time, Y. Zhang is the author with the most documents, 9 in total, as well as the author with the highest link strength (17), as shown in table 6.3.

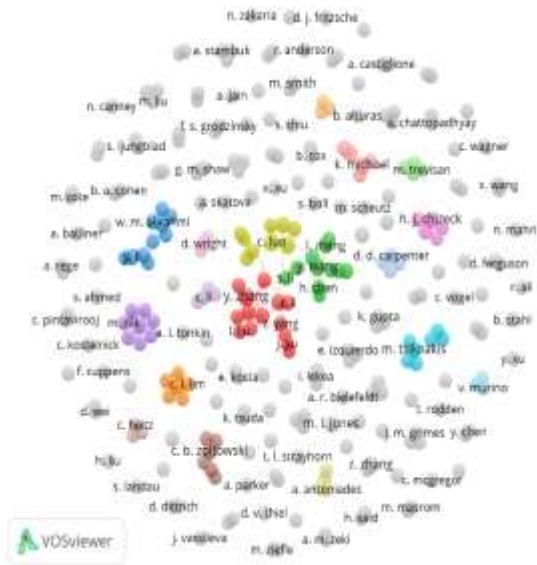


Figure 6.7: Co-authorship Analysis

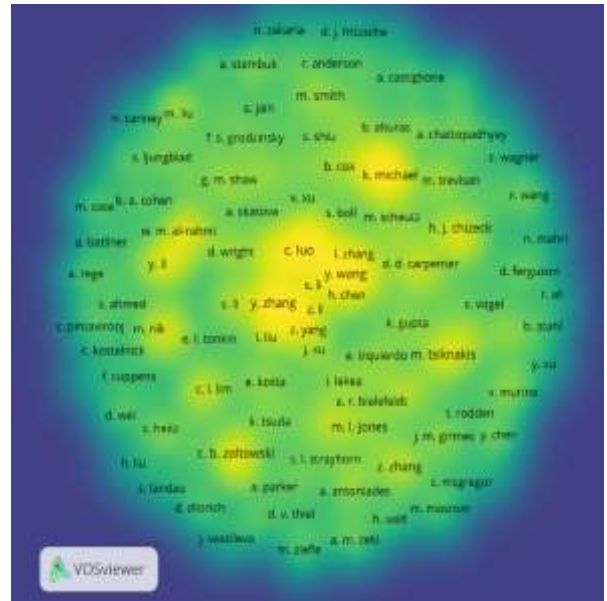


Figure 6.8: Author's density visualization map

Table 6.3: Top 10 Authors of data ethics publications

#	Author	Documents	Total Link Strength
1	Y. Zhang	9	17
2	K. Michael	8	5
3	Y. Wang	8	8
4	K. W. Miller	7	4
5	C. Luo	6	15
6	P. Li	6	17
7	Z. Zhang	6	1
8	C. B. Zoltowski	5	7
9	L. Wang	5	6
10	S. Salinas	5	14

In contrast, the writers with the most citations are those who worked alone. According to Table 6.4, the most cited one is Chengjun Liu from the New Jersey Institute of Technology in the United States, who has 363 citations. The second ranked is Ross Anderson from the Computer Laboratory at Cambridge University in the United Kingdom, who has 221 citations, and the third is Donovan Artz from the Los Alamos National Laboratory in New Mexico, USA. Before 2005, they had all made contributions to the field (Luma-Osmani, Ismaili, & Raufi, 2020).

**Table 6.4:** Top 10 Cited Papers of data ethics publications

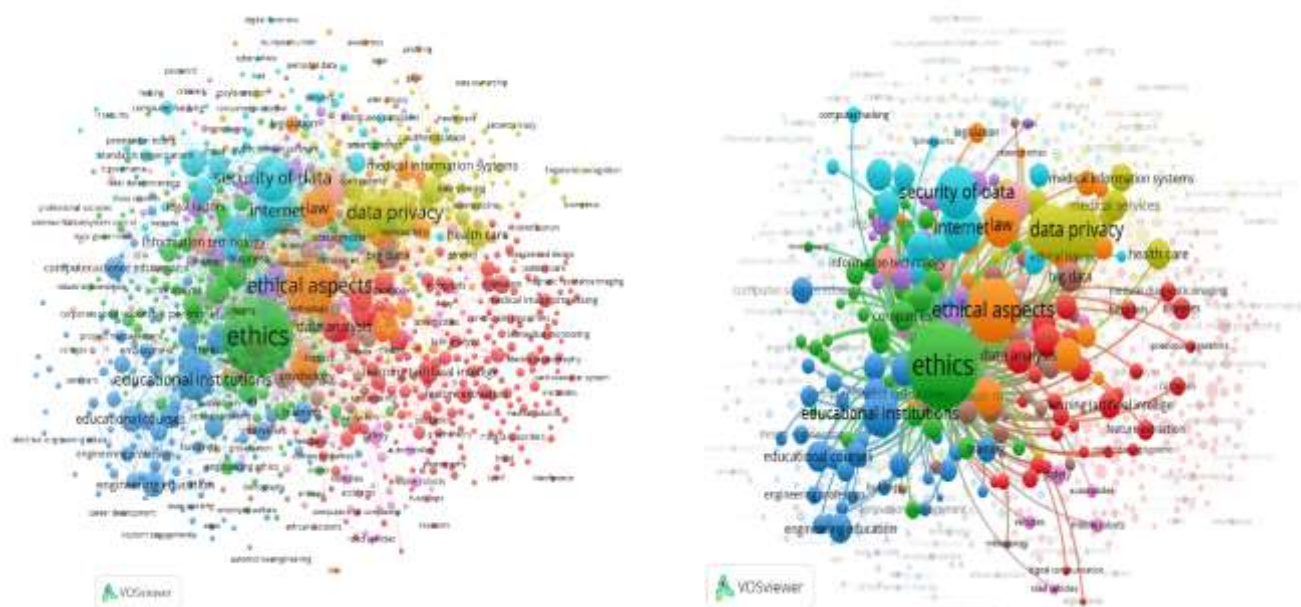
#	Document Title	Authors	Author Affiliations	Year	Cit.
1	Gabor-based kernel PCA with fractional power polynomial models for face recognition	Chengjun Liu	Dept. of Comput. Sci., New Jersey Inst. of Technol., Newark, NJ, USA	2004	363
2	Why information security is hard - an economic perspective	R. Anderson	Comput. Lab., Cambridge Univ., UK	2001	221
3	Digital steganography: hiding data within data	D. Artz	Los Alamos Nat. Lab., NM, USA	2001	207
4	A Review of Basic to Clinical Studies of Irreversible Electroporation Therapy	C. Jiang; R. V. Davalos; J. C. Bischof	Mechanical Engineering Department, Univ. of Minnesota, Minneapolis, MN, USA	2015	154
5	Big Data: New Opportunities and New Challenges	K. Michael; K. W. Miller	Univ. of Wollongong, Australia	2013	146
6	What Do Mobile App Users Complain About?	H. Khalid; E. Shihab; M. Nagappan; A. E. Hassan	Shopify; Concordia Univ., Canada	2015	144
7	Towards Achieving Data Security with the CC Adoption Framework	V. Chang; M. Ramachandran	School of Computing, Leeds Beckett Univ., UK	2016	128
8	UAVs for smart cities: Opportunities and challenges	F. Mohammed; A. Idries; N. Mohamed; J. Al-Jaroodi; I. Jawhar	College of Information Technology, UAE	2014	88
9	Technological learning, strategic flexibility, and new product development in the pharmaceutical industry	P. E. Bierly; A. K. Chakrabarti	Sch. of Bus. Admin., Monmouth Univ., West Long Branch, NJ, USA	1996	85
10	A security policy model for clinical information systems	R. J. Anderson	Comput. Lab., Cambridge Univ., UK	1996	84

This section of the research is dedicated to providing a comprehensive knowledge of Data Ethics by studying the prevalence of keyword co-occurrence. To demonstrate the research hotspot in the topic of data ethics, 15557 terms were extracted from 1483 articles. The threshold was set

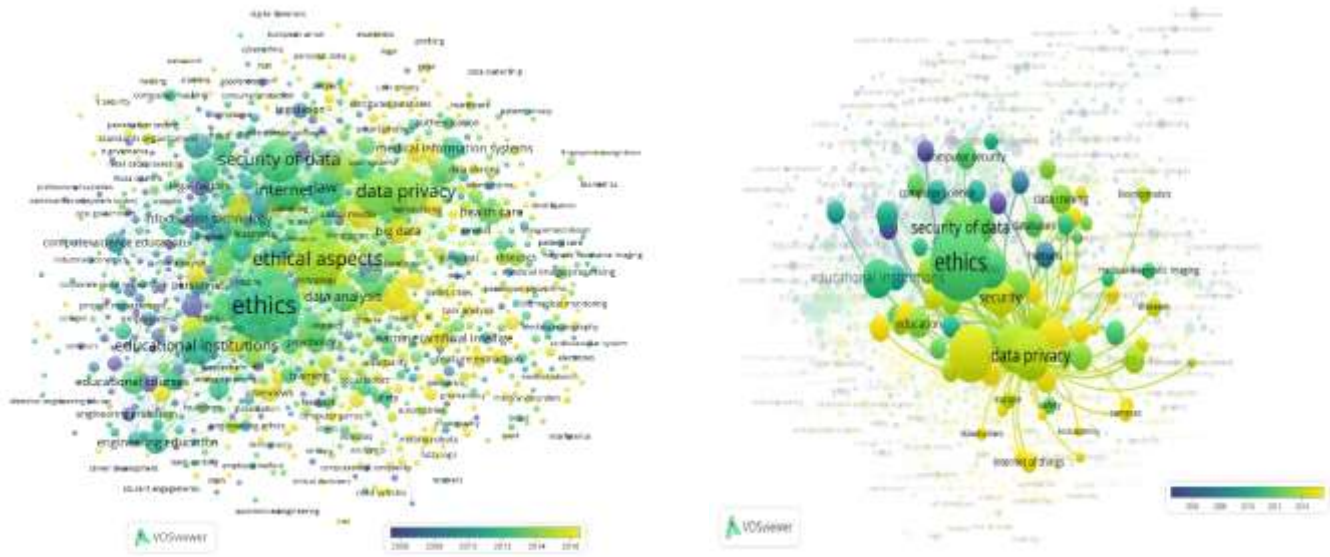
at five, therefore terms with less than five occurrences are eliminated, and so 799 keywords were included in the study, split into nine clearly defined clusters (Luma-Osmani, Ismaili, & Raufi, 2020).

Each diameter of the circle reflects the frequency of recurrence of a specific phrase. The bigger a circle, the more times the same term has been chosen. Of course, the most often used keywords “ethics” and “ethical aspects” occupied the greatest circles, with co-occurrences of 685 and 380, accordingly (Luma-Osmani, Ismaili, & Raufi, 2020). Table 6.5 shows the incidence but also the total strength connections of the top 10 nouns of data ethics articles.

Circles of the same hue indicate a related subject. The azure blue cluster contains keywords such as “internet”, “security of data” “computer hacking”, “IP networks”, “ethical issues” and appears to be more “Security” focused, whereas the orange cluster is focused in keywords such as “ethical aspects”, “legislation”, “law”, “research ethics”, “decision making”, targeting more “legal issues” and the green cluster comprises keywords like “ethics”, “government”, “companies”, “investment” are more associated with “business ethics” (Luma-Osmani, Ismaili, & Raufi, 2020).



**Figure 6.9:** Keyword co-occurrence network



**Figure 6.10:** Keyword co-occurrence timeline

(a) All keywords (b) Highest frequency keywords

(a) All keywords (b) Highest frequency keywords

Figure 6.10 shows the same keyword co-occurrence. To depict the years 2006 and 2016, terms are organized in variable phenomena in a time gradient color gradient of dark purple to yellow. It is apparent that from 2008 to 2012, the major focus was on ethics as a generic philosophical issue, but recently, the research hotspot focus has shifted to data privacy, security, and IoT (Luma-Osmani, Ismaili, & Raufi, 2020).

As shown in table 6.5, the top 10 keywords for data ethics articles are: ethics, ethical aspects, data privacy, privacy, security of data, internet, educational institutions, law, security and decision making (Luma-Osmani, Ismaili, & Raufi, 2020).

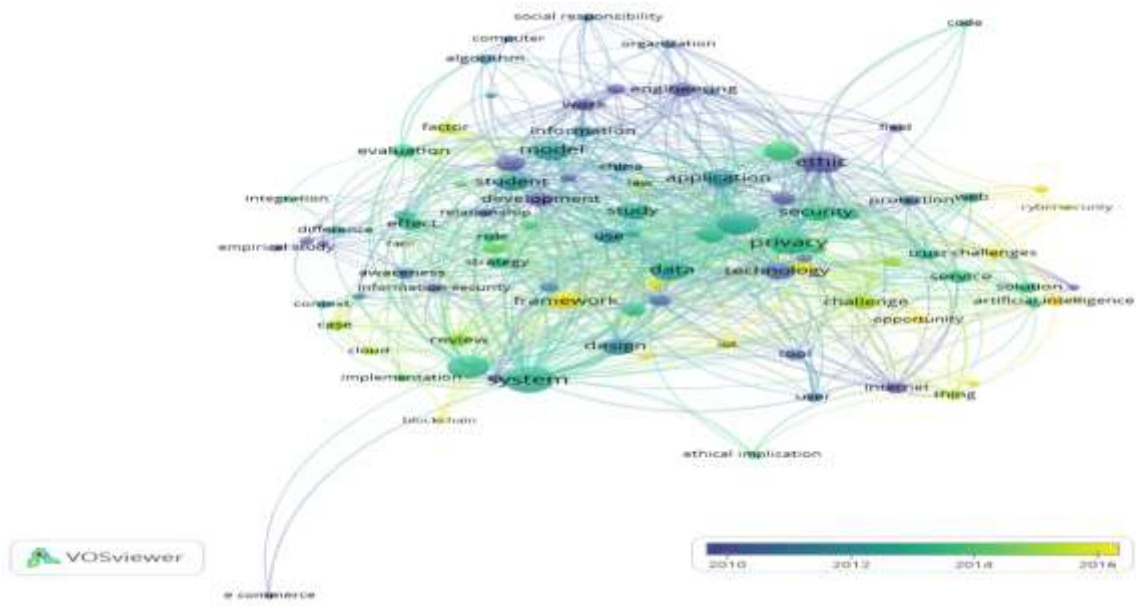


**Table 6.5:** Data ethics keywords

#	Keyword	Occurrences	Total Link Strength
1	ethics	685	6825
2	ethical aspects	380	4222
3	data privacy	299	3253
4	privacy	266	2896
5	security of data	248	2727
6	internet	207	2363
7	educational institutions	175	1820
8	law	169	1813
9	security	157	1766
10	decision making	117	1234

Lately, studies of textual data have been performed, using the word identification based on the publication's title.

Figure 6.11 shows how 3789 keywords were retrieved from the study, with 100 of them meeting the minimum five occurrences' criteria. Terms with a high relevance value are more likely to reflect specific themes covered by the text data, whereas those with a low relevance score are more likely to depict more generalized topics (Eck & Waltman, 2020). The bigger the tinted circle, the more frequently that word is used. Table 6 reveals that the phrase "system" had the most instances (99), followed by "privacy" (75), and "analysis" (69). "Student" gets the greatest relevance score among the top ten most often occurring keywords (Luma-Osmani, Ismaili, & Raufi, 2020).



**Figure 6.11:** Title terms timespan

**Table 6.6:** Top 10 title terms of data ethics publications

#	Term	Occurrences	Relevance Score
1	system	99	0.26
2	privacy	75	0.4219
3	analysis	69	0.3794
4	research	66	0.1887
5	ethic	65	0.2586
6	data	60	0.2385
7	model	57	0.3409
8	ethics	47	0.3103
9	student	44	0.5315
10	framework	43	0.277

## 6.5 Chapter Discussion

To summarize, ethical concerns are an inherent aspect of the digital world, and they have become a global concern in recent years. As a result, decision makers must be explicitly aware of these dangers and raise awareness of these concerns with their respective teams, even if the immediate impacts are not obvious. Notably, computer experts who create tools that facilitate this relationship are obliged to pay close attention, analyze these potential outcomes, and take steps to mitigate any possible harm (Luma-Osmani, Ismaili, & Raufi, 2020).

Co-authorship, co-occurrence of keywords, and co-occurrence of title phrases were examined along with building a network visualization map with VOSviewer to highlight research hotspots in the subject of data ethics. Since 2010, the number of articles addressing ethical issues in data has risen exponentially. In this regard, it should be noted that China's publications require quality enhancement in order to be referenced more frequently. There were nations with low H-Index scores, such as Canada and Australia, that published little but were heavily referenced; moreover, they need to raise the amount of articles they publish (Luma-Osmani, Ismaili, & Raufi, 2020). The keyword analysis revealed that ethics is shifting from a philosophical study to a business and data-driven one. The author collaboration clusters revealed that writers who worked alone were mentioned more frequently (Luma-Osmani, Ismaili, & Raufi, 2020).

Lastly, certain drawbacks of the paper should be noted. To begin, in comparison to other academic disciplines, just a modest number of articles (1483) were published throughout the five-decade period chosen. It indicates that just 30 articles were published worldwide per year, or 2.5 papers every month. As a result, the threshold was not raised to a higher number. Second, the VOSviewer program fully supports Web of Science, Scopus, Dimensions, and PubMed file formats. As a result, because our study used data from IEEE, only a few analyses were permitted, therefore excluding key analyses such as country and institute co-authorship, journal and reference co-citation analysis, and so on. Those databases will almost certainly be included in future research. Finally, because all of the articles chosen for bibliometric analysis were written in English, papers in other languages may have been overlooked (Luma-Osmani, Ismaili, & Raufi, 2020).

# CONCLUSIONS, CONTRIBUTIONS AND FUTURE WORK

---

## 7.1 Conclusions

Machine learning approaches, strategies, and algorithms have sparked intense scientific research in the discovery of causal connections across a wide range of datasets and domains. We review the most current and noteworthy works on this issue from relevant journals and conference proceedings from a variety of views. The findings show that causal reasoning is gaining momentum in research for a variety of fields, including smart agriculture. Unfortunately, a lot of issues must yet be solved (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020). The first critical issue is the creation, availability, and ownership of relevant agricultural datasets. Policies and standards for high-quality data that encourages trust are still being developed. Furthermore, the efficiency, dependability, cost-effectiveness, and utility of causal reasoning techniques in smart agriculture have a long way to go before reaching full maturity. Finally, the possible abuse of data and causal reasoning results provides an extra ethical and legal problem that need normative framework formulation and control (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

Some publications compare Structural Equation Models (SEMs) to Structural Causal Models (SCMs), but others focus on purely probabilistic models for explaining causality using Bayes Nets without including confounding, DE confounding, or counterfactuals. A significant progress in this regard, presents the application of Bayesian nets, as a tool for causality identification.

Furthermore, based on the findings of the review, we may infer that causal reasoning is only partially studied in this field of agriculture (Luma-Osmani, Ismaili, Raufi, & Zenuni, 2020).

Previously, the study of phenomena was primarily done by linear regression, with the assumption that the dependent attribute is of continuous type. However, as we know from real life, the presentation of a problem is usually impacted by two or more elements (Luma-Osmani, Ismaili, & Ram Pal, 2021)

As a result, as technology has advanced, logistic regression has become more extensively utilized, and it is being used as much as feasible while decreasing the usage of linear regression in clarifying numerous phenomena. By giving a probability score for data, logistic regression declares and estimates the rules among one dependent variable of binary type, with multiple independent ones (Luma-Osmani, Ismaili, & Ram Pal, 2021).

The statistical quantitative studies for understanding this data are getting more flawless, with numerous statistical programs providing a clearer and more equitable picture of many societal concerns. The logistic regression analysis is used in statistics to explain how a variable is affected by the change of two or more independent variables. The created model examines data to determine how various factors have influenced Cyberstalks' behavior. As a result, because the coefficient of the variable gender is negative, females are more likely to be harassed and cyberstalked. The older you get, the more likely you are to be a victim of cyberstalking, and harassing someone makes you more likely to be a victim of cyberstalking (Luma-Osmani, Ismaili, & Ram Pal, Building a Model in Discovering Multivariate Causal Rules for Exploratory Analyses, 2021).

## 7.2 Thesis Contributions

1. A systematic Literature Review (SLR) of causality is offered.
2. Bibliometrics Analyses based on key words, titles, co-authors, number of publications and sum of citations has been performed accordingly.
3. Unknown causality areas were explored, such as:
  - a. Smart Agriculture
  - b. Computer Crime
4. Novel joint entropy-based Algorithm named COJEC was presented.
5. Empirical representation of Cyberstalking has been made based on multivariate logit model, in regard to three independent variables.
6. Questionnaire results were publicly issued on the web data repositories marking a minor contribution to the research community.
7. Ethical concerns related to publicly accessible data has been presented.

## 7.3 Future Work

Whereas if sample was expanded to include new variables such as education, culture, other ethnicities, and a larger sample size, the results might alter. It is also ideal to divide participants into focus groups and test groups. Similarly, we plan to do more research with currently available public databases, with a focus on unexplored domains in causal discovery (Luma-Osmani, Ismaili, & Ram Pal, 2021).

## **APPENDIX A**

The questionnaire

This questionnaire is only for research purposes and all the credentials will be kept confidential.

In order to complete the survey, please circle the appropriate letters or write your answer on the given space. Your cooperation will be highly appreciated.

**1. Gender:**

a) Male    b) Female

**2. Age:** \_\_\_\_\_

Cyber stalking is the process of harassing, false accusation, defamation of a person or a group using social networks, electronic devices or anything else that in fact it's the product of ICT (Information and Communication Technology).

**3. Have you ever been a victim of cyber stalking?**

a) Yes        b) No

NOTE: If your answer to question 3 is YES continue with the following questions, else if your answer is NO jump to question 10.

**4. If your stalking took any electronic, form please indicate which form your stalker used most often:**

a) Facebook    b) Instagram    c) Snapchat    d) Twitter    e) Other

**5. In which form was the harassment done?**

a) Threatening or abusive emails        b) Threatened you in chat rooms or comments

c) Posted false information        d) encouraged others to harass or insult you

e) Ordered goods online in your name    f) any other behavior you found distressing in any way

**6. Do you know who the person who harassed you was?**

a) No        b) ex-partner    c) friend    d) other

**7. Did your stalker communicated with you via these social media platforms or in any other way?**

a) Yes        b) No

**8.** On average, how often did you receive some sort of contact from your stalker?

a) Hourly      b) Daily      c) Weekly      d) Monthly      e) 2-3 months

**9.** Did you managed to get rid of the stalker or he/she stopped harassing you?

a) Yes      b) No      c) They stopped harassing me

**10.** Have you ever harassed someone?

a) Yes      b) No

NOTE: If your answer to question 10 is YES continue with the following questions, else if your answer is NO jump to question 14.

**11.** What was the purpose that made you cyber stalk someone?

a) Personal issues      b) They harassed me earlier      c) For gossip in group chat      d) Other

**12.** Did you reach your goal that made you stalk someone?

a) Yes      b) No

**13.** Did you find pleasure whilst you were stalking someone?

a) Yes      b) No

**14.** Did you know that cyber stalking is in fact a criminal offense?

a) Yes      b) No

Thank you!



## APPENDIX B

### Apriori Association Rules from "Cyberstalking" dataset

1. Social\_media\_communication=Yes 41 ==> Victim\_of\_cyberstalking=Yes 41 <conf:(1)> lift:(2.15)  
lev:(0.15) [21] conv:(21.94)
2. Get\_rid\_of\_the\_cyberstalker=Yes 59 ==> Victim\_of\_cyberstalking=Yes 59 <conf:(1)> lift:(2.15)  
lev:(0.22) [31] conv:(31.58)
3. Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes 38 ==>  
Victim\_of\_cyberstalking=Yes 38 <conf:(1)> lift:(2.15) lev:(0.14) [20] conv:(20.34)
4. Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=Yes 38 ==> Victim\_of\_cyberstalking=Yes 38  
<conf:(1)> lift:(2.15) lev:(0.14) [20] conv:(20.34)
5. Cyberstalking\_achieving\_goal=Yes 37 ==> You\_harassed\_someone=Yes 37 <conf:(1)> lift:(2.41)  
lev:(0.15) [21] conv:(21.63)
6. Gender=Female Get\_rid\_of\_the\_cyberstalker=Yes 36 ==> Victim\_of\_cyberstalking=Yes 36  
<conf:(1)> lift:(2.15) lev:(0.14) [19] conv:(19.27)
7. Cyberstalking\_pleasure=No 33 ==> You\_harassed\_someone=Yes 33 <conf:(1)> lift:(2.41)  
lev:(0.14) [19] conv:(19.29)
8. Social\_media\_cyberstalking=Instagram 31 ==> Victim\_of\_cyberstalking=Yes 31 <conf:(1)>  
lift:(2.15) lev:(0.12) [16] conv:(16.59)
9. Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=Yes 31 ==>  
Victim\_of\_cyberstalking=Yes 31 <conf:(1)> lift:(2.15) lev:(0.12) [16] conv:(16.59)
10. Social\_media\_cyberstalking=Instagram Get\_rid\_of\_the\_cyberstalker=Yes 30 ==>  
Victim\_of\_cyberstalking=Yes 30 <conf:(1)> lift:(2.15) lev:(0.11) [16] conv:(16.06)
11. Know\_the\_stalker=No Get\_rid\_of\_the\_cyberstalker=Yes 30 ==> Victim\_of\_cyberstalking=Yes 30  
<conf:(1)> lift:(2.15) lev:(0.11) [16] conv:(16.06)
12. Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=No 28 ==> Victim\_of\_cyberstalking=Yes  
28 <conf:(1)> lift:(2.15) lev:(0.11) [14] conv:(14.99)
13. Cyberstalking\_pleasure=Yes 26 ==> You\_harassed\_someone=Yes 26 <conf:(1)> lift:(2.41) lev:(0.11)  
[15] conv:(15.2)
14. Social\_media\_communication=Yes Criminal\_offense=Yes 26 ==> Victim\_of\_cyberstalking=Yes 26  
<conf:(1)> lift:(2.15) lev:(0.1) [13] conv:(13.92)

15. Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=Yes 24 ==> Victim\_of\_cyberstalking=Yes 24 <conf:(1)> lift:(2.15) lev:(0.09) [12] conv:(12.85)

16. Gender=Male Get\_rid\_of\_the\_cyberstalker=Yes 23 ==> Victim\_of\_cyberstalking=Yes 23 <conf:(1)> lift:(2.15) lev:(0.09) [12] conv:(12.31)

17. Gender=Female Social\_media\_communication=Yes 23 ==> Victim\_of\_cyberstalking=Yes 23 <conf:(1)> lift:(2.15) lev:(0.09) [12] conv:(12.31)

18. Social\_media\_cyberstalking=Instagram Social\_media\_communication=Yes 23 ==> Victim\_of\_cyberstalking=Yes 23 <conf:(1)> lift:(2.15) lev:(0.09) [12] conv:(12.31)

19. Social\_media\_cyberstalking=Instagram Criminal\_offense=Yes 23 ==> Victim\_of\_cyberstalking=Yes 23 <conf:(1)> lift:(2.15) lev:(0.09) [12] conv:(12.31)

20. Cyberstalking\_achieving\_goal=Yes Cyberstalking\_pleasure=Yes 23 ==> You\_harassed\_someone=Yes 23 <conf:(1)> lift:(2.41) lev:(0.09) [13] conv:(13.44)

21. Gender=Female Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=Yes 23 ==> Victim\_of\_cyberstalking=Yes 23 <conf:(1)> lift:(2.15) lev:(0.09) [12] conv:(12.31)

22. Form\_of\_harassment=Posting false information 22 ==> Victim\_of\_cyberstalking=Yes 22 <conf:(1)> lift:(2.15) lev:(0.08) [11] conv:(11.77)

23. Cyberstalking\_achieving\_goal=No 22 ==> You\_harassed\_someone=Yes 22 <conf:(1)> lift:(2.41) lev:(0.09) [12] conv:(12.86)

24. School=Kiril Pejcinoviq Get\_rid\_of\_the\_cyberstalker=Yes 22 ==> Victim\_of\_cyberstalking=Yes 22 <conf:(1)> lift:(2.15) lev:(0.08) [11] conv:(11.77)

25. School=7Marsi Get\_rid\_of\_the\_cyberstalker=Yes 22 ==> Victim\_of\_cyberstalking=Yes 22 <conf:(1)> lift:(2.15) lev:(0.08) [11] conv:(11.77)

26. School=7Marsi Victim\_of\_cyberstalking=Yes 22 ==> Get\_rid\_of\_the\_cyberstalker=Yes 22 <conf:(1)> lift:(2.41) lev:(0.09) [12] conv:(12.86)

27. Social\_media\_communication=Yes You\_harassed\_someone=Yes 22 ==> Victim\_of\_cyberstalking=Yes 22 <conf:(1)> lift:(2.15) lev:(0.08) [11] conv:(11.77)

28. Victim\_of\_cyberstalking=Yes Cyberstalking\_pleasure=No 22 ==> You\_harassed\_someone=Yes 22 <conf:(1)> lift:(2.41) lev:(0.09) [12] conv:(12.86)

29. Gender=Female Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes 22 ==> Victim\_of\_cyberstalking=Yes 22 <conf:(1)> lift:(2.15) lev:(0.08) [11] conv:(11.77)

30. Social\_media\_cyberstalking=Instagram Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes 22 ==> Victim\_of\_cyberstalking=Yes 22 <conf:(1)> lift:(2.15) lev:(0.08) [11] conv:(11.77)

31. Social\_media\_cyberstalking=Instagram Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=Yes 22  
 ==> Victim\_of\_cyberstalking=Yes 22 <conf:(1)> lift:(2.15) lev:(0.08) [11] conv:(11.77)

32. Gender=Female Cyberstalking\_achieving\_goal=Yes 21 ==> You\_harassed\_someone=Yes 21  
 <conf:(1)> lift:(2.41) lev:(0.09) [12] conv:(12.27)

33. Gender=Female Cyberstalking\_pleasure=No 21 ==> You\_harassed\_someone=Yes 21 <conf:(1)>  
 lift:(2.41) lev:(0.09) [12] conv:(12.27)

34. Social\_media\_communication=No Get\_rid\_of\_the\_cyberstalker=Yes 21 ==>  
 Victim\_of\_cyberstalking=Yes 21 <conf:(1)> lift:(2.15) lev:(0.08) [11] conv:(11.24)

35. Cyberstalking\_achieving\_goal=Yes Criminal\_offense=Yes 21 ==> You\_harassed\_someone=Yes 21  
 <conf:(1)> lift:(2.41) lev:(0.09) [12] conv:(12.27)

36. Gender=Female Know\_the\_stalker=No 20 ==> Victim\_of\_cyberstalking=Yes 20 <conf:(1)>  
 lift:(2.15) lev:(0.08) [10] conv:(10.7)

37. Form\_of\_harassment=Posting false information Get\_rid\_of\_the\_cyberstalker=Yes 20 ==>  
 Victim\_of\_cyberstalking=Yes 20 <conf:(1)> lift:(2.15) lev:(0.08) [10] conv:(10.7)

38. Know\_the\_stalker=No Social\_media\_communication=Yes 20 ==> Victim\_of\_cyberstalking=Yes 20  
 <conf:(1)> lift:(2.15) lev:(0.08) [10] conv:(10.7)

39. Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=No 20 ==> Victim\_of\_cyberstalking=Yes 20  
 <conf:(1)> lift:(2.15) lev:(0.08) [10] conv:(10.7)

40. Cyberstalking\_pleasure=No Criminal\_offense=Yes 20 ==> You\_harassed\_someone=Yes 20  
 <conf:(1)> lift:(2.41) lev:(0.08) [11] conv:(11.69)

41. Gender=Female Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=Yes 20 ==>  
 Victim\_of\_cyberstalking=Yes 20 <conf:(1)> lift:(2.15) lev:(0.08) [10] conv:(10.7)

42. Cyberstalking\_purpose=Other 19 ==> You\_harassed\_someone=Yes 19 <conf:(1)> lift:(2.41)  
 lev:(0.08) [11] conv:(11.11)

43. School=Kiril Pejcinoviq Social\_media\_communication=Yes 19 ==> Victim\_of\_cyberstalking=Yes 19  
 <conf:(1)> lift:(2.15) lev:(0.07) [10] conv:(10.17)

44. Victim\_of\_cyberstalking=No Cyberstalking\_achieving\_goal=Yes 19 ==> You\_harassed\_someone=Yes  
 19 <conf:(1)> lift:(2.41) lev:(0.08) [11] conv:(11.11)

45. Social\_media\_communication=Yes You\_harassed\_someone=No 19 ==> Victim\_of\_cyberstalking=Yes  
 19 <conf:(1)> lift:(2.15) lev:(0.07) [10] conv:(10.17)

46. Victim\_of\_cyberstalking=Yes Cyberstalking\_achieving\_goal=No 19 ==> You\_harassed\_someone=Yes  
 19 <conf:(1)> lift:(2.41) lev:(0.08) [11] conv:(11.11)

47. Social\_media\_communication=Yes You\_harassed\_someone=No 19 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 19 <conf:(1)> lift:(2.41) lev:(0.08) [11] conv:(11.11)

48. Cyberstalking\_achieving\_goal=No Cyberstalking\_pleasure=No 19 ==> You\_harassed\_someone=Yes  
19 <conf:(1)> lift:(2.41) lev:(0.08) [11] conv:(11.11)

49. Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=No  
19 ==> Victim\_of\_cyberstalking=Yes 19 <conf:(1)> lift:(2.15) lev:(0.07) [10] conv:(10.17)

50. Victim\_of\_cyberstalking=Yes Social\_media\_communication=Yes You\_harassed\_someone=No 19 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 19 <conf:(1)> lift:(2.41) lev:(0.08) [11] conv:(11.11)

51. Social\_media\_communication=Yes You\_harassed\_someone=No 19 ==> Victim\_of\_cyberstalking=Yes  
Get\_rid\_of\_the\_cyberstalker=Yes 19 <conf:(1)> lift:(2.41) lev:(0.08) [11] conv:(11.11)

52. Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=Yes  
19 ==> Victim\_of\_cyberstalking=Yes 19 <conf:(1)> lift:(2.15) lev:(0.07) [10] conv:(10.17)

53. Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=No Criminal\_offense=Yes 19 ==>  
Victim\_of\_cyberstalking=Yes 19 <conf:(1)> lift:(2.15) lev:(0.07) [10] conv:(10.17)

54. Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=Yes Criminal\_offense=Yes 19 ==>  
Victim\_of\_cyberstalking=Yes 19 <conf:(1)> lift:(2.15) lev:(0.07) [10] conv:(10.17)

55. Form\_of\_harassment=Any other behavior founded distressing in any way 18 ==>  
Victim\_of\_cyberstalking=Yes 18 <conf:(1)> lift:(2.15) lev:(0.07) [9] conv:(9.63)

56. Cyberstalking\_purpose=For gossip in group chats 18 ==> You\_harassed\_someone=Yes 18  
<conf:(1)> lift:(2.41) lev:(0.07) [10] conv:(10.52)

57. School=Nikola Shtejn Cyberstalking\_achieving\_goal=Yes 18 ==> You\_harassed\_someone=Yes 18  
<conf:(1)> lift:(2.41) lev:(0.07) [10] conv:(10.52)

58. Gender=Male Social\_media\_communication=Yes 18 ==> Victim\_of\_cyberstalking=Yes 18  
<conf:(1)> lift:(2.15) lev:(0.07) [9] conv:(9.63)

59. Gender=Female Social\_media\_cyberstalking=Instagram 18 ==> Victim\_of\_cyberstalking=Yes 18  
<conf:(1)> lift:(2.15) lev:(0.07) [9] conv:(9.63)

60. Gender=Female Social\_media\_cyberstalking=Instagram 18 ==> Get\_rid\_of\_the\_cyberstalker=Yes 18  
<conf:(1)> lift:(2.41) lev:(0.07) [10] conv:(10.52)

61. Social\_media\_cyberstalking=Instagram You\_harassed\_someone=No 18 ==>  
Victim\_of\_cyberstalking=Yes 18 <conf:(1)> lift:(2.15) lev:(0.07) [9] conv:(9.63)

62. Know\_the\_stalker=No You\_harassed\_someone=No 18 ==> Victim\_of\_cyberstalking=Yes 18  
<conf:(1)> lift:(2.15) lev:(0.07) [9] conv:(9.63)

63. Get\_rid\_of\_the\_cyberstalker=Yes Cyberstalking\_pleasure=No 18 ==> Victim\_of\_cyberstalking=Yes  
18 <conf:(1)> lift:(2.15) lev:(0.07) [9] conv:(9.63)

64. Victim\_of\_cyberstalking=Yes Cyberstalking\_achieving\_goal=Yes 18 ==> You\_harassed\_someone=Yes  
18 <conf:(1)> lift:(2.41) lev:(0.07) [10] conv:(10.52)

65. Social\_media\_cyberstalking=Instagram You\_harassed\_someone=No 18 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 18 <conf:(1)> lift:(2.41) lev:(0.07) [10] conv:(10.52)

66. Get\_rid\_of\_the\_cyberstalker=Yes Cyberstalking\_pleasure=No 18 ==> You\_harassed\_someone=Yes  
18 <conf:(1)> lift:(2.41) lev:(0.07) [10] conv:(10.52)

67. Gender=Female Social\_media\_cyberstalking=Instagram Get\_rid\_of\_the\_cyberstalker=Yes 18 ==>  
Victim\_of\_cyberstalking=Yes 18 <conf:(1)> lift:(2.15) lev:(0.07) [9] conv:(9.63)

68. Gender=Female Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram 18 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 18 <conf:(1)> lift:(2.41) lev:(0.07) [10] conv:(10.52)

69. Gender=Female Social\_media\_cyberstalking=Instagram 18 ==> Victim\_of\_cyberstalking=Yes  
Get\_rid\_of\_the\_cyberstalker=Yes 18 <conf:(1)> lift:(2.41) lev:(0.07) [10] conv:(10.52)

70. Social\_media\_cyberstalking=Instagram Get\_rid\_of\_the\_cyberstalker=Yes  
You\_harassed\_someone=No 18 ==> Victim\_of\_cyberstalking=Yes 18 <conf:(1)> lift:(2.15) lev:(0.07) [9]  
conv:(9.63)

71. Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram You\_harassed\_someone=No 18  
==> Get\_rid\_of\_the\_cyberstalker=Yes 18 <conf:(1)> lift:(2.41) lev:(0.07) [10] conv:(10.52)

72. Social\_media\_cyberstalking=Instagram You\_harassed\_someone=No 18 ==>  
Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes 18 <conf:(1)> lift:(2.41) lev:(0.07) [10]  
conv:(10.52)

73. Know\_the\_stalker=No Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes 18 ==>  
Victim\_of\_cyberstalking=Yes 18 <conf:(1)> lift:(2.15) lev:(0.07) [9] conv:(9.63)

74. Know\_the\_stalker=No Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=Yes 18 ==>  
Victim\_of\_cyberstalking=Yes 18 <conf:(1)> lift:(2.15) lev:(0.07) [9] conv:(9.63)

75. Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=Yes Cyberstalking\_pleasure=No 18 ==>  
Victim\_of\_cyberstalking=Yes 18 <conf:(1)> lift:(2.15) lev:(0.07) [9] conv:(9.63)

76. Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes Cyberstalking\_pleasure=No 18 ==>  
You\_harassed\_someone=Yes 18 <conf:(1)> lift:(2.41) lev:(0.07) [10] conv:(10.52)

77. Get\_rid\_of\_the\_cyberstalker=Yes Cyberstalking\_pleasure=No 18 ==> Victim\_of\_cyberstalking=Yes  
You\_harassed\_someone=Yes 18 <conf:(1)> lift:(3.84) lev:(0.09) [13] conv:(13.31)

78. Gender=Female Social\_media\_communication=No 17 ==> Victim\_of\_cyberstalking=Yes 17  
<conf:(1)> lift:(2.15) lev:(0.06) [9] conv:(9.1)

79. Social\_media\_cyberstalking=Instagram Know\_the\_stalker=No 17 ==> Victim\_of\_cyberstalking=Yes 17  
<conf:(1)> lift:(2.15) lev:(0.06) [9] conv:(9.1)

80. School=Kiril Pejcinoviq Gender=Female Get\_rid\_of\_the\_cyberstalker=Yes 17 ==> Victim\_of\_cyberstalking=Yes 17  
<conf:(1)> lift:(2.15) lev:(0.06) [9] conv:(9.1)

81. School=Kiril Pejcinoviq Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes 17 ==> Victim\_of\_cyberstalking=Yes 17  
<conf:(1)> lift:(2.15) lev:(0.06) [9] conv:(9.1)

82. School=Kiril Pejcinoviq Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=No 17 ==> Victim\_of\_cyberstalking=Yes 17  
<conf:(1)> lift:(2.15) lev:(0.06) [9] conv:(9.1)

83. Gender=Female Know\_the\_stalker=No Get\_rid\_of\_the\_cyberstalker=Yes 17 ==> Victim\_of\_cyberstalking=Yes 17  
<conf:(1)> lift:(2.15) lev:(0.06) [9] conv:(9.1)

84. Know\_the\_stalker=No Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=No 17 ==> Victim\_of\_cyberstalking=Yes 17  
<conf:(1)> lift:(2.15) lev:(0.06) [9] conv:(9.1)

85. School=Kiril Pejcinoviq Know\_the\_stalker=No 16 ==> Victim\_of\_cyberstalking=Yes 16  
<conf:(1)> lift:(2.15) lev:(0.06) [8] conv:(8.56)

86. Gender=Male Cyberstalking\_achieving\_goal=Yes 16 ==> You\_harassed\_someone=Yes 16  
<conf:(1)> lift:(2.41) lev:(0.07) [9] conv:(9.35)

87. Get\_rid\_of\_the\_cyberstalker=Yes Cyberstalking\_achieving\_goal=No 16 ==> Victim\_of\_cyberstalking=Yes 16  
<conf:(1)> lift:(2.15) lev:(0.06) [8] conv:(8.56)

88. Get\_rid\_of\_the\_cyberstalker=Yes Cyberstalking\_achieving\_goal=No 16 ==> You\_harassed\_someone=Yes 16  
<conf:(1)> lift:(2.41) lev:(0.07) [9] conv:(9.35)

89. Cyberstalking\_achieving\_goal=Yes Criminal\_offense=No 16 ==> You\_harassed\_someone=Yes 16  
<conf:(1)> lift:(2.41) lev:(0.07) [9] conv:(9.35)

90. School=7Marsi Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=Yes 16 ==> Victim\_of\_cyberstalking=Yes 16  
<conf:(1)> lift:(2.15) lev:(0.06) [8] conv:(8.56)

91. School=7Marsi Victim\_of\_cyberstalking=Yes You\_harassed\_someone=Yes 16 ==> Get\_rid\_of\_the\_cyberstalker=Yes 16  
<conf:(1)> lift:(2.41) lev:(0.07) [9] conv:(9.35)

92. Gender=Male Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes 16 ==> Victim\_of\_cyberstalking=Yes 16  
<conf:(1)> lift:(2.15) lev:(0.06) [8] conv:(8.56)

93. Gender=Female Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=No 16 ==> Victim\_of\_cyberstalking=Yes 16  
<conf:(1)> lift:(2.15) lev:(0.06) [8] conv:(8.56)

94. Social\_media\_cyberstalking=Instagram Know\_the\_stalker=No Get\_rid\_of\_the\_cyberstalker=Yes 16  
==> Victim\_of\_cyberstalking=Yes 16 <conf:(1)> lift:(2.15) lev:(0.06) [8] conv:(8.56)

95. Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=Yes Cyberstalking\_achieving\_goal=No 16  
==> Victim\_of\_cyberstalking=Yes 16 <conf:(1)> lift:(2.15) lev:(0.06) [8] conv:(8.56)

96. Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes Cyberstalking\_achieving\_goal=No 16  
==> You\_harassed\_someone=Yes 16 <conf:(1)> lift:(2.41) lev:(0.07) [9] conv:(9.35)

97. Get\_rid\_of\_the\_cyberstalker=Yes Cyberstalking\_achieving\_goal=No 16 ==>  
Victim\_of\_cyberstalking=Yes You\_harassed\_someone=Yes 16 <conf:(1)> lift:(3.84) lev:(0.08) [11]  
conv:(11.83)

98. Victim\_of\_cyberstalking=Yes Cyberstalking\_achieving\_goal=No Cyberstalking\_pleasure=No 16 ==>  
You\_harassed\_someone=Yes 16 <conf:(1)> lift:(2.41) lev:(0.07) [9] conv:(9.35)

99. School=Nikola Shtejn Get\_rid\_of\_the\_cyberstalker=Yes 15 ==> Victim\_of\_cyberstalking=Yes 15  
<conf:(1)> lift:(2.15) lev:(0.06) [8] conv:(8.03)

100. School=Nikola Shtejn Cyberstalking\_pleasure=No 15 ==> You\_harassed\_someone=Yes 15  
<conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.77)

101. Form\_of\_harassment=Posting false information Social\_media\_communication=Yes 15 ==>  
Victim\_of\_cyberstalking=Yes 15 <conf:(1)> lift:(2.15) lev:(0.06) [8] conv:(8.03)

102. Social\_media\_communication=Yes Criminal\_offense=No 15 ==> Victim\_of\_cyberstalking=Yes 15  
<conf:(1)> lift:(2.15) lev:(0.06) [8] conv:(8.03)

103. Get\_rid\_of\_the\_cyberstalker=Yes Cyberstalking\_achieving\_goal=Yes 15 ==>  
Victim\_of\_cyberstalking=Yes 15 <conf:(1)> lift:(2.15) lev:(0.06) [8] conv:(8.03)

104. Victim\_of\_cyberstalking=Yes Cyberstalking\_pleasure=Yes 15 ==> You\_harassed\_someone=Yes 15  
<conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.77)

105. Get\_rid\_of\_the\_cyberstalker=Yes Cyberstalking\_achieving\_goal=Yes 15 ==>  
You\_harassed\_someone=Yes 15 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.77)

106. Cyberstalking\_pleasure=Yes Criminal\_offense=Yes 15 ==> You\_harassed\_someone=Yes 15  
<conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.77)

107. Gender=Male Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=Yes 15 ==>  
Victim\_of\_cyberstalking=Yes 15 <conf:(1)> lift:(2.15) lev:(0.06) [8] conv:(8.03)

108. Gender=Female Social\_media\_communication=Yes Criminal\_offense=Yes 15 ==>  
Victim\_of\_cyberstalking=Yes 15 <conf:(1)> lift:(2.15) lev:(0.06) [8] conv:(8.03)

109. Gender=Female Cyberstalking\_achieving\_goal=Yes Criminal\_offense=Yes 15 ==>  
You\_harassed\_someone=Yes 15 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.77)

110. Social\_media\_cyberstalking=Instagram Social\_media\_communication=Yes  
 You\_harassed\_someone=No 15 ==> Victim\_of\_cyberstalking=Yes 15 <conf:(1)> lift:(2.15) lev:(0.06) [8]  
 conv:(8.03)

111. Social\_media\_cyberstalking=Instagram Social\_media\_communication=Yes Criminal\_offense=Yes 15  
 ==> Victim\_of\_cyberstalking=Yes 15 <conf:(1)> lift:(2.15) lev:(0.06) [8] conv:(8.03)

112. Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=Yes Cyberstalking\_achieving\_goal=Yes  
 15 ==> Victim\_of\_cyberstalking=Yes 15 <conf:(1)> lift:(2.15) lev:(0.06) [8] conv:(8.03)

113. Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes Cyberstalking\_achieving\_goal=Yes  
 15 ==> You\_harassed\_someone=Yes 15 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.77)

114. Get\_rid\_of\_the\_cyberstalker=Yes Cyberstalking\_achieving\_goal=Yes 15 ==>  
 Victim\_of\_cyberstalking=Yes You\_harassed\_someone=Yes 15 <conf:(1)> lift:(3.84) lev:(0.08) [11]  
 conv:(11.09)

115. Social\_media\_cyberstalking=Instagram Social\_media\_communication=Yes  
 You\_harassed\_someone=No 15 ==> Get\_rid\_of\_the\_cyberstalker=Yes 15 <conf:(1)> lift:(2.41)  
 lev:(0.06) [8] conv:(8.77)

116. Social\_media\_cyberstalking=Instagram Social\_media\_communication=Yes  
 Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=No 15 ==> Victim\_of\_cyberstalking=Yes 15  
 <conf:(1)> lift:(2.15) lev:(0.06) [8] conv:(8.03)

117. Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram  
 Social\_media\_communication=Yes You\_harassed\_someone=No 15 ==>  
 Get\_rid\_of\_the\_cyberstalker=Yes 15 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.77)

118. Social\_media\_cyberstalking=Instagram Social\_media\_communication=Yes  
 You\_harassed\_someone=No 15 ==> Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes 15  
 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.77)

119. Social\_media\_cyberstalking=Snapchat 14 ==> Victim\_of\_cyberstalking=Yes 14 <conf:(1)>  
 lift:(2.15) lev:(0.05) [7] conv:(7.49)

120. Form\_of\_harassment=Threatened in chat rooms or comments 14 ==> Victim\_of\_cyberstalking=Yes  
 14 <conf:(1)> lift:(2.15) lev:(0.05) [7] conv:(7.49)

121. Know\_the\_stalker=Other 14 ==> Victim\_of\_cyberstalking=Yes 14 <conf:(1)> lift:(2.15) lev:(0.05)  
 [7] conv:(7.49)

122. Form\_of\_harassment=Threatened in chat rooms or comments 14 ==>  
 Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.18)

123. School=Kiril Pejcinoviq Social\_media\_cyberstalking=Instagram 14 ==> Victim\_of\_cyberstalking=Yes  
 14 <conf:(1)> lift:(2.15) lev:(0.05) [7] conv:(7.49)



124. School=Kiril Pejcinoviq Social\_media\_cyberstalking=Instagram 14 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.18)

125. School=7Marsi Social\_media\_communication=Yes 14 ==> Victim\_of\_cyberstalking=Yes 14  
<conf:(1)> lift:(2.15) lev:(0.05) [7] conv:(7.49)

126. School=7Marsi Social\_media\_communication=Yes 14 ==> Get\_rid\_of\_the\_cyberstalker=Yes 14  
<conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.18)

127. Gender=Female Cyberstalking\_purpose=For gossip in group chats 14 ==>  
You\_harassed\_someone=Yes 14 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.18)

128. Form\_of\_harassment=Threatened in chat rooms or comments Get\_rid\_of\_the\_cyberstalker=Yes 14  
==> Victim\_of\_cyberstalking=Yes 14 <conf:(1)> lift:(2.15) lev:(0.05) [7] conv:(7.49)

129. Victim\_of\_cyberstalking=Yes Form\_of\_harassment=Threatened in chat rooms or comments 14 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.18)

130. Form\_of\_harassment=Threatened in chat rooms or comments 14 ==> Victim\_of\_cyberstalking=Yes  
Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.18)

131. Form\_of\_harassment=Any other behavior founded distressing in any way  
Get\_rid\_of\_the\_cyberstalker=Yes 14 ==> Victim\_of\_cyberstalking=Yes 14 <conf:(1)> lift:(2.15)  
lev:(0.05) [7] conv:(7.49)

132. Cyberstalking\_purpose=For gossip in group chats Cyberstalking\_achieving\_goal=Yes 14 ==>  
You\_harassed\_someone=Yes 14 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.18)

133. Cyberstalking\_purpose=Other Cyberstalking\_pleasure=No 14 ==> You\_harassed\_someone=Yes 14  
<conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.18)

134. Cyberstalking\_achieving\_goal=Yes Cyberstalking\_pleasure=No 14 ==> You\_harassed\_someone=Yes  
14 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.18)

135. Cyberstalking\_achieving\_goal=No Criminal\_offense=Yes 14 ==> You\_harassed\_someone=Yes 14  
<conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.18)

136. School=Kiril Pejcinoviq Gender=Female Social\_media\_communication=Yes 14 ==>  
Victim\_of\_cyberstalking=Yes 14 <conf:(1)> lift:(2.15) lev:(0.05) [7] conv:(7.49)

137. School=Kiril Pejcinoviq Social\_media\_cyberstalking=Instagram Get\_rid\_of\_the\_cyberstalker=Yes 14  
==> Victim\_of\_cyberstalking=Yes 14 <conf:(1)> lift:(2.15) lev:(0.05) [7] conv:(7.49)

138. School=Kiril Pejcinoviq Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram 14 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.18)

139. School=Kiril Pejcinoviq Social\_media\_cyberstalking=Instagram 14 ==> Victim\_of\_cyberstalking=Yes  
Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.18)

140. School=Kiril Pejcinoviq Know\_the\_stalker=No Get\_rid\_of\_the\_cyberstalker=Yes 14 ==>  
Victim\_of\_cyberstalking=Yes 14 <conf:(1)> lift:(2.15) lev:(0.05) [7] conv:(7.49)

141. School=Kiril Pejcinoviq Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=Yes 14 ==>  
Victim\_of\_cyberstalking=Yes 14 <conf:(1)> lift:(2.15) lev:(0.05) [7] conv:(7.49)

142. School=7Marsi Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes 14 ==>  
Victim\_of\_cyberstalking=Yes 14 <conf:(1)> lift:(2.15) lev:(0.05) [7] conv:(7.49)

143. School=7Marsi Victim\_of\_cyberstalking=Yes Social\_media\_communication=Yes 14 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.18)

144. School=7Marsi Social\_media\_communication=Yes 14 ==> Victim\_of\_cyberstalking=Yes  
Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.18)

145. Gender=Female Social\_media\_cyberstalking=Instagram Criminal\_offense=Yes 14 ==>  
Victim\_of\_cyberstalking=Yes 14 <conf:(1)> lift:(2.15) lev:(0.05) [7] conv:(7.49)

146. Gender=Female Social\_media\_communication=No Get\_rid\_of\_the\_cyberstalker=Yes 14 ==>  
Victim\_of\_cyberstalking=Yes 14 <conf:(1)> lift:(2.15) lev:(0.05) [7] conv:(7.49)

147. Gender=Female Victim\_of\_cyberstalking=Yes Cyberstalking\_pleasure=No 14 ==>  
You\_harassed\_someone=Yes 14 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.18)

148. Gender=Female Social\_media\_cyberstalking=Instagram Criminal\_offense=Yes 14 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.18)

149. Gender=Female Cyberstalking\_pleasure=No Criminal\_offense=Yes 14 ==>  
You\_harassed\_someone=Yes 14 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.18)

150. Social\_media\_cyberstalking=Instagram Know\_the\_stalker=No Social\_media\_communication=Yes  
14 ==> Victim\_of\_cyberstalking=Yes 14 <conf:(1)> lift:(2.15) lev:(0.05) [7] conv:(7.49)

151. Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=No 14 ==>  
Victim\_of\_cyberstalking=Yes 14 <conf:(1)> lift:(2.15) lev:(0.05) [7] conv:(7.49)

152. Social\_media\_communication=Yes You\_harassed\_someone=Yes Criminal\_offense=Yes 14 ==>  
Victim\_of\_cyberstalking=Yes 14 <conf:(1)> lift:(2.15) lev:(0.05) [7] conv:(7.49)

153. Social\_media\_communication=No Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=Yes 14 ==>  
Victim\_of\_cyberstalking=Yes 14 <conf:(1)> lift:(2.15) lev:(0.05) [7] conv:(7.49)

154. Victim\_of\_cyberstalking=Yes Cyberstalking\_pleasure=No Criminal\_offense=Yes 14 ==>  
You\_harassed\_someone=Yes 14 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.18)

155. Cyberstalking\_achieving\_goal=Yes Cyberstalking\_pleasure=Yes Criminal\_offense=Yes 14 ==>  
You\_harassed\_someone=Yes 14 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.18)

156. Gender=Female Social\_media\_cyberstalking=Instagram Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=Yes 14 ==> Victim\_of\_cyberstalking=Yes 14 <conf:(1)> lift:(2.15) lev:(0.05) [7] conv:(7.49)

157. Gender=Female Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram Criminal\_offense=Yes 14 ==> Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.18)

158. Gender=Female Social\_media\_cyberstalking=Instagram Criminal\_offense=Yes 14 ==> Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(1)> lift:(2.41) lev:(0.06) [8] conv:(8.18)

159. Gender=Female Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=Yes 14 ==> Victim\_of\_cyberstalking=Yes 14 <conf:(1)> lift:(2.15) lev:(0.05) [7] conv:(7.49)

160. Gender=Female Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=Yes Criminal\_offense=Yes 14 ==> Victim\_of\_cyberstalking=Yes 14 <conf:(1)> lift:(2.15) lev:(0.05) [7] conv:(7.49)

161. Social\_media\_cyberstalking=Instagram Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=Yes 14 ==> Victim\_of\_cyberstalking=Yes 14 <conf:(1)> lift:(2.15) lev:(0.05) [7] conv:(7.49)

162. Know\_the\_stalker=No 36 ==> Victim\_of\_cyberstalking=Yes 35 <conf:(0.97)> lift:(2.09) lev:(0.13) [18] conv:(9.63)

163. Social\_media\_cyberstalking=Instagram 31 ==> Get\_rid\_of\_the\_cyberstalker=Yes 30 <conf:(0.97)> lift:(2.33) lev:(0.12) [17] conv:(9.06)

164. Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram 31 ==> Get\_rid\_of\_the\_cyberstalker=Yes 30 <conf:(0.97)> lift:(2.33) lev:(0.12) [17] conv:(9.06)

165. Social\_media\_cyberstalking=Instagram 31 ==> Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes 30 <conf:(0.97)> lift:(2.33) lev:(0.12) [17] conv:(9.06)

166. Victim\_of\_cyberstalking=Yes You\_harassed\_someone=No 29 ==> Get\_rid\_of\_the\_cyberstalker=Yes 28 <conf:(0.97)> lift:(2.32) lev:(0.11) [15] conv:(8.48)

167. Social\_media\_communication=No 26 ==> Victim\_of\_cyberstalking=Yes 25 <conf:(0.96)> lift:(2.07) lev:(0.09) [12] conv:(6.96)

168. Know\_the\_stalker=No Criminal\_offense=Yes 24 ==> Victim\_of\_cyberstalking=Yes 23 <conf:(0.96)> lift:(2.06) lev:(0.08) [11] conv:(6.42)

169. Gender=Female Social\_media\_communication=Yes 23 ==> Get\_rid\_of\_the\_cyberstalker=Yes 22 <conf:(0.96)> lift:(2.3) lev:(0.09) [12] conv:(6.72)

170. Social\_media\_cyberstalking=Instagram Social\_media\_communication=Yes 23 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 22 <conf:(0.96)> lift:(2.3) lev:(0.09) [12] conv:(6.72)

171. Social\_media\_cyberstalking=Instagram Criminal\_offense=Yes 23 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 22 <conf:(0.96)> lift:(2.3) lev:(0.09) [12] conv:(6.72)

172. Gender=Female Victim\_of\_cyberstalking=Yes Social\_media\_communication=Yes 23 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 22 <conf:(0.96)> lift:(2.3) lev:(0.09) [12] conv:(6.72)

173. Gender=Female Social\_media\_communication=Yes 23 ==> Victim\_of\_cyberstalking=Yes  
Get\_rid\_of\_the\_cyberstalker=Yes 22 <conf:(0.96)> lift:(2.3) lev:(0.09) [12] conv:(6.72)

174. Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram  
Social\_media\_communication=Yes 23 ==> Get\_rid\_of\_the\_cyberstalker=Yes 22 <conf:(0.96)> lift:(2.3)  
lev:(0.09) [12] conv:(6.72)

175. Social\_media\_cyberstalking=Instagram Social\_media\_communication=Yes 23 ==>  
Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes 22 <conf:(0.96)> lift:(2.3) lev:(0.09) [12]  
conv:(6.72)

176. Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram Criminal\_offense=Yes 23 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 22 <conf:(0.96)> lift:(2.3) lev:(0.09) [12] conv:(6.72)

177. Social\_media\_cyberstalking=Instagram Criminal\_offense=Yes 23 ==> Victim\_of\_cyberstalking=Yes  
Get\_rid\_of\_the\_cyberstalker=Yes 22 <conf:(0.96)> lift:(2.3) lev:(0.09) [12] conv:(6.72)

178. Victim\_of\_cyberstalking=Yes You\_harassed\_someone=No Criminal\_offense=Yes 20 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 19 <conf:(0.95)> lift:(2.29) lev:(0.08) [10] conv:(5.85)

179. Know\_the\_stalker=No You\_harassed\_someone=Yes 18 ==> Victim\_of\_cyberstalking=Yes 17  
<conf:(0.94)> lift:(2.03) lev:(0.06) [8] conv:(4.82)

180. Social\_media\_communication=No Criminal\_offense=Yes 18 ==> Victim\_of\_cyberstalking=Yes 17  
<conf:(0.94)> lift:(2.03) lev:(0.06) [8] conv:(4.82)

181. Know\_the\_stalker=No You\_harassed\_someone=No 18 ==> Get\_rid\_of\_the\_cyberstalker=Yes 17  
<conf:(0.94)> lift:(2.27) lev:(0.07) [9] conv:(5.26)

182. School=Kiril Pejcinoviq Victim\_of\_cyberstalking=Yes You\_harassed\_someone=No 18 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 17 <conf:(0.94)> lift:(2.27) lev:(0.07) [9] conv:(5.26)

183. Victim\_of\_cyberstalking=Yes Know\_the\_stalker=No You\_harassed\_someone=No 18 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 17 <conf:(0.94)> lift:(2.27) lev:(0.07) [9] conv:(5.26)

184. Know\_the\_stalker=No You\_harassed\_someone=No 18 ==> Victim\_of\_cyberstalking=Yes  
Get\_rid\_of\_the\_cyberstalker=Yes 17 <conf:(0.94)> lift:(2.27) lev:(0.07) [9] conv:(5.26)

185. Social\_media\_cyberstalking=Instagram Know\_the\_stalker=No 17 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 16 <conf:(0.94)> lift:(2.27) lev:(0.06) [8] conv:(4.97)

186. Gender=Female Victim\_of\_cyberstalking=Yes You\_harassed\_someone=No 17 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 16 <conf:(0.94)> lift:(2.27) lev:(0.06) [8] conv:(4.97)

187. Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram Know\_the\_stalker=No 17 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 16 <conf:(0.94)> lift:(2.27) lev:(0.06) [8] conv:(4.97)

188. Social\_media\_cyberstalking=Instagram Know\_the\_stalker=No 17 ==> Victim\_of\_cyberstalking=Yes  
Get\_rid\_of\_the\_cyberstalker=Yes 16 <conf:(0.94)> lift:(2.27) lev:(0.06) [8] conv:(4.97)

189. Gender=Male Know\_the\_stalker=No 16 ==> Victim\_of\_cyberstalking=Yes 15 <conf:(0.94)>  
lift:(2.02) lev:(0.05) [7] conv:(4.28)

190. Know\_the\_stalker=No Social\_media\_communication=No 16 ==> Victim\_of\_cyberstalking=Yes 15  
<conf:(0.94)> lift:(2.02) lev:(0.05) [7] conv:(4.28)

191. Social\_media\_communication=No You\_harassed\_someone=Yes 16 ==>  
Victim\_of\_cyberstalking=Yes 15 <conf:(0.94)> lift:(2.02) lev:(0.05) [7] conv:(4.28)

192. School=Kiril Pejcinoviq Gender=Male Victim\_of\_cyberstalking=No Criminal\_offense=Yes 16 ==>  
You\_harassed\_someone=No 15 <conf:(0.94)> lift:(1.6) lev:(0.04) [5] conv:(3.32)

193. Social\_media\_communication=Yes Criminal\_offense=No 15 ==> Get\_rid\_of\_the\_cyberstalker=Yes  
14 <conf:(0.93)> lift:(2.25) lev:(0.05) [7] conv:(4.38)

194. Cyberstalking\_pleasure=Yes Criminal\_offense=Yes 15 ==> Cyberstalking\_achieving\_goal=Yes 14  
<conf:(0.93)> lift:(3.58) lev:(0.07) [10] conv:(5.55)

195. Gender=Female Social\_media\_communication=Yes Criminal\_offense=Yes 15 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(0.93)> lift:(2.25) lev:(0.05) [7] conv:(4.38)

196. Victim\_of\_cyberstalking=Yes Social\_media\_communication=Yes Criminal\_offense=No 15 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(0.93)> lift:(2.25) lev:(0.05) [7] conv:(4.38)

197. Social\_media\_communication=Yes Criminal\_offense=No 15 ==> Victim\_of\_cyberstalking=Yes  
Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(0.93)> lift:(2.25) lev:(0.05) [7] conv:(4.38)

198. Social\_media\_cyberstalking=Instagram Social\_media\_communication=Yes Criminal\_offense=Yes 15  
==> Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(0.93)> lift:(2.25) lev:(0.05) [7] conv:(4.38)

199. You\_harassed\_someone=Yes Cyberstalking\_pleasure=Yes Criminal\_offense=Yes 15 ==>  
Cyberstalking\_achieving\_goal=Yes 14 <conf:(0.93)> lift:(3.58) lev:(0.07) [10] conv:(5.55)

200. Cyberstalking\_pleasure=Yes Criminal\_offense=Yes 15 ==> You\_harassed\_someone=Yes  
Cyberstalking\_achieving\_goal=Yes 14 <conf:(0.93)> lift:(3.58) lev:(0.07) [10] conv:(5.55)

201. Gender=Female Victim\_of\_cyberstalking=Yes Social\_media\_communication=Yes  
Criminal\_offense=Yes 15 ==> Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(0.93)> lift:(2.25) lev:(0.05) [7]  
conv:(4.38)

202. Gender=Female Social\_media\_communication=Yes Criminal\_offense=Yes 15 ==>  
Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(0.93)> lift:(2.25) lev:(0.05) [7]  
conv:(4.38)

203. Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram  
Social\_media\_communication=Yes Criminal\_offense=Yes 15 ==> Get\_rid\_of\_the\_cyberstalker=Yes 14  
<conf:(0.93)> lift:(2.25) lev:(0.05) [7] conv:(4.38)

204. Social\_media\_cyberstalking=Instagram Social\_media\_communication=Yes Criminal\_offense=Yes 15  
==> Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(0.93)> lift:(2.25)  
lev:(0.05) [7] conv:(4.38)

205. Social\_media\_communication=Yes 41 ==> Get\_rid\_of\_the\_cyberstalker=Yes 38 <conf:(0.93)>  
lift:(2.23) lev:(0.15) [20] conv:(5.99)

206. Victim\_of\_cyberstalking=Yes Social\_media\_communication=Yes 41 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 38 <conf:(0.93)> lift:(2.23) lev:(0.15) [20] conv:(5.99)

207. Social\_media\_communication=Yes 41 ==> Victim\_of\_cyberstalking=Yes  
Get\_rid\_of\_the\_cyberstalker=Yes 38 <conf:(0.93)> lift:(2.23) lev:(0.15) [20] conv:(5.99)

208. Social\_media\_communication=Yes Criminal\_offense=Yes 26 ==> Get\_rid\_of\_the\_cyberstalker=Yes  
24 <conf:(0.92)> lift:(2.22) lev:(0.09) [13] conv:(5.07)

209. Victim\_of\_cyberstalking=Yes Social\_media\_communication=Yes Criminal\_offense=Yes 26 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 24 <conf:(0.92)> lift:(2.22) lev:(0.09) [13] conv:(5.07)

210. Social\_media\_communication=Yes Criminal\_offense=Yes 26 ==> Victim\_of\_cyberstalking=Yes  
Get\_rid\_of\_the\_cyberstalker=Yes 24 <conf:(0.92)> lift:(2.22) lev:(0.09) [13] conv:(5.07)

211. Form\_of\_harassment=Posting false information 22 ==> Get\_rid\_of\_the\_cyberstalker=Yes 20  
<conf:(0.91)> lift:(2.19) lev:(0.08) [10] conv:(4.29)

212. Victim\_of\_cyberstalking=Yes Form\_of\_harassment=Posting false information 22 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 20 <conf:(0.91)> lift:(2.19) lev:(0.08) [10] conv:(4.29)

213. Form\_of\_harassment=Posting false information 22 ==> Victim\_of\_cyberstalking=Yes  
Get\_rid\_of\_the\_cyberstalker=Yes 20 <conf:(0.91)> lift:(2.19) lev:(0.08) [10] conv:(4.29)

214. Victim\_of\_cyberstalking=Yes Criminal\_offense=No 22 ==> Get\_rid\_of\_the\_cyberstalker=Yes 20  
<conf:(0.91)> lift:(2.19) lev:(0.08) [10] conv:(4.29)

215. Gender=Female Victim\_of\_cyberstalking=Yes 40 ==> Get\_rid\_of\_the\_cyberstalker=Yes 36  
<conf:(0.9)> lift:(2.17) lev:(0.14) [19] conv:(4.68)

216. Know\_the\_stalker=No Social\_media\_communication=Yes 20 ==> Get\_rid\_of\_the\_cyberstalker=Yes 18 <conf:(0.9)> lift:(2.17) lev:(0.07) [9] conv:(3.9)

217. School=Kiril Pejcinoviq Gender=Male Criminal\_offense=Yes 20 ==> You\_harassed\_someone=No 18 <conf:(0.9)> lift:(1.54) lev:(0.04) [6] conv:(2.77)

218. Victim\_of\_cyberstalking=Yes Know\_the\_stalker=No Social\_media\_communication=Yes 20 ==> Get\_rid\_of\_the\_cyberstalker=Yes 18 <conf:(0.9)> lift:(2.17) lev:(0.07) [9] conv:(3.9)

219. Know\_the\_stalker=No Social\_media\_communication=Yes 20 ==> Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes 18 <conf:(0.9)> lift:(2.17) lev:(0.07) [9] conv:(3.9)

220. School=Kiril Pejcinoviq Social\_media\_communication=Yes 19 ==> Get\_rid\_of\_the\_cyberstalker=Yes 17 <conf:(0.89)> lift:(2.15) lev:(0.06) [9] conv:(3.7)

221. School=Kiril Pejcinoviq Victim\_of\_cyberstalking=Yes Social\_media\_communication=Yes 19 ==> Get\_rid\_of\_the\_cyberstalker=Yes 17 <conf:(0.89)> lift:(2.15) lev:(0.06) [9] conv:(3.7)

222. School=Kiril Pejcinoviq Social\_media\_communication=Yes 19 ==> Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes 17 <conf:(0.89)> lift:(2.15) lev:(0.06) [9] conv:(3.7)

223. Victim\_of\_cyberstalking=Yes 66 ==> Get\_rid\_of\_the\_cyberstalker=Yes 59 <conf:(0.89)> lift:(2.15) lev:(0.22) [31] conv:(4.82)

224. School=7Marsi You\_harassed\_someone=Yes 18 ==> Victim\_of\_cyberstalking=Yes 16 <conf:(0.89)> lift:(1.91) lev:(0.05) [7] conv:(3.21)

225. School=7Marsi You\_harassed\_someone=Yes 18 ==> Get\_rid\_of\_the\_cyberstalker=Yes 16 <conf:(0.89)> lift:(2.14) lev:(0.06) [8] conv:(3.51)

226. Gender=Male Social\_media\_communication=Yes 18 ==> Get\_rid\_of\_the\_cyberstalker=Yes 16 <conf:(0.89)> lift:(2.14) lev:(0.06) [8] conv:(3.51)

227. School=7Marsi You\_harassed\_someone=Yes 18 ==> Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes 16 <conf:(0.89)> lift:(2.14) lev:(0.06) [8] conv:(3.51)

228. Gender=Male Victim\_of\_cyberstalking=Yes Social\_media\_communication=Yes 18 ==> Get\_rid\_of\_the\_cyberstalker=Yes 16 <conf:(0.89)> lift:(2.14) lev:(0.06) [8] conv:(3.51)

229. Gender=Male Social\_media\_communication=Yes 18 ==> Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes 16 <conf:(0.89)> lift:(2.14) lev:(0.06) [8] conv:(3.51)

230. Cyberstalking\_pleasure=Yes 26 ==> Cyberstalking\_achieving\_goal=Yes 23 <conf:(0.88)> lift:(3.4) lev:(0.11) [16] conv:(4.81)

231. Gender=Male Victim\_of\_cyberstalking=Yes 26 ==> Get\_rid\_of\_the\_cyberstalker=Yes 23 <conf:(0.88)> lift:(2.13) lev:(0.09) [12] conv:(3.8)

232. You\_harassed\_someone=Yes Cyberstalking\_pleasure=Yes 26 ==> Cyberstalking\_achieving\_goal=Yes  
 23 <conf:(0.88)> lift:(3.4) lev:(0.11) [16] conv:(4.81)

233. Cyberstalking\_pleasure=Yes 26 ==> You\_harassed\_someone=Yes Cyberstalking\_achieving\_goal=Yes  
 23 <conf:(0.88)> lift:(3.4) lev:(0.11) [16] conv:(4.81)

234. Gender=Female Victim\_of\_cyberstalking=Yes Criminal\_offense=Yes 26 ==>  
 Get\_rid\_of\_the\_cyberstalker=Yes 23 <conf:(0.88)> lift:(2.13) lev:(0.09) [12] conv:(3.8)

235. Victim\_of\_cyberstalking=Yes Criminal\_offense=Yes 43 ==> Get\_rid\_of\_the\_cyberstalker=Yes 38  
 <conf:(0.88)> lift:(2.13) lev:(0.14) [20] conv:(4.19)

236. Gender=Male Victim\_of\_cyberstalking=Yes Criminal\_offense=Yes 17 ==>  
 Get\_rid\_of\_the\_cyberstalker=Yes 15 <conf:(0.88)> lift:(2.12) lev:(0.06) [7] conv:(3.31)

237. School=Kiril Pejcinoviq Know\_the\_stalker=No 16 ==> Get\_rid\_of\_the\_cyberstalker=Yes 14  
 <conf:(0.88)> lift:(2.11) lev:(0.05) [7] conv:(3.12)

238. School=Kiril Pejcinoviq Victim\_of\_cyberstalking=Yes Know\_the\_stalker=No 16 ==>  
 Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(0.88)> lift:(2.11) lev:(0.05) [7] conv:(3.12)

239. School=Kiril Pejcinoviq Know\_the\_stalker=No 16 ==> Victim\_of\_cyberstalking=Yes  
 Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(0.88)> lift:(2.11) lev:(0.05) [7] conv:(3.12)

240. School=Kiril Pejcinoviq Victim\_of\_cyberstalking=Yes Criminal\_offense=Yes 16 ==>  
 Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(0.88)> lift:(2.11) lev:(0.05) [7] conv:(3.12)

241. Gender=Female Victim\_of\_cyberstalking=Yes You\_harassed\_someone=Yes Criminal\_offense=Yes  
 16 ==> Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(0.88)> lift:(2.11) lev:(0.05) [7] conv:(3.12)

242. Gender=Female Victim\_of\_cyberstalking=Yes You\_harassed\_someone=Yes 23 ==>  
 Get\_rid\_of\_the\_cyberstalker=Yes 20 <conf:(0.87)> lift:(2.09) lev:(0.07) [10] conv:(3.36)

243. Cyberstalking\_achieving\_goal=No 22 ==> Victim\_of\_cyberstalking=Yes 19 <conf:(0.86)> lift:(1.86)  
 lev:(0.06) [8] conv:(2.94)

244. Cyberstalking\_achieving\_goal=No 22 ==> Cyberstalking\_pleasure=No 19 <conf:(0.86)> lift:(3.72)  
 lev:(0.1) [13] conv:(4.22)

245. Victim\_of\_cyberstalking=No You\_harassed\_someone=Yes 22 ==>  
 Cyberstalking\_achieving\_goal=Yes 19 <conf:(0.86)> lift:(3.31) lev:(0.09) [13] conv:(4.07)

246. You\_harassed\_someone=Yes Cyberstalking\_achieving\_goal=No 22 ==>  
 Victim\_of\_cyberstalking=Yes 19 <conf:(0.86)> lift:(1.86) lev:(0.06) [8] conv:(2.94)

247. Cyberstalking\_achieving\_goal=No 22 ==> Victim\_of\_cyberstalking=Yes  
 You\_harassed\_someone=Yes 19 <conf:(0.86)> lift:(3.31) lev:(0.09) [13] conv:(4.07)



248. Social\_media\_communication=Yes You\_harassed\_someone=Yes 22 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 19 <conf:(0.86)> lift:(2.08) lev:(0.07) [9] conv:(3.21)

249. You\_harassed\_someone=Yes Cyberstalking\_achieving\_goal=No 22 ==> Cyberstalking\_pleasure=No  
19 <conf:(0.86)> lift:(3.72) lev:(0.1) [13] conv:(4.22)

250. Cyberstalking\_achieving\_goal=No 22 ==> You\_harassed\_someone=Yes Cyberstalking\_pleasure=No  
19 <conf:(0.86)> lift:(3.72) lev:(0.1) [13] conv:(4.22)

251. Victim\_of\_cyberstalking=Yes Social\_media\_communication=Yes You\_harassed\_someone=Yes 22  
==> Get\_rid\_of\_the\_cyberstalker=Yes 19 <conf:(0.86)> lift:(2.08) lev:(0.07) [9] conv:(3.21)

252. Social\_media\_communication=Yes You\_harassed\_someone=Yes 22 ==>  
Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes 19 <conf:(0.86)> lift:(2.08) lev:(0.07) [9]  
conv:(3.21)

253. Victim\_of\_cyberstalking=Yes Know\_the\_stalker=No 35 ==> Get\_rid\_of\_the\_cyberstalker=Yes 30  
<conf:(0.86)> lift:(2.06) lev:(0.11) [15] conv:(3.41)

254. Gender=Female Know\_the\_stalker=No 20 ==> Get\_rid\_of\_the\_cyberstalker=Yes 17 <conf:(0.85)>  
lift:(2.05) lev:(0.06) [8] conv:(2.92)

255. School=Kiril Pejcinoviq Gender=Female Victim\_of\_cyberstalking=Yes 20 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 17 <conf:(0.85)> lift:(2.05) lev:(0.06) [8] conv:(2.92)

256. Gender=Female Victim\_of\_cyberstalking=Yes Know\_the\_stalker=No 20 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 17 <conf:(0.85)> lift:(2.05) lev:(0.06) [8] conv:(2.92)

257. Gender=Female Know\_the\_stalker=No 20 ==> Victim\_of\_cyberstalking=Yes  
Get\_rid\_of\_the\_cyberstalker=Yes 17 <conf:(0.85)> lift:(2.05) lev:(0.06) [8] conv:(2.92)

258. School=Kiril Pejcinoviq Victim\_of\_cyberstalking=Yes 26 ==> Get\_rid\_of\_the\_cyberstalker=Yes 22  
<conf:(0.85)> lift:(2.04) lev:(0.08) [11] conv:(3.04)

259. Victim\_of\_cyberstalking=Yes Cyberstalking\_achieving\_goal=No 19 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 16 <conf:(0.84)> lift:(2.03) lev:(0.06) [8] conv:(2.78)

260. Cyberstalking\_achieving\_goal=No Cyberstalking\_pleasure=No 19 ==> Victim\_of\_cyberstalking=Yes  
16 <conf:(0.84)> lift:(1.81) lev:(0.05) [7] conv:(2.54)

261. Victim\_of\_cyberstalking=Yes Cyberstalking\_achieving\_goal=No 19 ==> Cyberstalking\_pleasure=No  
16 <conf:(0.84)> lift:(3.62) lev:(0.08) [11] conv:(3.65)

262. Victim\_of\_cyberstalking=Yes You\_harassed\_someone=Yes Cyberstalking\_achieving\_goal=No 19  
==> Get\_rid\_of\_the\_cyberstalker=Yes 16 <conf:(0.84)> lift:(2.03) lev:(0.06) [8] conv:(2.78)

263. Victim\_of\_cyberstalking=Yes Cyberstalking\_achieving\_goal=No 19 ==>  
 Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=Yes 16 <conf:(0.84)> lift:(3.86) lev:(0.08)  
 [11] conv:(3.71)

264. You\_harassed\_someone=Yes Cyberstalking\_achieving\_goal=No Cyberstalking\_pleasure=No 19 ==>  
 Victim\_of\_cyberstalking=Yes 16 <conf:(0.84)> lift:(1.81) lev:(0.05) [7] conv:(2.54)

265. Victim\_of\_cyberstalking=Yes You\_harassed\_someone=Yes Cyberstalking\_achieving\_goal=No 19  
 ==> Cyberstalking\_pleasure=No 16 <conf:(0.84)> lift:(3.62) lev:(0.08) [11] conv:(3.65)

266. Cyberstalking\_achieving\_goal=No Cyberstalking\_pleasure=No 19 ==> Victim\_of\_cyberstalking=Yes  
 You\_harassed\_someone=Yes 16 <conf:(0.84)> lift:(3.23) lev:(0.08) [11] conv:(3.51)

267. Victim\_of\_cyberstalking=Yes Cyberstalking\_achieving\_goal=No 19 ==>  
 You\_harassed\_someone=Yes Cyberstalking\_pleasure=No 16 <conf:(0.84)> lift:(3.62) lev:(0.08) [11]  
 conv:(3.65)

268. Victim\_of\_cyberstalking=Yes Social\_media\_communication=No 25 ==>  
 Get\_rid\_of\_the\_cyberstalker=Yes 21 <conf:(0.84)> lift:(2.02) lev:(0.07) [10] conv:(2.92)

269. Gender=Male Victim\_of\_cyberstalking=No Criminal\_offense=Yes 31 ==>  
 You\_harassed\_someone=No 26 <conf:(0.84)> lift:(1.43) lev:(0.06) [7] conv:(2.15)

270. Victim\_of\_cyberstalking=Yes You\_harassed\_someone=Yes 37 ==> Get\_rid\_of\_the\_cyberstalker=Yes  
 31 <conf:(0.84)> lift:(2.02) lev:(0.11) [15] conv:(3.09)

271. Know\_the\_stalker=No 36 ==> Get\_rid\_of\_the\_cyberstalker=Yes 30 <conf:(0.83)> lift:(2.01)  
 lev:(0.11) [15] conv:(3.01)

272. Know\_the\_stalker=No 36 ==> Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes 30  
 <conf:(0.83)> lift:(2.01) lev:(0.11) [15] conv:(3.01)

273. School=Kiril Pejcinoviq Victim\_of\_cyberstalking=No Criminal\_offense=Yes 24 ==>  
 You\_harassed\_someone=No 20 <conf:(0.83)> lift:(1.43) lev:(0.04) [5] conv:(1.99)

274. School=Nikola Shtejn Victim\_of\_cyberstalking=Yes 18 ==> Get\_rid\_of\_the\_cyberstalker=Yes 15  
 <conf:(0.83)> lift:(2.01) lev:(0.05) [7] conv:(2.63)

275. School=Nikola Shtejn Victim\_of\_cyberstalking=Yes 18 ==> Criminal\_offense=Yes 15 <conf:(0.83)>  
 lift:(1.3) lev:(0.02) [3] conv:(1.62)

276. Victim\_of\_cyberstalking=Yes Cyberstalking\_achieving\_goal=Yes 18 ==>  
 Get\_rid\_of\_the\_cyberstalker=Yes 15 <conf:(0.83)> lift:(2.01) lev:(0.05) [7] conv:(2.63)

277. Social\_media\_cyberstalking=Instagram You\_harassed\_someone=No 18 ==>  
 Social\_media\_communication=Yes 15 <conf:(0.83)> lift:(2.89) lev:(0.07) [9] conv:(3.2)

278. Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram You\_harassed\_someone=No 18 ==> Social\_media\_communication=Yes 15 <conf:(0.83)> lift:(2.89) lev:(0.07) [9] conv:(3.2)

279. Social\_media\_cyberstalking=Instagram You\_harassed\_someone=No 18 ==> Victim\_of\_cyberstalking=Yes Social\_media\_communication=Yes 15 <conf:(0.83)> lift:(2.89) lev:(0.07) [9] conv:(3.2)

280. Victim\_of\_cyberstalking=Yes You\_harassed\_someone=Yes Cyberstalking\_achieving\_goal=Yes 18 ==> Get\_rid\_of\_the\_cyberstalker=Yes 15 <conf:(0.83)> lift:(2.01) lev:(0.05) [7] conv:(2.63)

281. Victim\_of\_cyberstalking=Yes Cyberstalking\_achieving\_goal=Yes 18 ==> Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=Yes 15 <conf:(0.83)> lift:(3.82) lev:(0.08) [11] conv:(3.52)

282. Social\_media\_cyberstalking=Instagram Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=No 18 ==> Social\_media\_communication=Yes 15 <conf:(0.83)> lift:(2.89) lev:(0.07) [9] conv:(3.2)

283. Social\_media\_cyberstalking=Instagram You\_harassed\_someone=No 18 ==> Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes 15 <conf:(0.83)> lift:(3.11) lev:(0.07) [10] conv:(3.3)

284. School=Kiril Pejcinoviq Gender=Male You\_harassed\_someone=No Criminal\_offense=Yes 18 ==> Victim\_of\_cyberstalking=No 15 <conf:(0.83)> lift:(1.56) lev:(0.04) [5] conv:(2.09)

285. School=Kiril Pejcinoviq Gender=Male Victim\_of\_cyberstalking=No You\_harassed\_someone=No 18 ==> Criminal\_offense=Yes 15 <conf:(0.83)> lift:(1.3) lev:(0.02) [3] conv:(1.62)

286. Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=No 18 ==> Social\_media\_communication=Yes 15 <conf:(0.83)> lift:(2.89) lev:(0.07) [9] conv:(3.2)

287. Social\_media\_cyberstalking=Instagram Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=No 18 ==> Victim\_of\_cyberstalking=Yes Social\_media\_communication=Yes 15 <conf:(0.83)> lift:(2.89) lev:(0.07) [9] conv:(3.2)

288. Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram You\_harassed\_someone=No 18 ==> Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes 15 <conf:(0.83)> lift:(3.11) lev:(0.07) [10] conv:(3.3)

289. Social\_media\_cyberstalking=Instagram You\_harassed\_someone=No 18 ==> Victim\_of\_cyberstalking=Yes Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes 15 <conf:(0.83)> lift:(3.11) lev:(0.07) [10] conv:(3.3)

290. Victim\_of\_cyberstalking=Yes You\_harassed\_someone=Yes Criminal\_offense=Yes 23 ==> Get\_rid\_of\_the\_cyberstalker=Yes 19 <conf:(0.83)> lift:(1.99) lev:(0.07) [9] conv:(2.69)

291. Gender=Female Social\_media\_communication=No 17 ==> Get\_rid\_of\_the\_cyberstalker=Yes 14  
<conf:(0.82)> lift:(1.98) lev:(0.05) [6] conv:(2.48)

292. Social\_media\_cyberstalking=Instagram Know\_the\_stalker=No 17 ==>  
Social\_media\_communication=Yes 14 <conf:(0.82)> lift:(2.85) lev:(0.06) [9] conv:(3.02)

293. Gender=Female Victim\_of\_cyberstalking=Yes You\_harassed\_someone=No 17 ==> School=Kiril  
Pejcinoviq 14 <conf:(0.82)> lift:(1.75) lev:(0.04) [5] conv:(2.24)

294. Gender=Female Victim\_of\_cyberstalking=Yes Social\_media\_communication=No 17 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(0.82)> lift:(1.98) lev:(0.05) [6] conv:(2.48)

295. Gender=Female Social\_media\_communication=No 17 ==> Victim\_of\_cyberstalking=Yes  
Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(0.82)> lift:(1.98) lev:(0.05) [6] conv:(2.48)

296. Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram Know\_the\_stalker=No 17 ==>  
Social\_media\_communication=Yes 14 <conf:(0.82)> lift:(2.85) lev:(0.06) [9] conv:(3.02)

297. Social\_media\_cyberstalking=Instagram Know\_the\_stalker=No 17 ==> Victim\_of\_cyberstalking=Yes  
Social\_media\_communication=Yes 14 <conf:(0.82)> lift:(2.85) lev:(0.06) [9] conv:(3.02)

298. Victim\_of\_cyberstalking=Yes Social\_media\_communication=No Criminal\_offense=Yes 17 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(0.82)> lift:(1.98) lev:(0.05) [6] conv:(2.48)

299. Gender=Male You\_harassed\_someone=No 44 ==> Criminal\_offense=Yes 36 <conf:(0.82)>  
lift:(1.28) lev:(0.05) [7] conv:(1.76)

300. Victim\_of\_cyberstalking=Yes Cyberstalking\_pleasure=No 22 ==> Get\_rid\_of\_the\_cyberstalker=Yes  
18 <conf:(0.82)> lift:(1.97) lev:(0.06) [8] conv:(2.57)

301. School=Kiril Pejcinoviq Gender=Male You\_harassed\_someone=No 22 ==>  
Victim\_of\_cyberstalking=No 18 <conf:(0.82)> lift:(1.53) lev:(0.04) [6] conv:(2.05)

302. School=Kiril Pejcinoviq Gender=Male Victim\_of\_cyberstalking=No 22 ==>  
You\_harassed\_someone=No 18 <conf:(0.82)> lift:(1.4) lev:(0.04) [5] conv:(1.83)

303. School=Kiril Pejcinoviq Gender=Male You\_harassed\_someone=No 22 ==> Criminal\_offense=Yes 18  
<conf:(0.82)> lift:(1.28) lev:(0.03) [3] conv:(1.58)

304. Victim\_of\_cyberstalking=Yes You\_harassed\_someone=Yes Cyberstalking\_pleasure=No 22 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes 18 <conf:(0.82)> lift:(1.97) lev:(0.06) [8] conv:(2.57)

305. Victim\_of\_cyberstalking=Yes Cyberstalking\_pleasure=No 22 ==> Get\_rid\_of\_the\_cyberstalker=Yes  
You\_harassed\_someone=Yes 18 <conf:(0.82)> lift:(3.75) lev:(0.09) [13] conv:(3.44)

306. Gender=Male Victim\_of\_cyberstalking=No You\_harassed\_someone=No 32 ==>  
Criminal\_offense=Yes 26 <conf:(0.81)> lift:(1.27) lev:(0.04) [5] conv:(1.64)

307. Social\_media\_communication=No 26 ==> Get\_rid\_of\_the\_cyberstalker=Yes 21 <conf:(0.81)> lift:(1.94) lev:(0.07) [10] conv:(2.53)

308. Social\_media\_communication=No 26 ==> Victim\_of\_cyberstalking=Yes  
Get\_rid\_of\_the\_cyberstalker=Yes 21 <conf:(0.81)> lift:(1.94) lev:(0.07) [10] conv:(2.53)

309. School=Kiril Pejcinoviq Criminal\_offense=Yes 40 ==> You\_harassed\_someone=No 32 <conf:(0.8)> lift:(1.37) lev:(0.06) [8] conv:(1.85)

310. School=Kiril Pejcinoviq Gender=Male Criminal\_offense=Yes 20 ==> Victim\_of\_cyberstalking=No 16  
<conf:(0.8)> lift:(1.49) lev:(0.04) [5] conv:(1.86)

311. Social\_media\_communication=Yes You\_harassed\_someone=No 19 ==>  
Social\_media\_cyberstalking=Instagram 15 <conf:(0.79)> lift:(3.62) lev:(0.08) [10] conv:(2.97)

312. Victim\_of\_cyberstalking=Yes Social\_media\_communication=Yes You\_harassed\_someone=No 19  
==> Social\_media\_cyberstalking=Instagram 15 <conf:(0.79)> lift:(3.62) lev:(0.08) [10] conv:(2.97)

313. Social\_media\_communication=Yes You\_harassed\_someone=No 19 ==>  
Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram 15 <conf:(0.79)> lift:(3.62)  
lev:(0.08) [10] conv:(2.97)

314. Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=No  
19 ==> Social\_media\_cyberstalking=Instagram 15 <conf:(0.79)> lift:(3.62) lev:(0.08) [10] conv:(2.97)

315. Social\_media\_communication=Yes You\_harassed\_someone=No 19 ==>  
Social\_media\_cyberstalking=Instagram Get\_rid\_of\_the\_cyberstalker=Yes 15 <conf:(0.79)> lift:(3.74)  
lev:(0.08) [10] conv:(3)

316. Victim\_of\_cyberstalking=Yes Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes  
You\_harassed\_someone=No 19 ==> Social\_media\_cyberstalking=Instagram 15 <conf:(0.79)> lift:(3.62)  
lev:(0.08) [10] conv:(2.97)

317. Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=No  
19 ==> Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram 15 <conf:(0.79)> lift:(3.62)  
lev:(0.08) [10] conv:(2.97)

318. Victim\_of\_cyberstalking=Yes Social\_media\_communication=Yes You\_harassed\_someone=No 19  
==> Social\_media\_cyberstalking=Instagram Get\_rid\_of\_the\_cyberstalker=Yes 15 <conf:(0.79)>  
lift:(3.74) lev:(0.08) [10] conv:(3)

319. Social\_media\_communication=Yes You\_harassed\_someone=No 19 ==>  
Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram Get\_rid\_of\_the\_cyberstalker=Yes  
15 <conf:(0.79)> lift:(3.74) lev:(0.08) [10] conv:(3)

320. School=Kiril Pejcinoviq Gender=Male 28 ==> Victim\_of\_cyberstalking=No 22 <conf:(0.79)>  
lift:(1.47) lev:(0.05) [7] conv:(1.86)

321. School=Kiril Pejcinoviq Gender=Male 28 ==> You\_harassed\_someone=No 22 <conf:(0.79)> lift:(1.34) lev:(0.04) [5] conv:(1.66)

322. School=Nikola Shtejn Gender=Male 23 ==> Criminal\_offense=Yes 18 <conf:(0.78)> lift:(1.22) lev:(0.02) [3] conv:(1.38)

323. School=Nikola Shtejn You\_harassed\_someone=Yes 23 ==> Cyberstalking\_achieving\_goal=Yes 18 <conf:(0.78)> lift:(3) lev:(0.08) [12] conv:(2.83)

324. School=Nikola Shtejn You\_harassed\_someone=Yes 23 ==> Criminal\_offense=Yes 18 <conf:(0.78)> lift:(1.22) lev:(0.02) [3] conv:(1.38)

325. Victim\_of\_cyberstalking=Yes Know\_the\_stalker=No Criminal\_offense=Yes 23 ==> Get\_rid\_of\_the\_cyberstalker=Yes 18 <conf:(0.78)> lift:(1.88) lev:(0.06) [8] conv:(2.24)

326. Cyberstalking\_purpose=For gossip in group chats 18 ==> Gender=Female 14 <conf:(0.78)> lift:(1.51) lev:(0.03) [4] conv:(1.75)

327. Form\_of\_harassment=Any other behavior founded distressing in any way 18 ==> Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(0.78)> lift:(1.87) lev:(0.05) [6] conv:(2.1)

328. Cyberstalking\_purpose=For gossip in group chats 18 ==> Cyberstalking\_achieving\_goal=Yes 14 <conf:(0.78)> lift:(2.98) lev:(0.07) [9] conv:(2.66)

329. Gender=Female Social\_media\_cyberstalking=Instagram 18 ==> Criminal\_offense=Yes 14 <conf:(0.78)> lift:(1.21) lev:(0.02) [2] conv:(1.29)

330. You\_harassed\_someone=Yes Cyberstalking\_purpose=For gossip in group chats 18 ==> Gender=Female 14 <conf:(0.78)> lift:(1.51) lev:(0.03) [4] conv:(1.75)

331. Cyberstalking\_purpose=For gossip in group chats 18 ==> Gender=Female You\_harassed\_someone=Yes 14 <conf:(0.78)> lift:(3.25) lev:(0.07) [9] conv:(2.74)

332. Victim\_of\_cyberstalking=Yes Form\_of\_harassment=Any other behavior founded distressing in any way 18 ==> Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(0.78)> lift:(1.87) lev:(0.05) [6] conv:(2.1)

333. Form\_of\_harassment=Any other behavior founded distressing in any way 18 ==> Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(0.78)> lift:(1.87) lev:(0.05) [6] conv:(2.1)

334. Social\_media\_communication=No Criminal\_offense=Yes 18 ==> Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(0.78)> lift:(1.87) lev:(0.05) [6] conv:(2.1)

335. You\_harassed\_someone=Yes Cyberstalking\_purpose=For gossip in group chats 18 ==> Cyberstalking\_achieving\_goal=Yes 14 <conf:(0.78)> lift:(2.98) lev:(0.07) [9] conv:(2.66)

336. Cyberstalking\_purpose=For gossip in group chats 18 ==> You\_harassed\_someone=Yes Cyberstalking\_achieving\_goal=Yes 14 <conf:(0.78)> lift:(2.98) lev:(0.07) [9] conv:(2.66)

337. School=Kiril Pejcinoviq Victim\_of\_cyberstalking=Yes You\_harassed\_someone=No 18 ==>  
Gender=Female 14 <conf:(0.78)> lift:(1.51) lev:(0.03) [4] conv:(1.75)

338. Gender=Female Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram 18 ==>  
Criminal\_offense=Yes 14 <conf:(0.78)> lift:(1.21) lev:(0.02) [2] conv:(1.29)

339. Gender=Female Social\_media\_cyberstalking=Instagram 18 ==> Victim\_of\_cyberstalking=Yes  
Criminal\_offense=Yes 14 <conf:(0.78)> lift:(2.57) lev:(0.06) [8] conv:(2.51)

340. Gender=Female Social\_media\_cyberstalking=Instagram Get\_rid\_of\_the\_cyberstalker=Yes 18 ==>  
Criminal\_offense=Yes 14 <conf:(0.78)> lift:(1.21) lev:(0.02) [2] conv:(1.29)

341. Gender=Female Social\_media\_cyberstalking=Instagram 18 ==> Get\_rid\_of\_the\_cyberstalker=Yes  
Criminal\_offense=Yes 14 <conf:(0.78)> lift:(2.91) lev:(0.06) [9] conv:(2.64)

342. Social\_media\_communication=No Criminal\_offense=Yes 18 ==> Victim\_of\_cyberstalking=Yes  
Get\_rid\_of\_the\_cyberstalker=Yes 14 <conf:(0.78)> lift:(1.87) lev:(0.05) [6] conv:(2.1)

343. Gender=Female Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram  
Get\_rid\_of\_the\_cyberstalker=Yes 18 ==> Criminal\_offense=Yes 14 <conf:(0.78)> lift:(1.21) lev:(0.02) [2]  
conv:(1.29)

344. Gender=Female Social\_media\_cyberstalking=Instagram Get\_rid\_of\_the\_cyberstalker=Yes 18 ==>  
Victim\_of\_cyberstalking=Yes Criminal\_offense=Yes 14 <conf:(0.78)> lift:(2.57) lev:(0.06) [8] conv:(2.51)

345. Gender=Female Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram 18 ==>  
Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=Yes 14 <conf:(0.78)> lift:(2.91) lev:(0.06) [9]  
conv:(2.64)

346. Gender=Female Social\_media\_cyberstalking=Instagram 18 ==> Victim\_of\_cyberstalking=Yes  
Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=Yes 14 <conf:(0.78)> lift:(2.91) lev:(0.06) [9]  
conv:(2.64)

347. School=Kiril Pejcinoviq Get\_rid\_of\_the\_cyberstalker=Yes 22 ==> Gender=Female 17 <conf:(0.77)>  
lift:(1.5) lev:(0.04) [5] conv:(1.78)

348. School=Kiril Pejcinoviq Get\_rid\_of\_the\_cyberstalker=Yes 22 ==> Social\_media\_communication=Yes  
17 <conf:(0.77)> lift:(2.68) lev:(0.07) [10] conv:(2.61)

349. School=Kiril Pejcinoviq Get\_rid\_of\_the\_cyberstalker=Yes 22 ==> You\_harassed\_someone=No 17  
<conf:(0.77)> lift:(1.32) lev:(0.03) [4] conv:(1.52)

350. School=Kiril Pejcinoviq Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes 22 ==>  
Gender=Female 17 <conf:(0.77)> lift:(1.5) lev:(0.04) [5] conv:(1.78)

351. School=Kiril Pejcinoviq Get\_rid\_of\_the\_cyberstalker=Yes 22 ==> Gender=Female  
Victim\_of\_cyberstalking=Yes 17 <conf:(0.77)> lift:(2.74) lev:(0.08) [10] conv:(2.63)

352. School=Kiril Pejcinoviq Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes 22 ==> Social\_media\_communication=Yes 17 <conf:(0.77)> lift:(2.68) lev:(0.07) [10] conv:(2.61)

353. School=Kiril Pejcinoviq Get\_rid\_of\_the\_cyberstalker=Yes 22 ==> Victim\_of\_cyberstalking=Yes Social\_media\_communication=Yes 17 <conf:(0.77)> lift:(2.68) lev:(0.07) [10] conv:(2.61)

354. School=Kiril Pejcinoviq Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes 22 ==> You\_harassed\_someone=No 17 <conf:(0.77)> lift:(1.32) lev:(0.03) [4] conv:(1.52)

355. School=Kiril Pejcinoviq Get\_rid\_of\_the\_cyberstalker=Yes 22 ==> Victim\_of\_cyberstalking=Yes You\_harassed\_someone=No 17 <conf:(0.77)> lift:(3.78) lev:(0.09) [12] conv:(2.92)

356. School=Kiril Pejcinoviq Victim\_of\_cyberstalking=Yes 26 ==> Gender=Female 20 <conf:(0.77)> lift:(1.5) lev:(0.05) [6] conv:(1.8)

357. School=Kiril Pejcinoviq Victim\_of\_cyberstalking=No 41 ==> You\_harassed\_someone=No 31 <conf:(0.76)> lift:(1.29) lev:(0.05) [7] conv:(1.55)

358. Gender=Male Criminal\_offense=Yes 48 ==> You\_harassed\_someone=No 36 <conf:(0.75)> lift:(1.28) lev:(0.06) [7] conv:(1.53)

359. Victim\_of\_cyberstalking=No Criminal\_offense=Yes 48 ==> You\_harassed\_someone=No 36 <conf:(0.75)> lift:(1.28) lev:(0.06) [7] conv:(1.53)

360. Know\_the\_stalker=No Criminal\_offense=Yes 24 ==> Get\_rid\_of\_the\_cyberstalker=Yes 18 <conf:(0.75)> lift:(1.81) lev:(0.06) [8] conv:(2)

361. Know\_the\_stalker=No Criminal\_offense=Yes 24 ==> Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes 18 <conf:(0.75)> lift:(1.81) lev:(0.06) [8] conv:(2)

362. School=Nikola Shtejn You\_harassed\_someone=No 20 ==> Victim\_of\_cyberstalking=No 15 <conf:(0.75)> lift:(1.4) lev:(0.03) [4] conv:(1.55)

363. School=Kiril Pejcinoviq Victim\_of\_cyberstalking=No You\_harassed\_someone=No Criminal\_offense=Yes 20 ==> Gender=Male 15 <conf:(0.75)> lift:(1.54) lev:(0.04) [5] conv:(1.71)

364. School=Kiril Pejcinoviq Gender=Male Criminal\_offense=Yes 20 ==> Victim\_of\_cyberstalking=No You\_harassed\_someone=No 15 <conf:(0.75)> lift:(1.97) lev:(0.05) [7] conv:(2.07)

365. School=Nikola Shtejn 43 ==> Criminal\_offense=Yes 32 <conf:(0.74)> lift:(1.16) lev:(0.03) [4] conv:(1.29)

366. Gender=Male Victim\_of\_cyberstalking=No 43 ==> You\_harassed\_someone=No 32 <conf:(0.74)> lift:(1.27) lev:(0.05) [6] conv:(1.49)

367. Social\_media\_cyberstalking=Instagram 31 ==> Social\_media\_communication=Yes 23 <conf:(0.74)> lift:(2.57) lev:(0.1) [14] conv:(2.45)



368. Social\_media\_cyberstalking=Instagram 31 ==> Criminal\_offense=Yes 23 <conf:(0.74)> lift:(1.16) lev:(0.02) [3] conv:(1.24)

369. Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram 31 ==> Social\_media\_communication=Yes 23 <conf:(0.74)> lift:(2.57) lev:(0.1) [14] conv:(2.45)

370. Social\_media\_cyberstalking=Instagram 31 ==> Victim\_of\_cyberstalking=Yes Social\_media\_communication=Yes 23 <conf:(0.74)> lift:(2.57) lev:(0.1) [14] conv:(2.45)

371. Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram 31 ==> Criminal\_offense=Yes 23 <conf:(0.74)> lift:(1.16) lev:(0.02) [3] conv:(1.24)

372. Social\_media\_cyberstalking=Instagram 31 ==> Victim\_of\_cyberstalking=Yes Criminal\_offense=Yes 23 <conf:(0.74)> lift:(2.45) lev:(0.1) [13] conv:(2.4)

373. Cyberstalking\_purpose=Other 19 ==> Cyberstalking\_pleasure=No 14 <conf:(0.74)> lift:(3.17) lev:(0.07) [9] conv:(2.43)

374. School=Kiril Pejcinoviq Social\_media\_communication=Yes 19 ==> Gender=Female 14 <conf:(0.74)> lift:(1.43) lev:(0.03) [4] conv:(1.54)

375. You\_harassed\_someone=Yes Cyberstalking\_purpose=Other 19 ==> Cyberstalking\_pleasure=No 14 <conf:(0.74)> lift:(3.17) lev:(0.07) [9] conv:(2.43)

376. Cyberstalking\_purpose=Other 19 ==> You\_harassed\_someone=Yes Cyberstalking\_pleasure=No 14 <conf:(0.74)> lift:(3.17) lev:(0.07) [9] conv:(2.43)

377. School=Kiril Pejcinoviq Victim\_of\_cyberstalking=Yes Social\_media\_communication=Yes 19 ==> Gender=Female 14 <conf:(0.74)> lift:(1.43) lev:(0.03) [4] conv:(1.54)

378. School=Kiril Pejcinoviq Social\_media\_communication=Yes 19 ==> Gender=Female Victim\_of\_cyberstalking=Yes 14 <conf:(0.74)> lift:(2.62) lev:(0.06) [8] conv:(2.27)

379. Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=Yes Criminal\_offense=Yes 19 ==> Gender=Female 14 <conf:(0.74)> lift:(1.43) lev:(0.03) [4] conv:(1.54)

380. Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=Yes Criminal\_offense=Yes 19 ==> Gender=Female 14 <conf:(0.74)> lift:(1.43) lev:(0.03) [4] conv:(1.54)

381. Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=Yes Criminal\_offense=Yes 19 ==> Gender=Female Victim\_of\_cyberstalking=Yes 14 <conf:(0.74)> lift:(2.62) lev:(0.06) [8] conv:(2.27)

382. Social\_media\_cyberstalking=Instagram Get\_rid\_of\_the\_cyberstalker=Yes 30 ==> Social\_media\_communication=Yes 22 <conf:(0.73)> lift:(2.54) lev:(0.09) [13] conv:(2.37)

383. Social\_media\_cyberstalking=Instagram Get\_rid\_of\_the\_cyberstalker=Yes 30 ==> Criminal\_offense=Yes 22 <conf:(0.73)> lift:(1.14) lev:(0.02) [2] conv:(1.2)

384. Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram  
 Get\_rid\_of\_the\_cyberstalker=Yes 30 ==> Social\_media\_communication=Yes 22 <conf:(0.73)> lift:(2.54)  
 lev:(0.09) [13] conv:(2.37)

385. Social\_media\_cyberstalking=Instagram Get\_rid\_of\_the\_cyberstalker=Yes 30 ==>  
 Victim\_of\_cyberstalking=Yes Social\_media\_communication=Yes 22 <conf:(0.73)> lift:(2.54) lev:(0.09)  
 [13] conv:(2.37)

386. Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram  
 Get\_rid\_of\_the\_cyberstalker=Yes 30 ==> Criminal\_offense=Yes 22 <conf:(0.73)> lift:(1.14) lev:(0.02) [2]  
 conv:(1.2)

387. Social\_media\_cyberstalking=Instagram Get\_rid\_of\_the\_cyberstalker=Yes 30 ==>  
 Victim\_of\_cyberstalking=Yes Criminal\_offense=Yes 22 <conf:(0.73)> lift:(2.42) lev:(0.09) [12]  
 conv:(2.32)

388. School=Kiril Pejcinoviq 67 ==> You\_harassed\_someone=No 49 <conf:(0.73)> lift:(1.25) lev:(0.07)  
 [9] conv:(1.47)

389. School=Kiril Pejcinoviq Victim\_of\_cyberstalking=Yes 26 ==> Social\_media\_communication=Yes 19  
 <conf:(0.73)> lift:(2.53) lev:(0.08) [11] conv:(2.31)

390. Gender=Male You\_harassed\_someone=No 44 ==> Victim\_of\_cyberstalking=No 32 <conf:(0.73)>  
 lift:(1.36) lev:(0.06) [8] conv:(1.57)

391. Cyberstalking\_achieving\_goal=No 22 ==> Get\_rid\_of\_the\_cyberstalker=Yes 16 <conf:(0.73)>  
 lift:(1.75) lev:(0.05) [6] conv:(1.84)

392. School=7Marsi Victim\_of\_cyberstalking=Yes 22 ==> You\_harassed\_someone=Yes 16 <conf:(0.73)>  
 lift:(1.75) lev:(0.05) [6] conv:(1.84)

393. School=7Marsi Get\_rid\_of\_the\_cyberstalker=Yes 22 ==> You\_harassed\_someone=Yes 16  
 <conf:(0.73)> lift:(1.75) lev:(0.05) [6] conv:(1.84)

394. Cyberstalking\_achieving\_goal=No 22 ==> Victim\_of\_cyberstalking=Yes  
 Get\_rid\_of\_the\_cyberstalker=Yes 16 <conf:(0.73)> lift:(1.75) lev:(0.05) [6] conv:(1.84)

395. Victim\_of\_cyberstalking=Yes Cyberstalking\_pleasure=No 22 ==> Cyberstalking\_achieving\_goal=No  
 16 <conf:(0.73)> lift:(4.69) lev:(0.09) [12] conv:(2.66)

396. Cyberstalking\_achieving\_goal=No 22 ==> Victim\_of\_cyberstalking=Yes Cyberstalking\_pleasure=No  
 16 <conf:(0.73)> lift:(4.69) lev:(0.09) [12] conv:(2.66)

397. You\_harassed\_someone=Yes Cyberstalking\_achieving\_goal=No 22 ==>  
 Get\_rid\_of\_the\_cyberstalker=Yes 16 <conf:(0.73)> lift:(1.75) lev:(0.05) [6] conv:(1.84)

398. Cyberstalking\_achieving\_goal=No 22 ==> Get\_rid\_of\_the\_cyberstalker=Yes  
 You\_harassed\_someone=Yes 16 <conf:(0.73)> lift:(3.33) lev:(0.08) [11] conv:(2.46)

399. School=Kiril Pejcinoviq Gender=Male Victim\_of\_cyberstalking=No 22 ==> Criminal\_offense=Yes 16  
<conf:(0.73)> lift:(1.13) lev:(0.01) [1] conv:(1.13)

400. School=7Marsi Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes 22 ==>  
You\_harassed\_someone=Yes 16 <conf:(0.73)> lift:(1.75) lev:(0.05) [6] conv:(1.84)

401. School=7Marsi Get\_rid\_of\_the\_cyberstalker=Yes 22 ==> Victim\_of\_cyberstalking=Yes  
You\_harassed\_someone=Yes 16 <conf:(0.73)> lift:(2.79) lev:(0.07) [10] conv:(2.32)

402. School=7Marsi Victim\_of\_cyberstalking=Yes 22 ==> Get\_rid\_of\_the\_cyberstalker=Yes  
You\_harassed\_someone=Yes 16 <conf:(0.73)> lift:(3.33) lev:(0.08) [11] conv:(2.46)

403. You\_harassed\_someone=Yes Cyberstalking\_achieving\_goal=No 22 ==>  
Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes 16 <conf:(0.73)> lift:(1.75) lev:(0.05) [6]  
conv:(1.84)

404. Cyberstalking\_achieving\_goal=No 22 ==> Victim\_of\_cyberstalking=Yes  
Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=Yes 16 <conf:(0.73)> lift:(3.33) lev:(0.08)  
[11] conv:(2.46)

405. Victim\_of\_cyberstalking=Yes You\_harassed\_someone=Yes Cyberstalking\_pleasure=No 22 ==>  
Cyberstalking\_achieving\_goal=No 16 <conf:(0.73)> lift:(4.69) lev:(0.09) [12] conv:(2.66)

406. You\_harassed\_someone=Yes Cyberstalking\_achieving\_goal=No 22 ==>  
Victim\_of\_cyberstalking=Yes Cyberstalking\_pleasure=No 16 <conf:(0.73)> lift:(4.69) lev:(0.09) [12]  
conv:(2.66)

407. Victim\_of\_cyberstalking=Yes Cyberstalking\_pleasure=No 22 ==> You\_harassed\_someone=Yes  
Cyberstalking\_achieving\_goal=No 16 <conf:(0.73)> lift:(4.69) lev:(0.09) [12] conv:(2.66)

408. Cyberstalking\_achieving\_goal=No 22 ==> Victim\_of\_cyberstalking=Yes  
You\_harassed\_someone=Yes Cyberstalking\_pleasure=No 16 <conf:(0.73)> lift:(4.69) lev:(0.09) [12]  
conv:(2.66)

409. Victim\_of\_cyberstalking=No You\_harassed\_someone=No Criminal\_offense=Yes 36 ==>  
Gender=Male 26 <conf:(0.72)> lift:(1.49) lev:(0.06) [8] conv:(1.68)

410. Gender=Male You\_harassed\_someone=No Criminal\_offense=Yes 36 ==>  
Victim\_of\_cyberstalking=No 26 <conf:(0.72)> lift:(1.35) lev:(0.05) [6] conv:(1.52)

411. Gender=Male Victim\_of\_cyberstalking=No 43 ==> Criminal\_offense=Yes 31 <conf:(0.72)>  
lift:(1.12) lev:(0.02) [3] conv:(1.19)

412. School=Kiril Pejcinoviq Gender=Male 28 ==> Criminal\_offense=Yes 20 <conf:(0.71)> lift:(1.11)  
lev:(0.01) [2] conv:(1.12)

413. Cyberstalking\_achieving\_goal=Yes Criminal\_offense=Yes 21 ==> Gender=Female 15 <conf:(0.71)>  
lift:(1.39) lev:(0.03) [4] conv:(1.46)

414. Gender=Female Cyberstalking\_achieving\_goal=Yes 21 ==> Criminal\_offense=Yes 15 <conf:(0.71)> lift:(1.11) lev:(0.01) [1] conv:(1.08)

415. You\_harassed\_someone=Yes Cyberstalking\_achieving\_goal=Yes Criminal\_offense=Yes 21 ==> Gender=Female 15 <conf:(0.71)> lift:(1.39) lev:(0.03) [4] conv:(1.46)

416. Gender=Female You\_harassed\_someone=Yes Cyberstalking\_achieving\_goal=Yes 21 ==> Criminal\_offense=Yes 15 <conf:(0.71)> lift:(1.11) lev:(0.01) [1] conv:(1.08)

417. Cyberstalking\_achieving\_goal=Yes Criminal\_offense=Yes 21 ==> Gender=Female You\_harassed\_someone=Yes 15 <conf:(0.71)> lift:(2.98) lev:(0.07) [9] conv:(2.28)

418. Gender=Female Cyberstalking\_achieving\_goal=Yes 21 ==> You\_harassed\_someone=Yes Criminal\_offense=Yes 15 <conf:(0.71)> lift:(2.9) lev:(0.07) [9] conv:(2.26)

419. Victim\_of\_cyberstalking=No 76 ==> You\_harassed\_someone=No 54 <conf:(0.71)> lift:(1.22) lev:(0.07) [9] conv:(1.37)

420. Social\_media\_cyberstalking=Instagram 31 ==> Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes 22 <conf:(0.71)> lift:(2.65) lev:(0.1) [13] conv:(2.27)

421. Social\_media\_cyberstalking=Instagram 31 ==> Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=Yes 22 <conf:(0.71)> lift:(2.65) lev:(0.1) [13] conv:(2.27)

422. Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram 31 ==> Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes 22 <conf:(0.71)> lift:(2.65) lev:(0.1) [13] conv:(2.27)

423. Social\_media\_cyberstalking=Instagram 31 ==> Victim\_of\_cyberstalking=Yes Social\_media\_communication=Yes Get\_rid\_of\_the\_cyberstalker=Yes 22 <conf:(0.71)> lift:(2.65) lev:(0.1) [13] conv:(2.27)

424. Victim\_of\_cyberstalking=Yes Social\_media\_cyberstalking=Instagram 31 ==> Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=Yes 22 <conf:(0.71)> lift:(2.65) lev:(0.1) [13] conv:(2.27)

425. Social\_media\_cyberstalking=Instagram 31 ==> Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=Yes 22 <conf:(0.71)> lift:(2.65) lev:(0.1) [13] conv:(2.27)

426. School=Nikola Shtejn Gender=Female 20 ==> Criminal\_offense=Yes 14 <conf:(0.7)> lift:(1.09) lev:(0.01) [1] conv:(1.03)

427. School=Nikola Shtejn You\_harassed\_someone=No 20 ==> Criminal\_offense=Yes 14 <conf:(0.7)> lift:(1.09) lev:(0.01) [1] conv:(1.03)

428. Cyberstalking\_pleasure=No Criminal\_offense=Yes 20 ==> Gender=Female 14 <conf:(0.7)> lift:(1.36) lev:(0.03) [3] conv:(1.39)

429. Cyberstalking\_pleasure=No Criminal\_offense=Yes 20 ==> Victim\_of\_cyberstalking=Yes 14  
<conf:(0.7)> lift:(1.51) lev:(0.03) [4] conv:(1.53)

430. Know\_the\_stalker=No Social\_media\_communication=Yes 20 ==>  
Social\_media\_cyberstalking=Instagram 14 <conf:(0.7)> lift:(3.21) lev:(0.07) [9] conv:(2.23)

431. Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=No 20 ==> Social\_media\_communication=Yes  
14 <conf:(0.7)> lift:(2.42) lev:(0.06) [8] conv:(2.03)

432. School=Kiril Pejcinoviq Gender=Female Victim\_of\_cyberstalking=Yes 20 ==>  
Social\_media\_communication=Yes 14 <conf:(0.7)> lift:(2.42) lev:(0.06) [8] conv:(2.03)

433. School=Kiril Pejcinoviq Gender=Female Victim\_of\_cyberstalking=Yes 20 ==>  
You\_harassed\_someone=No 14 <conf:(0.7)> lift:(1.2) lev:(0.02) [2] conv:(1.19)

434. Gender=Female You\_harassed\_someone=No Criminal\_offense=Yes 20 ==> School=Kiril Pejcinoviq  
14 <conf:(0.7)> lift:(1.48) lev:(0.03) [4] conv:(1.51)

435. School=Kiril Pejcinoviq Gender=Female Criminal\_offense=Yes 20 ==> You\_harassed\_someone=No  
14 <conf:(0.7)> lift:(1.2) lev:(0.02) [2] conv:(1.19)

436. Gender=Female Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=Yes 20 ==>  
Criminal\_offense=Yes 14 <conf:(0.7)> lift:(1.09) lev:(0.01) [1] conv:(1.03)

437. You\_harassed\_someone=Yes Cyberstalking\_pleasure=No Criminal\_offense=Yes 20 ==>  
Gender=Female 14 <conf:(0.7)> lift:(1.36) lev:(0.03) [3] conv:(1.39)

438. Cyberstalking\_pleasure=No Criminal\_offense=Yes 20 ==> Gender=Female  
You\_harassed\_someone=Yes 14 <conf:(0.7)> lift:(2.92) lev:(0.06) [9] conv:(2.17)

439. Victim\_of\_cyberstalking=Yes Know\_the\_stalker=No Social\_media\_communication=Yes 20 ==>  
Social\_media\_cyberstalking=Instagram 14 <conf:(0.7)> lift:(3.21) lev:(0.07) [9] conv:(2.23)

440. Know\_the\_stalker=No Social\_media\_communication=Yes 20 ==> Victim\_of\_cyberstalking=Yes  
Social\_media\_cyberstalking=Instagram 14 <conf:(0.7)> lift:(3.21) lev:(0.07) [9] conv:(2.23)

441. Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=No 20 ==>  
Social\_media\_communication=Yes 14 <conf:(0.7)> lift:(2.42) lev:(0.06) [8] conv:(2.03)

442. Get\_rid\_of\_the\_cyberstalker=Yes Criminal\_offense=No 20 ==> Victim\_of\_cyberstalking=Yes  
Social\_media\_communication=Yes 14 <conf:(0.7)> lift:(2.42) lev:(0.06) [8] conv:(2.03)

443. You\_harassed\_someone=Yes Cyberstalking\_pleasure=No Criminal\_offense=Yes 20 ==>  
Victim\_of\_cyberstalking=Yes 14 <conf:(0.7)> lift:(1.51) lev:(0.03) [4] conv:(1.53)

444. Cyberstalking\_pleasure=No Criminal\_offense=Yes 20 ==> Victim\_of\_cyberstalking=Yes  
You\_harassed\_someone=Yes 14 <conf:(0.7)> lift:(2.69) lev:(0.06) [8] conv:(2.11)

445. Gender=Female Victim\_of\_cyberstalking=Yes Get\_rid\_of\_the\_cyberstalker=Yes  
You\_harassed\_someone=Yes 20 ==> Criminal\_offense=Yes 14 <conf:(0.7)> lift:(1.09) lev:(0.01) [1]  
conv:(1.03)

446. Gender=Female Get\_rid\_of\_the\_cyberstalker=Yes You\_harassed\_someone=Yes 20 ==>  
Victim\_of\_cyberstalking=Yes Criminal\_offense=Yes 14 <conf:(0.7)> lift:(2.31) lev:(0.06) [7] conv:(1.99)

## APPENDIX C

### COJEC Algorithm's Source Code

```
1  #include <iostream>
2  #include <fstream>
3  #include <string>
4  #include <vector>
5  #include <cmath>
6  #include "strutils.h"
7
8  using namespace std;
9
10
11 bool exists (string **& FeatureValues, int i, string word)
12 {
13     for(int j = 0; j<15; j++)
14     {
15         if(FeatureValues[i][j]==word)
16             return true;
17     }
18     return false;
19 }
20
21 int addFeatureValue (string **& FeatureValues, int i, string word)
22 {
23     for(int j = 0; j<15; j++)
24     {
25         if(FeatureValues[i][j]=="CTRL")
26         {
27             FeatureValues[i][j]=word;
28             return 0;
29         }
30     }
31 }
32
33 void printFeatureValues (const vector<string> & Features, string **& FeatureValues)
34 {
35     for(int i=0; i<15; i++)
36     {
37         cout<<endl<<"Feature " <<Features[i]<<": "<<endl;
38         for(int j = 0; j<15; j++)
39         {
40             if(FeatureValues[i][j]!="CTRL")
41             {
42                 cout<<FeatureValues[i][j]<<";";
43             }
44         }
45         cout<<endl;
46     }
47 }
48
```

```

49 void printTheTable(const vector<vector<string>> & theTable)
50 {
51     for(int i=0; i<theTable.size(); i++)
52     {
53         cout<<"Row "<<i<<": "<<endl;
54         for(int j = 0; j<theTable[i].size()-1; j++)
55         {
56             cout<<theTable[i][j]<<",";
57         }
58         cout<<theTable[i][theTable[i].size()-1];
59         cout<<endl<<endl;
60     }
61 }
62
63 void processFile(ifstream &input /*,ofstream &output*/)
64 {
65     ofstream output;
66     string outFileName = "casual rules.txt";
67     if (output.fail())
68     {
69         cout << "Error: Cannot open file " << outFileName << endl;
70         system ("pause");
71     }
72
73     cout<<"processing the file"<<endl;
74     string strLine, word;
75     vector<string> Features;
76
77     getline(input,strLine);
78     StripWhite(strLine);
79     ToUpper(strLine);
80
81     unsigned int smileys = 0, index = 0, length, firstIndex = 0;
82     string searchPattern = ",";
83     if (strLine != "")
84     {
85         cout<<strLine<<endl<<endl<<endl;
86         length = strLine.length();
87
88         while (index < length)
89         {
90             index = strLine.find(",", index);
91
92             if (index != string::npos)
93             {
94                 index += searchPattern.length();
95                 word = strLine.substr(firstIndex, index-firstIndex-1);
96                 Features.push_back(word);
97                 cout<<"The found word: "<<word<<endl;
98                 firstIndex = index;
99                 smileys++;
100             }

```



```

101     }
102     word = strLine.substr(firstIndex);
103     Features.push_back(word);
104     cout<<"The found word: "<<word<<endl;
105     cout<<"Number of commas is. "<<smileys<<endl;
106 }
107
108 cout<<endl<<endl<<"The features vector is:"<<endl;
109 for(int i =0; i<Features.size(); i++)
110     cout<<Features[i]<<endl;
111 //system("pause");
112
113
114 //DEALING WITH FEATURE VALUES
115 string ** FeatureValues = new string * [15];
116 for(int i =0; i<15; i++)
117 {
118     FeatureValues[i] = new string [15];
119     for(int j = 0; j<15; j++)
120     {
121         FeatureValues[i][j]="CTRL";
122     }
123 }
124
125
126 vector<vector<string>> theTable;
127 int lineCount = 0;
128 while (getline(input,strLine))
129 {
130     StripWhite(strLine);
131     ToUpper(strLine);
132     lineCount++;
133     //cout<<endl<<endl<<"processing line "<<lineCount<<endl;
134     vector<string> row;
135     unsigned int smileys = 0, index = 0, length, firstIndex = 0;
136     string searchPattern = ",";
137     if (strLine != "")
138     {
139         //cout<<strLine<<endl;
140         length = strLine.length();
141
142         while (index < length)
143         {
144             if(strLine.at(firstIndex) == '"')
145             {
146                 index = strLine.find("\\"", index+1);
147                 if (index != string::npos)
148                 {
149                     index += searchPattern.length();
150                     word = strLine.substr(firstIndex+1, index-firstIndex-2);
151                     StripWhite(word);
152                     if(word=="")
153                         word = "CTRL";
154                     //cout<<"The found word: "<<word<<endl;
155                     row.push_back(word);
156                     //cout<<"firstIndex="<<firstIndex<<", index="<<index<<endl;
157                     index++;
158                     firstIndex = index;
159                     if(!exists(FeatureValues,smileys,word))
160                         addFeatureValue(FeatureValues,smileys,word);

```

```

161         smileys++;
162     }
163     //system("pause");
164 }
165 else
166 {
167     index = strLine.find(",", index);
168     if (index != string::npos)
169     {
170         index += searchPattern.length();
171         word = strLine.substr(firstIndex, index-firstIndex-1);
172         StripWhite(word);
173         if(word=="")
174             word = "CTRL";
175         //cout<<"The found word: "<<word<<endl;
176         row.push_back(word);
177         firstIndex = index;
178         if(!exists(FeatureValues,smileys,word))
179             addFeatureValue(FeatureValues,smileys,word);
180         smileys++;
181     }
182 }
183
184 }//while
185
186 word = strLine.substr(firstIndex);
187 StripWhite(word);
188 //cout<<"The found word: "<<word<<endl;
189 row.push_back(word);
190 if(!exists(FeatureValues,smileys,word))
191
192         addFeatureValue(FeatureValues,smileys,word);
193 if(smileys!=14)
194 {
195     cout<<"Problem Houston"<<endl;
196     system("pause");
197 }
198 //cout<<"Number of commas is "<<smileys<<endl;
199 theTable.push_back(row);
200 }//if (strLine != "")
201 }//while (getline(input,strLine))
202
203 cout<<endl<<endl<<"Now printing the feature values"<<endl;
204 printFeatureValues(Features, FeatureValues);
205
206 cout<<endl<<endl<<"Now printing the table"<<endl;
207 printTheTable(theTable);
208
209
210

```

```

211 cout<<"NOW FINDING THE CASUAL RULES"<<endl;
212 long double tau = 0.08;
213 int input1CTRL = 0, input2CTRL = 0, outputCTRL = 0;
214 //system("pause");
215 for(int input1_i = 0; input1_i<15; input1_i++)
216 {
217     for(int input1_j = 0; input1_j<15; input1_j++)
218     {
219         if(FeatureValues[input1_i][input1_j]=="CTRL")
220         {
221             input1CTRL++;
222             if(input1CTRL>1)
223                 continue;
224         }
225         for(int input2_i = input1_i+1; input2_i<15; input2_i++)
226         {
227             for(int input2_j = 0; input2_j<15; input2_j++)
228             {
229                 if(FeatureValues[input2_i][input2_j]=="CTRL")
230                 {
231                     input2CTRL++;
232                     if(input2CTRL>1)
233                         continue;
234                 }
235                 for(int output_i = 0; output_i<15; output_i++)
236                 {
237                     if((output_i == input1_i)|| (output_i == input2_i))
238                         continue;
239                     for(int output_j = 0; output_j<15; output_j++)
240                     {
241                         if(FeatureValues[output_i][output_j]=="CTRL")
242                         {
243                             outputCTRL++;
244                             if(outputCTRL>1)
245                                 continue;
246                         }
247                     }
248                     int inputsMatchOnly=0, outputsAndInputsMatch = 0;
249                     for(int table_i = 0; table_i<theTable.size(); table_i++)
250                     {
251                         if((theTable[table_i][input1_i]==FeatureValues[input1_i][input1_j])&&
252                             (theTable[table_i][input2_i]==FeatureValues[input2_i][input2_j]))
253                         {
254                             inputsMatchOnly++;
255                             /*cout<<"-----"<<endl;
256                             cout<<"theTable["<<table_i<<"]["<<input1_i<<"]="
257                             "<<theTable[table_i][input1_i]<<endl;
258                             cout<<"FeatureValues["<<input1_i<<"]
259                             ["<<input1_j<<"]="<<FeatureValues[input1_i][input1_j]<<endl<<endl;

```

```

260         cout<<"theTable["<<table_i<<"]["<<input2_i<<"]="
261         "<<theTable[table_i][input2_i]<<endl;
262         cout<<"FeatureValues["<<input2_i<<"]["<<input2_j<<"]="
263         "<<FeatureValues[input2_i][input2_j]<<endl<<endl;
264         cout<<"-----"<<endl;*/
265         if(theTable[table_i][output_i]==FeatureValues[output_i][output_j])
266         {
267             outputsAndInputsMatch++;
268         }
269     }
270 }//for(int table_i = 0; table_i<theTable.size(); table_i++)
271 //cout<<"-----"<<endl;
272 //cout<<"inputsMatchOnly="<<inputsMatchOnly<<endl;
273 //cout<<"outputsAndInputsMatchy="<<outputsAndInputsMatch<<endl<<endl;
274 long double PX = (long double) inputsMatchOnly/theTable.size();
275 long double PXY = (long double) outputsAndInputsMatch/theTable.size();
276 long double HYX = PXY*(log(PXY)/log(PX));
277 long double tau = 0.08;

278
279
280
281 if((HYX<tau) && (outputsAndInputsMatch>9))
282 {
283     cout<<"-----"<<endl;
284     cout<<"HYX="<<HYX<<endl;
285     cout<<"inputsMatchOnly="<<inputsMatchOnly<<endl;
286     cout<<"outputsAndInputsMatchy="<<outputsAndInputsMatch<<endl<<endl;
287     cout<<"Feature 1="<<Features[input1_i]
288     <<", value="<<FeatureValues[input1_i][input1_j]<<endl;
289
290     cout<<"Feature 2="<<Features[input2_i]
291
292     <<", value="<<FeatureValues[input2_i][input2_j]<<endl;
293
294     cout<<"> "<<"Feature="<<Features[output_i]
295     <<", value="<<FeatureValues[output_i][output_j]<<endl<<endl;
296
297     output<<"-----"<<endl;
298     output<<"HYX="<<HYX<<endl;
299     output<<"inputsMatchOnly="<<inputsMatchOnly<<endl;
300     output<<"outputsAndInputsMatchy="<<outputsAndInputsMatch<<endl<<endl;
301     output<<"Feature 1="<<Features[input1_i]
302     <<", value="<<FeatureValues[input1_i][input1_j]<<endl;
303
304     output<<"Feature 2="<<Features[input2_i]
305     <<", value="<<FeatureValues[input2_i][input2_j]<<endl;
306
307     output<<">"<<endl;
308     output<<"Feature="<<Features[output_i]
309     <<", value="<<FeatureValues[output_i][output_j]<<endl<<endl;
310 }

```

```

311     }
312     }
313     }
314     }
315     }
316     }
317     }
318     output.close();
319 }
320
321 int main()
322 {
323     ifstream input;
324     //ofstream output;
325     string strLine;
326
327     string inFileName = "shkurte.csv";
328     //string outFileName = "casual rules.txt";
329
330     input.open(inFileName.c_str());
331     if (input.fail())
332     {
333         cout << "Error: Cannot open file " << inFileName << endl ;
334     }
335
336     processFile(input/*,output*/);
337     //else
338     //{
339     //    output.open(outFileName.c_str());
340     //    if (output.fail())
341     //    {
342     //        cout << "Error: Cannot open file " << outFileName << endl ;
343     //    }
344     //    else
345     //    {
346     //        processFile(input/*,output*/);
347     //    }
348     //    output.close();
349     //}
350     system("pause");
351     return 0;
352 }
353
354

```

## REFERENCES

- Abdullah, A. T., & Jahan, I. (2020). Causes of Cybercrime Victimization: A Systematic Literature Review. *International Journal of Research and Review*, 7(5), 89- 98.
- Abela, S., Tang, Y., Singh, J., & Paek, E. (2020). Applications of Causal Modeling in Cybersecurity: An Exploratory Approach. *Advances in Science, Technology and Engineering Systems Journal*, 380-387.
- Abin, D., Mahajan, T. C., Bhoj, M. S., Bagde, S., & Rajeswari, K. (2015). Causal Association Mining for Detection of Adverse Drug. *International Conference on Computing Communication Control and Automation* (pp. 382-385). India: IEEE, ACM. doi:10.1109/ICCUBEA.2015.80
- Abualigah, L., Diabat, A., Mirjalili, S., Abd Elaziz, M., & Gandomih, A. H. (2021). The Arithmetic Optimization Algorithm. *Computer Methods in Applied Mechanics and Engineering*, 376. doi:doi.org/10.1016/j.cma.2020.113609
- Abu-Ulbeh, W., Altalhi, M., Abualigah, L., Ali Almazroi, A., Sumari, P., & Gandomi, A. H. (2021). Cyberstalking Victimization Model Using Criminological Theory: A Systematic Literature Review, Taxonomies Applications, Tools, and Validations. *Electronics*. doi:10.3390/electronics10141670
- ACM. (2018). ACM Code of Ethics and Professional Conduct. *Communications of the ACM*. doi:10.1145/3274591
- Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *20th International Conference on Very Large Data Bases VLDB '94* (pp. 478-499). California, USA: ACM.
- Agrawal, R., & Srikant, R. (2000). Privacy-Preserving Data Mining. *ACM SIGMOD international conference on Management of data* (pp. 439–450). ACM.
- Agrawal, S., & Agrawal, J. (2015). Survey on Anomaly Detection using Data Mining Techniques. *19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems*. 60, pp. 708-713. Singapore: Elsevier. doi:10.1016/j.procs.2015.08.220
- Aha, D., Murphy, P., Merz, C., Keogh, E., Blake, C., Hettich, S., & Newman, D. (Eds.). (1987). *UCI Machine Learning Repository*. Retrieved November 24, 2019, from <https://archive.ics.uci.edu/ml/index.php>
- Ait-Mlouk, A., Gharnati, F., & Agouti, T. (2017). An improved approach for association rule mining using a multi-criteria decision support system: a case study in road safety. *European Transport Research Review*, 9, 40. doi:10.1007/s12544-017-0257-5
- Al Mutawa, N., Bryce, J., Franqueira, V. N., Marrington, A., & Read, J. C. (2019). Behavioural Digital Forensics Model: Embedding Behavioural Evidence Analysis into the Investigation of Digital Crimes. 28, 70-82. doi:10.1016/j.diin.2018.12.003
- Alharbi, M., & Rajasekaran, S. (2015). Conjunctive Combined Causal Rules Mining. *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)* (pp. 28-33). Abu Dhabi: IEEE, ACM. doi:10.1109/ISSPIT.2015.7394344

- Alharbi, M., & Rajasekaran, S. (2015). Disjunctive Combined Causal Rules Mining. *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)* (pp. 40-45). Abu Dhabi: IEEE, ACM. doi:10.1109/ISSPIT.2015.7394368
- Ali, I., Cawkwell, F., Dwyer, E., & Green, S. (2016). Modeling Managed Grassland Biomass Estimation by Using Multitemporal Remote Sensing Data—A Machine Learning Approach. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(7), 3254-3264. doi:10.1109/JSTARS.2016.2561618
- Aliakbari, F., Hammad, K., Bahrami, M., & Aein, F. (2015). Ethical and legal challenges associated with disaster nursing. *Nursing Ethics*, 22(4), 493-503. doi:10.1177/0969733014534877
- Alonso, J., Castañón, Á. R., & Bahamonde, A. (2013). Support Vector Regression to predict carcass weight in beef cattle in advance of the slaughter. *Computers and Electronics in Agriculture*, 91, 116-120. doi:10.1016/j.compag.2012.08.009
- Alonso, J., Villa, A., & Bahamonde, A. (2015). Improved estimation of bovine weight trajectories using Support Vector Machine Classification. *Computers and Electronics in Agriculture*, 110, 36-41. doi:10.1016/j.compag.2014.10.001
- Amatya, S., Karkee, M., Gongal, A., Zhang, Q., & Whiting, M. D. (2016). Detection of cherry tree branches with full foliage in planar architecture for automated sweet-cherry harvesting. *Biosystems Engineering*, 146, 3–15. doi:10.1016/j.biosystemseng.2015.10.003
- Amazon AWS. (2020, January 29). Retrieved from <https://aws.amazon.com/opendata/?wwps-cards.sort-by=item.additionalFields.sortDate&wwps-cards.sort-order=desc>
- Anaconda Navigator. (2021). Retrieved March 9, 2021, from <https://docs.anaconda.com/anaconda/navigator/#:~:text=Anaconda%20Navigator%20is%20a%20desktop,in%20a%20local%20Anaconda%20Repository.>
- Association, W. (2013). Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. *J. Am. Med. Assoc.*, 310, E1-E4.
- Barth-Jones, D. (2012, August 10). *The debate over 'Re-Identification' of Health Information: What do we risk?* Retrieved August 21, 2020, from Health Affairs.
- Bello, N. M., Ferreira, V. C., Gianola, D., & Rosa, G. J. (2018). Conceptual framework for investigating causal effects from observational data in livestock. *Journal of Animal Science*, 96(10), 4045–4062. doi:10.1093/jas/sky277
- Bhoopathi, H., & Rama, B. (2016). Study of Causal Data Mining with Comparison of its Algorithms. *International Journal of Scientific Research And Education*, 4(3), 5118-5122. doi:10.18535/ijrsre/v4i03.18
- Bhoopathi, H., & Rama, B. (2017). Causal Rule Mining for Knowledge Discovery from Databases. *International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 978-984). India: IEEE. doi:10.1109/ICCONS.2017.8250611

- Bowes, J., Neufeld, E., Greer, J. E., & Cooke, J. (2000). A Comparison of Association Rule Discovery and Bayesian Network Causal Inference Algorithms to Discover Relationships in Discrete Data. *13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence* (pp. 326-336). Canada: Springer. doi:10.1007/3-540-45486-1\_27
- Boyd, D., Keller, E. F., & Tijerina, B. (2016). *Supporting Ethical Data Research: An Exploratory Study of Emerging Issues in Big Data and Technical Research*. New York: Data & Society Research Institute.
- Budhathoki, K., Boley, M., & Vreeken, J. (2018). Rule Discovery for Exploratory Causal Reasoning. *32nd Conference on Neural Information Processing Systems (NIPS2018)*. Canada.
- Busse, M., Kernecker, M. L., & Siebert, R. (2020). Ethical Issues in Poultry Production – Datasets from a German Consumer Survey. *Data in Brief*, 31. doi:10.1016/j.dib.2020.105748
- Cao, L., & Kevin Wang, S.-Y. (2020). Correlates of stalking victimization in Canada: A model of social support and comorbidity. *International Journal of Law, Crime and Justice*. doi:10.1016/j.ijlcrj.2020.100437
- Carbonnelle, P. (2021). *PYPL Index*. Retrieved 03 09, 2021, from <https://pypl.github.io/PYPL.html>
- Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, 151, 61-69. doi:10.1016/j.compag.2018.05.012
- Chung, C.-L., Huang, K.-J., Chen, S.-Y., Lai, M.-H., Chen, Y.-C., & Kuo, Y.-F. (2016). Detecting Bakanae disease in rice seedlings by machine vision. *Computers and Electronics in Agriculture*, 121(C), 404-411. doi:10.1016/j.compag.2016.01.008
- Clifton, C., Kantarcioglu, M., & Vaidya, J. (2002). Defining Privacy for Data Mining. *National Science Foundation Workshop on Next Generation Data Mining*, (pp. 126-133). Baltimore, Maryland, USA.
- Collecting Data: Surveys, Experiments, & Observational Studies*. (n.d.). (D. Roberts, Producer) Retrieved March 10, 2021, from <https://mathbitsnotebook.com/Algebra2/Statistics/STSsurveys.html>
- Cook, J. (2009). Ethics of Data Mining. In J. Wang, *Encyclopedia of Data Warehousing and Mining*, 2nd Edition (pp. 783-788). Hershey, Pennsylvania, USA. doi:10.4018/978-1-60566-010-3.ch121
- Cooper, A. K., & Coetzee, S. (2020). *On the Ethics of Using Publicly-Available Data*. Pretoria, South Africa.
- Cooper, G. F. (1997). A Simple Constraint-Based Algorithm for Efficiently Mining Observational Databases for Causal Relationship. *Data Mining and Knowledge Discovery*, 1(2), 203– 224. doi:10.1023/A:1009787925236
- Coopersmith, E. J., Minsker, B. S., & Wenzel, C. E. (2014). Machine learning assessments of soil drying for agricultural planning. *Computers and Electronics in Agriculture*, 104, 93-104. doi:10.1016/j.compag.2014.04.004
- Craninx, M., Fievez, V., Vlaeminck, B., & De Baets, B. (2008). Artificial neural network models of the rumen fermentation pattern in dairy cattle. *Computers and Electronics in Agriculture*, 60(2), 226-238. doi:10.1016/j.compag.2007.08.005



- Creswell, J. W. (2012). *Educational Research: Planning, Conducting and Evaluation Quantitative and Qualitative Research, 4th Ed.* Boston, Massachusetts, USA: Pearson.
- DataFlair, T. (2019, 02 15). *Data Flair*. Retrieved 12 5, 2019, from <https://data-flair.training/blogs/data-mining-and-knowledge-discovery/>
- Dehkharghani, R., Mercan, H., Javeed, A., & Saygn, Y. (2014). Sentimental Causal Rule Discovery from Twitter. *Expert Systems with Applications*, 49(10), 4950-4958. doi:10.1016/j.eswa.2014.02.024
- Digital 2021: Global Overview Report*. (2021, January 27). (S. Kemp, Producer) Retrieved March 10, 2021, from <https://datareportal.com/reports/digital-2021-global-overview-report>
- Doolan, D. M., Winters, J., & Nouredini, S. (2017). Answering Research Questions Using an Existing Data Set. *Medical Research Archives*, 5(9).
- Dorey, C. M., Baumann, H., & Andorno, N. B. (2018). Patient data and patient rights: Swiss healthcare stakeholders' ethical awareness regarding large patient data sets - a qualitative study. *BMC medical ethics*, 19(1). doi:10.1186/s12910-018-0261-x
- Dutta, R., Smith, D., Rawnsley, R., Bishop-Hurley, G., Hills, J., Timms, G., & Henry, D. (2015). Dynamic cattle behavioural classification using supervised ensemble classifiers. *Computers and Electronics in Agriculture*, 111, 18-28. doi:10.1016/j.compag.2014.12.002
- Ebrahimi, M. A., Khoshtaghaza, M.-H., Minaee, S., & Jamshidi, B. (2017). Vision-based pest detection based on SVM classification method. *Computers and Electronics in Agriculture*, 137, 52-58. doi:10.1016/j.compag.2017.03.016
- Eck, N. J., & Waltman, L. (2020). *Manual for VOSviewer version 1.6.15*. University of Leiden and CWTS Meaningfull Metrics.
- ElegantJ BI*. (2018, October 17). Retrieved March 30, 2021, from <https://www.elegantjbi.com/blog/what-is-karl-pearson-correlation-analysis-and-how-can-it-be-used-for-enterprise-analysis-needs.htm>
- EU. (2018). *Ethics and data protection*.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3), 37-54. doi:10.1609/aimag.v17i3.1230
- Feng, Y., Peng, Y., Cui, N., Gong, D., & Zhang, K. (2017). Modeling reference evapotranspiration using extreme learning machine and generalized regression neural network only with temperature data. *Computers and Electronics in Agriculture*, 136, 71–78. doi:10.1016/j.compag.2017.01.027
- Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, 145, 311–318. doi:10.1016/j.compag.2018.01.009
- Fiesler, C. (2019). Ethical Considerations for Research Involving (Speculative) Public Data. *Proceedings of the ACM on Human-Computer Interaction*, 3, 1–13. doi:10.1145/3370271

- Floridi, L., & Taddeo, M. (2016). What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374((2083):20160360). doi:10.1098/rsta.2016.0360
- Francis, R. A., Guikema, S. D., & Henneman, L. (2014). Bayesian belief networks for predicting drinking water distribution system pipe breaks. *Reliability Engineering & System Safety*, 130, 1-11. doi:10.1016/j.ress.2014.04.024
- Fule, P., & Roddick, J. (2004). Detecting Privacy and Ethical Sensitivity in Data Mining Results. *27th Australasian Computer Science Conference (ACSC2004)* (pp. 159-166). Dunedin, New Zealand: ACM, dbpl.
- Girju, R., & Moldovan, D. (2002). Mining Answers for Causation Questions. *American Association for Artificial Intelligence (AAAI) Spring Symposium on Mining Answers from Texts and Knowledge Bases*, (pp. 15-25). California, USA.
- Girotra, M., Nagpal, K., Minocha, S., & Sharma, N. (2013). Comparative Survey on Association Rule Mining Algorithms. *International Journal of Computer Applications*, 84(10), 18-22. doi:10.5120/14612-2862
- Glymour, I., Scheines, R., Spirtes, P., & Ramsey, J. (2017). *The Tetrad Project*. Retrieved December 24, 2019, from <http://www.phil.cmu.edu/tetrad/>
- Gray, D., Bowes, D., Davey, N., Sun, Y., & Christianson, B. (2011). The Misuse of the NASA Metrics Data Program Data Sets. *15th Annual Conference on Evaluation & Assessment in Software Engineering* (pp. 96 – 103). Durham, UK: IEEE. doi:10.1049/ic.2011.0012
- Great Learning. (2019, October 23). Retrieved January 28, 2020, from <https://www.mygreatlearning.com/blog/top-5-sources-for-analytics-and-machine-learning-datasets/>
- Gujarati, D. N., & Porter, D. C. (2009). *Basic Econometric, 5th Edition*. The McGraw-Hill Companies.
- Guo, R., Cheng, L., Li, J., Hahn, P. R., & Liu, H. (2010). A Survey of Learning Causality with Data: Problems and Methods. *ACM Transactions on the Web, The SCImago Journal & Country Rank*, 9(4). doi:arXiv:1809.09337v3
- Guo, Y., Xing, W., & Lee, H.-S. (2015). Identifying Students' Mechanistic Explanations in Textual Responses to Science Questions with Association Rule Mining. *IEEE 15th International Conference on Data Mining Workshops* (pp. 264-268). New Jersey, USA: IEEE. doi:10.1109/ICDMW.2015.225
- Guo, Y.-M., Huang, Z.-L., Guo, J., Li, H., & Guo, X.-R. (2019). Bibliometric Analysis on Smart Cities Research. *Sustainability*, 11(3606). doi:doi:10.3390/su11133606
- Gupta, A. (2018, February 11). *Legal and ethical implications of data accessibility for public welfare and AI research advancement*. Retrieved August 21, 2020
- Hand, D. J. (2018). Aspects of Data Ethics in a Changing World: Where Are We Now? *Big Data*, 6(3), 176–190. doi:10.1089/big.2018.0083

- Haron, H., & Bt. Mohd Yusof, F. (2010). Cyber stalking: The social impact of social networking technology. *International Conference on Education and Management Technology*, (pp. 237-241). Cairo, Egypt. doi:10.1109/ICEMT.2010.5657665
- Harris, R. B., Samberg, L. H., Yeh, E. T., Smith, A. T., Wenying, W., Junbang, W., . . . Bedunah, D. J. (2016). Rangeland responses to pastoralists' grazing management on a Tibetan steppe grassland, Qinghai Province, China. *The Rangeland Journal*, 38(1 ), 1-15. doi:10.1071/RJ15040
- Hassani, H., Huang, X., & Ghods, M. (2017). Big data and Causality. *Annals of Data Science*, 5(2), 133-156. doi:10.1007/s40745-017-0122-3
- Hastings, J., Branting, K., & Lockwood, J. (2002). CARMA: A Case-Based Rangeland Management Adviser. *AI Magazine*, 23(2). doi:10.1609/aimag.v23i2.1640
- Hira, S., & Deshpande, P. S. (2016). Mining precise cause and effect rules in large time series data of socio-economic indicators. *Springer Plus*, 5(1). doi:10.1186/s40064-016-3292-0
- Horne, A. C., Szemis, J. M., Webb, J. A., Kaur, S., Stewardson, M. J., Bond, N., & Nathan, R. (2018). Informing environmental water management decisions: using conditional probability networks to address the information needs of planning and implementation cycles. *Environmental Management*, 61(3), 347-357. doi:10.1007/s00267-017-0874-8
- Humphreys, S. (2013). Healthcare datasets: ethical concerns. 310–311. doi:10.3399/bjgp13X668230
- IEEE Xplore. (n.d.). Retrieved June 15, 2020, from <https://ieeexplore.ieee.org/Xplorehelp/overview-of-ieee-xplore/about-ieee-xplore>
- Institute for Work&Health. (2016, February). *Observational vs. Experimental studies*. (Toronto) Retrieved December 18, 2019, from <https://www.iwh.on.ca/what-researchers-mean-by/observational-vs-experimental-studies>
- Jin, Z., Li, J., Liu, L., Le, T. D., Sun, B., & Wang, R. (2012). Discovery of Causal Rules Using Partial Association. *IEEE 12th International Conference on Data Mining* (pp. 309-318). Belgium: IEEE. doi:10.1109/ICDM.2012.36
- Johann, A. L., Araújo, A. G., Delalibera, H. C., & Hirakawa, A. R. (2016). Soil moisture modeling based on stochastic behavior of forces on a no-till chisel opener. *Computers and Electronics in Agriculture*, 121, 420-428. doi:10.1016/j.compag.2015.12.020
- Jupyter Notebook. (2021, February 8). Retrieved March 9, 2021, from <https://jupyter.org/about>
- Karimi, K. (2010). *A Brief Introduction to Temporality and Causality*. Switzerland. doi:arXiv:1007.2449
- Kaur, C. (2013). Association Rule Mining using Apriori Algorithm: A Survey. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 2(6), 2081-2084.
- Khan, S., & Parkinson, S. (2017). Causal Connections Mining Within Security Event Logs. *9th International Conference on Knowledge Capture*. Texas, USA: ACM. doi:10.1145/3148011.3154476

- Kjamilji, A., Idrizi, A., Luma-Osmani, S., & Zenuni-Kjamilji, F. (2020). Secure Naïve Bayes classification without loss of accuracy. *International Conference on Science and Engineering*, 3, pp. 397-403. Yogyakarta, Indonesia. Retrieved from <http://sunankalijaga.org/prosiding/index.php/icse/article/view/536>
- Kung, H.-Y., Kuo, T.-H., Chen, C.-H., & Tsai, P.-Y. (2016). Accuracy Analysis Mechanism for Agriculture Data Using the Ensemble Neural Network Method. *Sustainability*, 8, 735. doi:10.3390/su8080735
- Lagos-Ortiz, K., Salas-Zárate, M. d., Paredes-Valverde, M. A., García-Díaz, J. A., & Valencia-García, R. (2020). AgriEnt: A Knowledge-Based Web Platform for Managing Insect Pests of Field Crops. *Applied Sciences*, 10(3), 1040. doi:10.3390/app10031040
- Lee, S., Cha, Y., Han, S., & Hyun, C. (2019). Application of Association Rule Mining and Social Network Analysis for Understanding Causality of Construction Defects. *Sustainability - Open Access Journal*, 11(3), 618-632. doi:10.3390/su11030618
- Leetaru, K. (2017, July 20). *Should Open Access And Open Data Come With Open Ethics*. Retrieved August 20, 2020, from Forbes: AI & Big Data: <https://www.forbes.com/sites/kalevleetaru/2017/07/20/should-open-access-and-open-data-come-with-open-ethics/>
- Li, J., Le, T. D., Liu, L., Liu, J., Jin, Z., Sun, B., & Ma, S. (2015). From Observational Studies to Causal Rule Mining. *ACM Transactions on Intelligent Systems and Technology (TIST) - Special Issue on Causal Discovery and Inference*, 7(2), 14:1–14:27. doi:10.1145/2746410
- Li, J., Le, T. D., Liu, L., Liu, J., Jinyz, Z., & Sun, B. (2013). Mining Causal Association Rules. *IEEE 13th International Conference on Data Mining Workshops* (pp. 114–123). Texas, USA: IEEE. doi:10.1109/ICDMW.2013.88
- Li, J., Ma, S., Le, T., Liu, L., & Liu, J. (2016). Causal Decision Trees. *IEEE Transactions on Knowledge and Data Engineering*, 29(2), 257–271. doi:10.1109/TKDE.2016.2619350
- Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine Learning in Agriculture: A Review. *Sensors*, 18(8). doi:10.3390/s18082674
- Liao, H., Tang, M., Luo, L., Li, C., Chiclana, F., & Zeng, X.-J. (2018). A Bibliometric Analysis and Visualization of Medical Big Data Research. *Sustainability*, 10(166). doi:10.3390/su10010166
- Liu, D. Z., Gao, Y., & Zhao, J. P. (2008). An association based approach to discovering ordering rules. *7th International Conference on Machine Learning and Cybernetics* (pp. 202-205). Kunming, China: IEEE. doi:10.1109/ICMLC.2008.4620404
- Luma-Osmani, S., Ismaili, F., & Ram Pal, P. (2021). Building a Model in Discovering Multivariate Causal Rules for Exploratory Analyses. *International Conference on Data Analytics for Business and Industry* (pp. 272-276). Sakheer, Bahrain: IEEE. doi:10.1109/ICDABI53623.2021.9655981
- Luma-Osmani, S., Ismaili, F., & Raufi, B. (2020). Bibliometric Analysis and Visualization of Ethical concerns on Publicly accessible data sets. *4th International Scientific Conference on Business and*

- Information Technologies* (pp. 168-179). Tetovo, Republic of North Macedonia: South East European University.
- Luma-Osmani, S., Ismaili, F., Pathak, P., & Zenuni, X. (2022, January). Identifying Causal Structures from Cyberstalking: Behaviors Severity and Association. *Journal of Communications Software and Systems*, 8(1), 1-8.
- Luma-Osmani, S., Ismaili, F., Raufi, B., & Zenuni, X. (2020). Causal Reasoning Application in Smart Farming and Ethics: A Systematic Review. *Annals of Emerging Technologies in Computing (AETiC)*, 4(4), 10-18. doi:10.33166/AETiC.2020.04.002
- Luma-Osmani, S., Ismaili, F., Zenuni, X., & Raufi, B. (2020). A Systematic Literature Review in Causal Association Rules Mining. In P. R. Paul (Ed.), *11th Annual IEEE Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (pp. 20-26). Vancouver, BC, Canada: IEEE. doi:10.1109/IEMCON51383.2020.9284908
- Maddock, J., Mason, R., & Starbird, K. (n.d.). *Using Historical Twitter Data for Research: Ethical Challenges of Tweet Deletions*. University of Washington.
- Mamet, S. D., Redlick, E., Brabant, M., Lamb, E. G., Helgason, B. L., Stanley, K., & Siciliano, S. D. (2019). Structural equation modeling of a winnowed soil microbiome identifies how invasive plants re-structure microbial networks. *The ISME Journal*, 13(8), 1988-1996. doi:10.1038/s41396-019-0407-y
- Mani, S., & Cooper, G. F. (2001). A Simulation Study of Three Related Causal Data Mining Algorithms. *8th International Workshop on Artificial Intelligence and Statistics* (pp. 73-80). California, USA: dblp.
- Matthews, S. G., Miller, A. L., Plötz, T., & Kyriazakis, I. (2017). Automated tracking to measure behavioural changes in pigs for health and welfare monitoring. *Scientific Reports*, 7, 17582. doi:10.1038/s41598-017-17451-6
- Mazlack, L. J. (2001). Considering Causality in Data Mining. *5th WSES/IEEE World Multiconference*, (pp. 493-498). Cincinnati, USA.
- Mehdizadeh, S., Behmanesh, J., & Khalili, K. (2017). Using MARS, SVM, GEP and empirical equations for estimation of monthly mean reference evapotranspiration. *Computers and Electronics in Agriculture*, 139, 103-114. doi:10.1016/j.compag.2017.05.002
- Mohammadi, K., Shamshirband, S., Motamedi, S., Petković, D. H., & Gocic, M. (2015). Extreme learning machine-based prediction of daily dew point temperature. *Computers and Electronics in Agriculture*, 117, 214-225. doi:10.1016/j.compag.2015.08.008
- Molina, J.-L., & Zazo, S. (2017). Causal Reasoning for the Analysis of Rivers Runoff Temporal Behavior. *Water Resources Management*, 31, 4669-4681. doi:10.1007/s11269-017-1772-9
- Morales, I. R., Cebrián, D. R., Blanco, E. F., & Sierra, A. P. (2016). Early warning in egg production curves from commercial hens: A SVM approach. *Computers and Electronics in Agriculture*, 121, 169-179. doi:10.1016/j.compag.2015.12.009

- Morellos, A., Pantazi, X.-E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., . . . Mouazen, A. M. (2016). Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. *Biosystems Engineering*, 152, 104-116. doi:10.1016/j.biosystemseng.2016.04.018
- Mori, I. (2016). *Public dialogue on the ethics of data science in government*. London: Ipsos Mori.
- Moshou, D., Bravo, C., Oberti, R., West, J., Bodria, L., McCartney, A., & Ramon, H. (2005). Plant disease detection based on data fusion of hyper-spectral and multi-spectral fluorescence imaging using Kohonen maps. *Real-Time Imaging*, 11(2), 75-83. doi:10.1016/j.rti.2005.03.003
- Moshou, D., Bravo, C., West, J., Wahlen, S., McCartney, A., & Ramon, H. (2004). Automatic detection of 'yellow rust' in wheat using reflectance measurements and neural networks. *Computers and Electronics in Agriculture*, 44(3), 173-188. doi:10.1016/j.compag.2004.04.003
- Moshou, D., Pantazi, X.-E., Kateris, D., & Gravalos, I. (2014). Water stress detection based on optical multisensor fusion with a least squares support vector machine classifier. *Biosystems Engineering*, 117, 15-22. doi:10.1016/j.biosystemseng.2013.07.008
- Mucherino, A., Papajorgji, P. J., & Pardalos, P. (2009). *Data Mining in Agriculture*. New York, USA: Springer-Verlag. doi:10.1007/978-0-387-88615-2
- Mugan, J. (2013). A Developmental Approach to Learning Causal Models for Cyber Security. *Machine Intelligence and Bio-inspired Computation: Theory and Applications VII*, 8751. Baltimore, Maryland, United States. doi:10.1117/12.2014418
- Nahvi, B., Habibi, J., Mohammadi, K., Shamsirband, S., & Razgan, O. S. (2016). Using self-adaptive evolutionary algorithm to improve the performance of an extreme learning machine for estimating soil temperature. *Computers and Electronics in Agriculture*, 124, 150-160. doi:10.1016/j.compag.2016.03.025
- Narayanan, A., & Shmatikov, V. (2008). *Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)*. Austin.
- Neuberg, L. G. (2003). Causality: Models, Reasoning, and Inference by Judea Pearl. *Econometric Theory*, 19(4), 675–685. doi:10.1017/S0266466603004109
- Newton, E., Sweeney, L., & Malin, B. (2003). *Preserving Privacy by De-identifying Facial Images*. Pittsburgh.
- Noh, B., Son, J., Park, H., & Chang, S. (2017). In-Depth Analysis of Energy Efficiency Related Factors in Commercial Buildings Using Data Cube and Association Rule Mining. *Sustainability*, 9, 2119. doi:10.3390/su9112119
- Ochoa, S., Rasmussen, J., Robson, C., & Salib, M. (2001). Reidentification of Individuals in Chicago's Homicide Database: A Technical and Legal Study. *Massachusetts Institute of Technology*.
- OECD. (2016). *Research Ethics and New Forms of Data for Social and Economic Research*. Paris: OECD.

- Pantazi, X. E., Moshou, D., Oberti, R., West, J., Mouazen, A., & Bochtis, D. (2017). Detection of biotic and abiotic stresses in crops by using hierarchical self-organizing classifiers. *Precision Agriculture*, 18(3), 383–393. doi:10.1007/s11119-017-9507-8
- Pantazi, X. E., Tamouridou, A., Alexandridis, T. K., Lagopodi, A., Kontouris, G., & Moshou, D. (2017). Detection of *Silybum marianum* infection with *Microbotryum silybum* using VNIR field spectroscopy. *Computers and Electronics in Agriculture*, 137, 130-137. doi:10.1016/j.compag.2017.03.017
- Pantazi, X., Moshou, D., Alexandridis, T., Whetton, R., & Mouazen, A. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture*, 121, 57-65. doi:10.1016/j.compag.2015.11.018
- Parker, M. (2015). *Exploring the Ethical Imperative for Data Sharing*. Washington : National Academy of Science.
- Patil, A. P., & Deka, P. C. (2016). An extreme learning machine approach for modeling evapotranspiration using extrinsic inputs. *Computers and Electronics in Agriculture*, 121, 385–392. doi:10.1016/j.compag.2016.01.016
- Pearl, J. (2010). An introduction to causal inference. *The International Journal of Biostatistics*, 6(2). doi:10.2202/1557-4679.1203
- Pearl, J. (2012). The causal foundations of structural equation modeling. In R. H. Hoyle, *Handbook of Structural Equation Modeling* (pp. 68-91). Los Angeles, USA: Dept of Computer Science, California University.
- Pedreschi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware Data Mining. *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Las Vegas, Nevada, USA: ACM. doi:"10.1145/1401890.1401959
- Pegorini, V., Karam, L. Z., Pitta, C. S., Cardoso, R., da Silva, J. C., Kalinowski, H. J., . . . Assmann, T. S. (2015). In Vivo Pattern Classification of Ingestive Behavior in Ruminants Using FBG Sensors and Machine Learning. *Sensors*, 15(11), 28456–28471. doi:10.3390/s151128456
- Pellet, J. P., & Elisseeff, A. (2008). Using Markov Blankets for Causal Structure Learning. *Journal of Machine Learning Research*, 9(7), 1295-1342.
- Petersen, M. L. (2011). Compound treatments, transportability, and the structural causal model: the power and simplicity of causal graphs. *Epidemiology*, 22(3), 378-381. doi:10.1097/EDE.0b013e3182126127
- Privacy Lives*. (2013, May 6). Retrieved August 10, 2020, from Forbes: Harvard Professor Re-Identifies Anonymous Volunteers In DNA Study: <https://www.privacylives.com/forbes-harvard-professor-re-identifies-anonymous-volunteers-in-dna-study/2013/05/06/>
- PyCharm*. (n.d.). Retrieved August 30, 2020, from <https://www.jetbrains.com/pycharm/>
- Python*. (2021, March 9). Retrieved from <https://docs.python.org/3/tutorial/index.html>

- Ram Pal, P., Pathak, P., & Luma-Osmani, S. (2021, March). IHAC: Incorporating Heuristics for Efficient Rule Generation & Rule Selection in Associative Classification. *Journal of Information & Knowledge Management*, 20(01), 2150010 - 1-13. doi:10.1142/S0219649221500106
- Ram Pal, P., Pathak, P., Yadav, V., & Ora, P. (2019). Classification of Pruning Methodologies for Model Development using Data Mining Techniques. *International Journal of Engineering and Advanced Technology (IJEAT)*, 9(2), 2043-2047.
- Rameshkumar, K., Sambath, M., & Ravi, S. (2013). Relevant Association Rule Mining from Medical Dataset Using New Irrelevant Rule Elimination Technique. *International Conference on Information Communication and Embedded Systems (ICICES)* (pp. 300-304). India: IEEE. doi:10.1109/ICICES.2013.6508351
- Rammohan, R. R. (2010). Three algorithms for causal learning. Retrieved from [https://digitalrepository.unm.edu/cs\\_etds/14](https://digitalrepository.unm.edu/cs_etds/14)
- Ramos, P. J., Prieto, F. A., Montoya, E., & Oliveros, C. E. (2017). Automatic fruit count on coffee branches using computer vision. *Computers and Electronics in Agriculture*, 137(C), 9-22. doi:10.1016/j.compag.2017.03.010
- Ramsey, J. D., Zhang, K., Glymour, M., Romero, R. S., Huang, B., Ebert-Uphoff, I., . . . Glymour, C. (2018). Tetrad - A toolbox for Causal Discovery. *8th international workshop on Climate Informatics (CI 2018)*, (pp. 89-92). Coldorado, USA. doi:10.5065/D6BZ64XQ
- Reynolds, G. W. (2015). *Ethics in Information Technology*. Massachusetts, USA: Cengage Learning.
- Reynolds, J. (n.d.). *Real Python*. Retrieved March 9, 2021, from 8 World-Class Software Companies That Use Python: <https://realpython.com/world-class-companies-using-python/>
- Rink, B., Bejan, C. A., & Harabagiu, S. (2010). Learning Textual Graph Patterns to Detect Causal Event Relations. *23rd International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010)*, (pp. 265-270). Florida, USA.
- Robinson, R. (1977). Counting unlabeled acyclic digraphs. *Little C.H.C. (eds) Combinatorial Mathematics V. Lecture Notes in Mathematics*, 622. doi:10.1007/BFb0069178
- Rosa, G. J., & Valente, B. D. (2013). Breeding and Genetics Symposium: Inferring causal effects from observational data in livestock. *Journal of Animal Science*, 91(2). doi:10.2527/jas.2012-5840
- Rothman, D. (1982). Were Tuskegee & Willowbrook 'studies in nature'? *American Journal of Medicine*, 5-7.
- Ruggiero, S., Pedreschi, D., & Turini, F. (2010). Data Mining for Discrimination Discovery. *TKDD*(4). doi:10.1145/1754428.1754432
- Salembier, C., Segrestin, B., Berthet, E., Weil, B., & Meynard, J.-M. (2018). Genealogy of design reasoning in agronomy: Lessons for supporting the design of agricultural systems. *Agricultural Systems*, 164, 277-290. doi:10.1016/j.agsy.2018.05.005



- Sengupta, S., & SukLee, W. (2014). Identification and determination of the number of immature green citrus fruit in a canopy under different ambient light conditions. *Biosystems Engineering*, 117, 51-61. doi:10.1016/j.biosystemseng.2013.07.007
- Shao, Y., Zhao, C., Bao, Y., & He, Y. (2012). Quantification of Nitrogen Status in Rice by Least Squares Support Vector Machines and Reflectance Spectroscopy. *Food and Bioprocess Technology*, 5(1), 100-107. doi: 10.1007/s11947-009-0267-y
- Shapiro, H. T., Backlar, P., Flynn, L. M., Brito, A., Greider, C. W., Brito, A., . . . Childress, J. F. (2001). *Ethical and Policy Issues in Research Involving Human Participants*. Maryland, USA.
- Sharma, A., & Kiciman, E. (2019). DoWhy: A Python package for causal inference. *Causal Data Science Meeting*.
- Shaw, G., Xu, Y., & Geva, S. (2008). Utilizing Non-Redundant Association Rules from Multi-Level Datasets. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (pp. 681-684). Sydney, Australia: IEEE, ACM. doi:10.1109/WIIAT.2008.39
- Shen, X. (2020). Challenges and Opportunities with Causal Discovery Algorithms: Application to Alzheimer's Pathophysiology. *Scientific Reports*, 10, 2975. doi:10.1038/s41598-020-59669-x
- Shen, Y., Liu, J., & Shen, J. (2010). The Further Development of Weka Base on Positive and Negative Association Rules. *International Conference on Intelligent Computation Technology and Automation* (pp. 811-814). Changsha, China: IEEE, ACM. doi:10.1109/ICICTA.2010.676
- Shi, D., Guo, Z., Johansson, K. H., & Shi, L. (2018). Causality Countermeasures for Anomaly Detection in Cyber-Physical Systems. *IEEE Transactions on Automatic Control*, 386-401.
- Shimizu, S., Hoyer, P. O., Hyvärinen, A., & J., K. A. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7, 2003-2030.
- Shukla, R., Yadav, V., Ram Pal, P., & Pathak, P. (2019). Machine Learning Techniques for Detecting and Predicting Breast Cancer. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 8(7), 2658-2662.
- Silverstein, C., Brin, S., Motwani, R., & Ullman, J. (1998). Scalable Techniques for Mining Causal Structures. *International conference on very large data bases (VLDB98)*, (pp. 594-605). New York, USA.
- Singh, A., Chaudhary, M., Rana, A., & Dubey, G. (2011). Online Mining of data to Generate Association Rule Mining in Large Databases. *International Conference on Recent Trends in Information Systems* (pp. 126-131). India: IEEE. doi:10.1109/ReTIS.2011.6146853
- Singh, K., Gupta, G., Tewari, V., & Shroff, G. (2018). Comparative Benchmarking of Causal Discovery Techniques. *ACM India Joint International Conference on Data Science and Management of Data* (pp. 46-56). Goa, India: ACM. doi:10.1145/3152494.3152499
- Sobhani, F., & Straccia, U. (2019). Towards a Forensic Event Ontology to Assist Video Surveillance-based Vandalism Detection. *CEUR Workshop*, 2396, pp. 30-47.

- Southern, M. (2020, January 23). *Search Engine Journal*. Retrieved January 30, 2020, from <https://www.searchenginejournal.com/googles-new-dataset-search-engine-comes-out-of-beta/344860/>
- Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, Prediction and Search 2nd Edition*. Massachusetts, USA.
- Stiles, P. G., & Boothroyd, R. A. (2015). Ethical Use of Administrative Data. *Actionable Intelligence: Using Integrated Data Systems to Achieve a More Effective, Efficient, and Ethical Government*, 125-155. doi:10.1057/9781137475114\_5
- Su, Y.-x., Xu, H., & Yan, L.-j. (2017). Support vector machine-based open crop model (SBOCM): Case of rice production in China. *Saudi Journal of Biological Sciences*, 27, 537-547.
- Sweeney, L. (2000). Simple Demographics Often Identify People Uniquely. 671.
- Sweeney, L., Abu, A., & Winn, J. (2013). Identifying Participants in the Personal Genome Project by Name. *SSRN Electronic Journal*. doi:10.2139/ssrn.2257732
- Tanner, A. (2013, April 25). *Harvard Professor Re-Identifies*. Retrieved August 21, 2020, from Forbes: Tech: <https://www.forbes.com/sites/adamtanner/2013/04/25/harvard-professor-re-identifies-anonymous-volunteers-in-dna-study/#5d02ce0992c9>
- Taylor, C. (2018, March 28). *Datamation*. Retrieved 01 09, 2020, from Structured vs. Unstructured Data: <https://www.datamation.com/big-data/structured-vs-unstructured-data.html>
- Thomas, D. R., Pastrana, S., Hutchings, A., Clayton, R., & Beresford, A. R. (2017). Ethical issues in research using datasets of illicit origin. 445-462. doi:10.1145/3131365.3131389
- Townsend, L., & Wallace, C. (2016). *Social Media Research: A Guide to Ethics*. University of Aberdeen.
- Vallor, S., & Rewak, W. J. (2018). *An Introduction to Data Ethics*.
- Van Der Walt, E., & Eloff, J. (2019). Identity deception detection: requirements and a model. *Information and Computer Security*, 26(4). doi:10.1108/ICS-01-2019-0017
- van Wel, L., & Royakkers, L. (2004). Ethical issues in web data mining. *Ethics and Information Technology*, 6(2), 129-140. doi:10.1023/B:ETIN.0000047476.05912.3d
- Varman, S. A., Baskaran, A. R., Aravindh, S., & Prabhu, E. (2017). Deep Learning and IoT for Smart Agriculture Using WSN. *IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* (pp. 1-6). Coimbatore, India: IEEE. doi:10.1109/ICCIC.2017.8524140
- Vayena, E., & Madof, L. (2019). *Navigating the Ethics of Big Data in Public Health - Oxford Handbooks*. Oxford. doi:10.1093/oxfordhb/9780190245191.013.31
- Wang, J., & Mueller, K. (2016). The Visual Causality Analyst: An Interactive Interface for Causal Reasoning. *IEEE Transactions on Visualization and Computer Graphics*, 22(1), 230 - 239. doi:10.1109/TVCG.2015.2467931

- Wang, Q., Liu, J., Chen, Z., Li, F., & Yu, H. (2018). A causation-based method developed for an integrated risk assessment of heavy metals in soil. *Science of The Total Environment*, 642, 1396-1405. doi:10.1016/j.scitotenv.2018.06.118
- Weyer, V. D., Waal, A. d., Lechner, A. M., Unger, C. J., O'Connor, T. G., Baumgartl, T., . . . Truter, W. F. (2019). Quantifying rehabilitation risks for surface-strip coal mines using a soil compaction Bayesian network in South Africa and Australia: To demonstrate the R2AIN Framework. *Integrated Environmental Assessment and Management*, 15(2), 190-208. doi:10.1002/ieam.4128
- Wheeler, J. (2018). Mining the First 100 Days: Human and Data. *Journal of Librarianship and Scholarly Communication*, 6(2), eP2235 | 1-23. doi:10.7710/2162-3309.2235
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques, 4th Edition*. Massachusetts, USA: Elsevier.
- Yadav, M. (2006). Legal and Ethical Aspects of Casualty Services in India. *Journal of Indian Academy of Forensic Medicine (JIAFM)*, 28(3), 114-120.
- Yu, P. S., Liang, B. C.-C., & Chen, M. S. (1998). *New York, USA Patent No. US005832482A*.
- Zhang, H., Yao, D. (., Ramakrishnan, N., & Zhang, Z. (2016, May). Causality reasoning about network events for detecting stealthy malware activities. *Computers & Security*, 58. doi:10.1016/j.cose.2016.01.002
- Ziauddin, Z., Kammal, S., Khan, K. Z., & Khan, M. I. (2012). Research on Association Rule Mining. *Advances in Computational Mathematics and its Applications (ACMA)*, 2(1), 226-236.