Analyzing offenses against life data:
a machine learning approach on data extracted from the Human Relations Area Files (HRAF) database

THESIS

submitted in partial fulfillment of the
requirements for the degree of

BACHELOR OF SCIENCE
in
COMPUTER SCIENCE

Author :	M. Lelasseux
Student ID :	s2479826
Supervisors :	Prof.dr. M.R. Spruit
	Prof.dr. M.C.A. Liem
	Dr. K.L. Syme

Leiden, The Netherlands, February 29, 2024

Analyzing offenses against life data:
a machine learning approach on data extracted from the Human Relations Area Files (HRAF) database

M. Lelasseux

Snellius building, Leiden University
P.O. Box 9500, 2333 CA Leiden, The Netherlands

February 29, 2024

Abstract

For this research machine learning techniques were used to analyse data extracted from the Human Relations Area Files (HRAF), a worldwide database with ethnographic collections. Much research has been conducted on other-directed harm (such as assault and homicide) and self-directed harm (such as self-harm and suicidal behaviours), but there has been little research on how to model available data on self-harm and other-directed harm and how to predict future events where self-harm and assault could occur using machine learning methods. The predictions of these events could help with preventing them and are relevant for educational purposes, for example for police training, and for psychologists to better understand the roots of self-harm and other-directed harm. Other-directed harm and self-directed harm have been framed by evolutionary researchers as bargaining strategies to influence conflict outcomes. This researched aimed to investigate what machine learning techniques can be implemented to analyse the differential causes and social contexts of other-directed harm and self-directed harm. For this analysis the CRISP-DM method was used. The HRAF is coded at the paragraph-level with OCM codes, which stands for 'outline cultural materials'. A datafile containing all the texts on Offenses Against Life (OAL) (OCM code 682) was used to conduct analyses. The covariation of OCM codes related to self-directed harm, other-directed harm, and types of conflicts, were analysed using machine learning techniques to target different OCM codes. Regression methods were used to research connections between the OCM codes and applied

on one-hot-encoded data (all the OCM codes were binary coded), with various models such as Bayesian Ridge, Light Gradient Boosting, and Orthogonal Matching Pursuit being the best models. From there, feature importance plots were created, each feature importance plot shows the top 10 of most important predictor variables. Lastly, the hierarchy of OCM code 762 (Suicide) was determined and cluster analysis was done on the OAL data file. No cluster forming was found between the individual cases in the OAL data file, nor were domain experts able to identify clusters between the OCM codes without information on the PCA components. For OCM code 762 (Suicide) are the most important variables impacting whether suicide would, or would not occur: Mortality, Special Burial Practices and Funerals, Sexual Stimulation, Personality Disorders, Sexuality, Physical Descriptions, Termination of Marriage, Conception, Pharmaceuticals and Cult of the Dead.

Keywords: Data-analysis, CRISP-DM, Self-harm, Other-directed harm, Suicide, Bargaining model

Contents

Chapter 1

Introduction

A decrease in violent crimes can be seen in recent years, both in the US[12] and in The Netherlands for instance[51]. However, an increase in suicides and suicidal behaviour can be seen across the globe in various studies (see [40], [20], and [4]), but it remains uncertain why suicide rates have risen as violent crimes rates have declined. Moreover, it remains unknown what situations lead to self-harm, other-directed harm, or a co-occurrence of self-harm and other-directed harm. Much research has been done on other-directed harm (such as assault and homicide) and self-directed harm (such as self-harm, suicide, and suicidal behaviours) events, but there has been little research on how to model available data and predict future events, specifically which dimensions of conflict predict other-directed harm and which dimensions predict self-directed harm. For this research machine learning techniques were used to analyse data extracted from the Human Relations Area Files (HRAF), a worldwide database with ethnographic collections covering all aspects of cultural and social life. The aim of this research is to predict other-directed harm and self-directed harm events in the future and therefore help preventing them.

Conflict does not have one definition. Jacoby (2008) discusses interdisciplinary approaches to violence and conflict, different dimensions of conflict, and various definitions of conflict used by researchers. He defines conflict as a situation in which two (or more) parties have mutually incompatible goals or perceive as such[13]. Evolutionary theorists, on the other hand, make a distinction between genetic conflict and overt behavioural conflict, which are related by separate phenomena. Genetic conflict refers to the fact that the optimal fitness outcomes (i.e., future genetic representation) of one or more organisms threatens to harm the fitness outcomes

of other organisms and occurs between genetic relatives[11, 47], sexual partners[29], and individuals who are not genetically related[3]. Overt conflict concerns the way organisms impose costs on each other to resolve conflicts between one another in their favour, this can include yelling and fighting for example, or more subtle behaviours such as avoidance (see [5] and [30]). Evolutionarily, both other-directed aggression and self-harm are associated with social conflict and are seen as strategies to influence how others behave[10, 38]. Sources of conflicts can vary between division of resources (i.e., land, food, money), status, or identity for example.

The aim of this interdisciplinary research is to investigate similarities and differences between instances of other-directed harm only (such as assault and homicide), self-directed harm only (such as self-harm and suicidal behaviours), and instances where other-directed harm and self-directed harm co-occur.

The main research question is:

RQ1: What machine learning techniques support the analysis of offenses against life data extracted from the HRAF, and can help predict future events where other-directed harm and self-harm could occur?

The domain questions (which are focused on violence studies) of this research are:

RQ2: What correlations, relations and patterns can be found in the provided data, focusing on suicide events and different dimensions of aggression?

RQ3: What information and relationships such as clusters can be deducted from the OCM codes given in the provided data?

RQ4: What variables are important when predicting specific events such as a suicide event or other cases important to other-directed and self-directed harm?

Chapter 2

Theory: violence studies

The specific behaviours individuals exhibit in overt behavioural conflict have also been termed bargaining strategies by evolutionary and game theorists[16, 38]. In the subsections below the bargaining model and other information on other-directed harm and self-harm will be given.

2.1 Conflict resolution strategies

As said before in section 1, the Introduction, Jacoby discusses in his book[13] the different ideas and dimensions of conflict. There are different strategies that can be applied when trying to solve a conflict. Persuasion, deception, physical aggression, offering or withholding resources, and withholding cooperation are all known as conflict resolution strategies[46]. These strategies do not directly imply that when resolving a conflict every party involved wins, even though it could be a possible outcome.
Sell et al.[38] studied in their research three components of bargaining power in males and females, namely: fighting ability, coalitional strength, and mate value. Fighting ability and mate value reliably predicted aggression, aggressive attitudes, and delinquent behaviour in both men and woman[38]. Sell was also the one who mentioned in various of his researches (see [39], [37], and [36]) that upper-body strength is a measure to determine whether a person would survive in a group or in the wild as greater upper-body strength is a predictor of violent aggression. This is because those who are physically stronger can more effectively use their strength. Analyses of ancient human weapons show that they all depended on the upper body strength for effectiveness[38].

2.2 Self-harm and other-directed harm

Self-harm can be described as a wide range of behaviours and intentions to harm oneself, including: attempted hanging, impulsive self-poisoning, and superficial cutting in response to intolerable tension[41]. When looking at the bargaining theory, self-harm can be seen as risk-taking behaviour. This could range from drinking too much[7] (and therefore being more prone to accidents), to committing suicide. Most self-harm and suicidal behaviour is non-fatal. As with suicide, definitions of self-harm vary greatly between countries. In some countries for instance, running away from one's home is a form of suicidal behaviour and therefore self-harm[46]. This is an example of a type of conflict with which suicidal behaviour occurs. There are many other types of conflict associated with suicidal behaviour, such as forced marriages, academic pressure, bullying, and sexual assault[9]. Main predictors of suicidal behaviour are extreme conflict[9] and powerlessness[45].

Other-directed harm can be described as harm caused by an actor towards another person, with examples such as forcing someone to have sex, robbing or mugging someone[33]. Other-directed harm could also be verbally abusing another person[54], or even only looking at another person with the intention to hurting them or letting them know you are angry. An example of this is looking angry at another person when they jump the queue in the grocery store.

2.3 Bargaining strategies

People engage in all kinds of behaviours to influence the outcomes of conflict, for instance giving them information, withholding information, or negotiating between parties[45]. These are all bargaining strategies, hence the bargaining model. Depression in women (associated with self-harm and suicidal behaviour) can co-occur with aggression (seen as other-directed harm), this ranges from being powerless as a component of depression and anger, to for example anger occurring as a result of expectations being violated[28]. The analyses done on self-harm in this research are focused on suicide and (non-fatal) suicidal behaviours.
The bargaining model proposes that suicidal behaviour and suicide attempts are a costly signal of need of one party to another, with completed suicides an unfortunate byproduct[45]. Though self-harm is distinct from

suicide, self-harm is the biggest known risk factor for completed suicide[45]. It is also so that in many cases where bargaining strategies are applied, both parties need each other, and that committing suicide is therefore only a threat. In the example of the girl threatening her family to commit suicide after being given away for marriage, the girl does not really want to die as she wants to marry her partner of choice. At the same time does her family not want her to die as they *need* her, they cannot give a dead girl away to another man[46].

2.4 Intergoup violence

Intergroup violence is defined as an act perpetrated by a member of one social group upon a (or multiple) member(s) of another social group[6]. Research[6] states, however, that the basic premise of the social identity approach suggests that any act of violence done by a group member could be either interpersonal or intergroup in nature. Now Van Vugt[21, 50] argues that the human psychology has been shaped by intergroup competition and conflict. In other words, he states that the evolutionary history of coalitional aggression between groups of men may have resulted in sex-specific differences in the way groups, specifically outgroups, are perceived. This creates ingroup versus outgroup tendencies, which are still observable nowadays and in modern history. An important implication of the warfare hypothesis that Van Vugt obtained, is that intergroup violence may have affected the evolved psychologies of both men and women differently[49]. Intergroup aggression has historically involved rival coalitions of men fighting over scarce resources, such as land (agriculture) or cattle breeding for example. As a consequence, this aspect of human coalitional psychology may be more pronounced among men, hence the term *male* warrior hypothesis[49]. Research[49] states that the male warrior hypothesis predicts potential sex differences in intragroup dynamics as a result of intergroup threat. Vugt et al.[50] says that being successful in intergroup competition requires membership of a strong, cohesive, and coordinated ingroup.

2.5 Internalizing and externalizing behaviour

Behaviour in people can be categorized into various categories, with two broad categories being: internalizing behaviour and externalizing behaviour[25]. Internalizing behaviour (defined as an over-control of emotions) consists

of for example: anxiety, depression, somatic complaints without known medical basis, and social withdrawal from contact[24]. Internalizing behaviour tends to be found more in women than in men[17].
Externalizing behaviour is known as acting out, including aggressive and destructive behaviours. Externalizing symptoms include for example impulsivity, hyperactivity, and temper tantrums[24]. In general do more men show externalizing behaviour than women[17]. An explanation to why men lean more towards externalizing behaviour and women more to internalizing behaviour can be explained by (among other things): upper body strength[38].

Sex differences

Research[26] has shown that consistent cross-national risk factors of suicidal behaviour and suicide include being female, younger, less educated, unmarried, and having a mental disorder. As discussed before, men are more physically aggressive and they die by suicide more often than women. However, being suicidal (and showing suicidal behaviour) and non-fatal suicide attempts are factors more associated with women. Thus, taken together, the evidence suggests that physical aggression and suicide death is associated with the male sex, whereas non-fatal suicidality is associated with being female and being young[26].

Chapter 3

Method: data science process

This chapter will first briefly explain some things on machine learning, with a focus on the programs used during this project and the data available. Then a detailed explanation on the method of this research will be given. The first CRISP-DM cycle analysed the OCM codes found in the Offenses Against Life (OAL) dataset. CRISP-DM stands for Cross-Industry Standard Process for Data Mining[44]. With the second CRISP-DM cycle, a more in depth analysis was done on the results found during the first cycle. A hierarchy of target OCM code 762 (Suicide) was created and cluster analysis was done on the OAL data file.

3.1 Machine learning

Machine learning can be described as the technique that improves system performance by learning from experience via computational methods[2]. The main task of machine learning in general is to evaluate data and then to develop learning algorithms that can then build models from the provided data. A correctly implemented model can make predictions on new observations. Machines learning is thus the subject of learning algorithms.

In machine learning there are two types of models, supervised and unsupervised learning models[2]. Supervised models are sub-categorised as a regression model or a classification model. Unsupervised models are sub-categorised into clustering, dimensionality reduction, association rule learning, principle component analysis (PCA), and t-distributed stochastic neighbour embedding (t-SNE)[19]. The first analyses of this research were

done using supervised models, regression models to be precise. The second analyses done during the second CRISP-DM cycle were done using unsupervised models, namely clustering and Principle Component Analysis (PCA). The reason that regression models were chosen to analyse the data is because for this research we are mainly interested in the relationships between OCM codes. Regression models can be used to understand the relationship between variables (in this case OCM codes) and identify important predictors for example[27]. However, for the second CRISP-DM cycle we were interested in whether there is any cluster forming between the OCM codes. As our unsupervised models, both clustering and principle component analysis (PCA) were chosen for this CRISP-DM cycle. Clustering is the act of grouping variables together into subsets, in such a manner that similar variables are grouped together, while different instances belong to other groups[34]. The Principal Component Analysis (PCA) method assists researchers in determining the optimal combination of data that most accurately captures the concept they aim to assess. By providing distinct components, PCA condenses the dimensions of a multivariate dataset into a reduced number of dimensions, effectively reducing its dimensionality[48]. In subsections 3.3.4, Modeling, and 3.3.5, Evaluation, a more detailed explanation of the models used will be given.

Hierarchy & clusters

One of the objects of this research is to determine whether there are any hierarchies, dependencies and clusters present in the provided data. A hierarchy is a data structure in the form of a tree where items are linked to each other in parent-child relationships. A dependency means that a variable or function is dependent on another variable or function, they are thus linked to each other. A cluster is a group of variables grouped together with similar characteristics, or when different variables are occurring closely together.

Programs

Multiple programs and libraries were used to obtain the desired results for this project. In this chapter a brief outline of the Python libraries Pandas and PyCaret, and the program RapidMiner can be found.

Pandas

Pandas is an open-source, fast, and flexible analysis tool in Python. It is used to do simple tasks such as finding duplicates in two columns of a file, or filtering rows from a data file. Pandas is one of the most used libraries in Python for data sets and therefore fairly standard in programming. Alongside pandas, other libraries are also frequently used in programming such as numpy, matplotlib and sklearn.

PyCaret

PyCaret is an open-source, low-code machine learning library in Python[1]. It can evaluate data, run different models, and evaluate multiple models at the same time. PyCaret is the supervised learning module used in this study to determine the best model to predict certain target values (the chosen OCM codes) and evaluate the provided data. It was chosen as it is a fast and reliable method to obtain results.

RapidMiner

The last program used for this research was RapidMiner. RapidMiner is a tool with which data mining and machine learning procedures can be done[22]. RapidMiner itself is written in Java but the program does not require any code or programming language. See Figure 3.1 to see a preview of how the program looks. This window shows how the decision tree of target 762 (Suicide) was obtained, used to investigate dependencies between variables and answer domain questions 2 and 4.

3.2 The data

This section will provide an in depth overview of the data used.
At first hand two data files, also called datasets, were provided by the ISGA department of Leiden University. The data was extracted from the HRAF[8] and provided in an `.xlsx` file and a `.csv` file. The original data files and all the newly created/adapted data files can all be found on GitHub, see the link in the Appendix on page 65 to access the data and code used for this project.

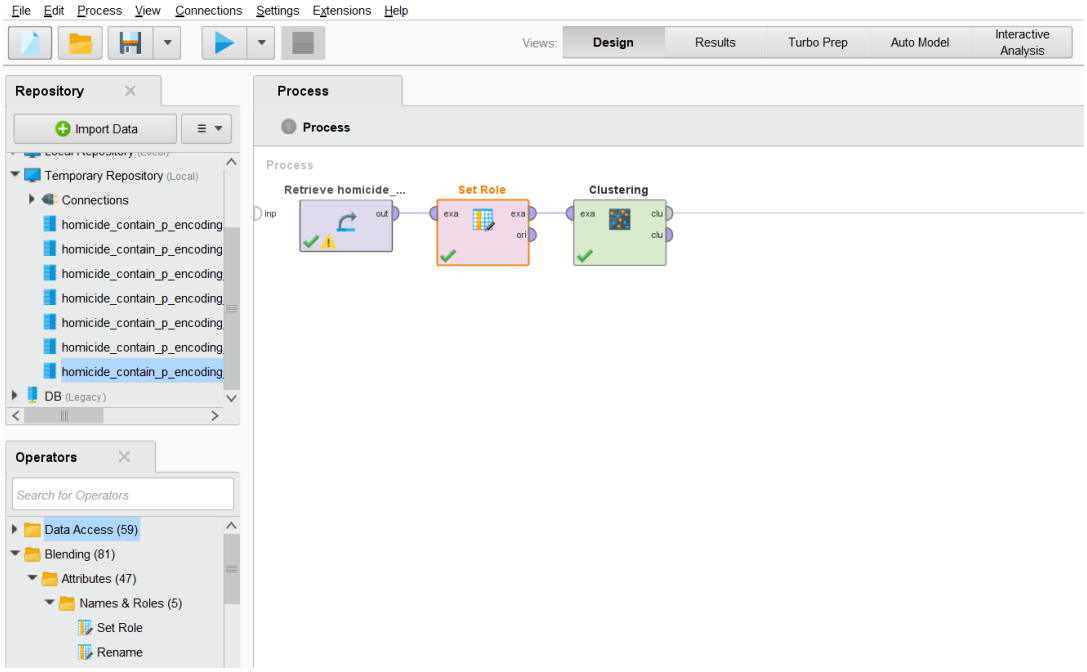

Figure 3.1: *The outlook of the program RapidMiner, a data mining and machine learning tool. Shown is a window of how the hierarchy (decision tree) of OCM code 762 (Suicide) was obtained*

Data description

The OAL data file and the suicide data file consist of different information. The `homicide.xlsx` file contains 7623 paragraphs/list-items/enotes/etc. of offenses against life data. This was the name of the file when given to the researchers, except that not all the cases in this file were explicit *homicide* cases, some cases only consisted of other-directed harm events. Therefore, it was chosen to call this data file the 'Offenses Against Life' data, in short OAL. From now on this data will be called the OAL data (file), but the name of the file when doing the analyses remains 'homicide', for convenience purposes only.
In Figure 3.2 a snippet of the original OAL datafile can be found.

This data file contained 7623 lines of data and 26 columns with more information on the cases mentioned. The first column showed a description of the text (data type: string), in the second column (with the columns being semicolon separated) the OCM codes (data type: list), which were used to code the texts, were noted. On average every text contained five different OCM codes, the OCM codes were separated by a comma. The third column showed what type of text the line was, among other things

	A	B	C	D	E	F	G	H
1	text	ocms	type	pageEid	prevPage	nextPage	sreid	sreprev
2	{{682 684 685 847}} Homicides are condemned tc	#682 #683 #684 #685 #689 #847	p	fa08-003-005132	fa08-003-005067	fa08-003-005176	fa08-003-005164	fa08-003-005157
3	{{682}} Murder brought exclusion from the grouf	#425 #682	p	fa16-001-004598	fa16-001-004548	fa16-001-004680	fa16-001-004672	fa16-001-004666
4	{{682}} Murder, as a matter of fact, seems to hav	#152 #157 #682	p	fa16-002-004212	fa16-002-004170	fa16-002-004252	fa16-002-004236	fa16-002-004229
5	{{181 682}} One can best explain the horror that	#181 #682	p	fa16-002-004212	fa16-002-004170	fa16-002-004252	fa16-002-004245	fa16-002-004236
6	The murderer was formerly, we are told, exclude	#181 #682	p	fa16-002-004252	fa16-002-004212	fa16-002-004282	fa16-002-004270	fa16-002-004245
7	Even today, it is very rare that a murderer goes t	#181 #682	p	fa16-002-004252	fa16-002-004212	fa16-002-004282	fa16-002-004276	fa16-002-004270
8	{{682}} If, later on, chance brings the murderer in	#682	p	fa16-002-004282	fa16-002-004252	fa16-002-004318	fa16-002-004293	fa16-002-004276
9	Sometimes, however, the murderer is permitted	#114 #171 #682 #783	p	fa16-002-004282	fa16-002-004252	fa16-002-004318	fa16-002-004300	fa16-002-004293
10	{{783 682}} This text by itself is very clear. If one	#682 #783 #825	p	fa16-002-004318	fa16-002-004282	fa16-002-004348	fa16-002-004332	fa16-002-004300
11	{{764}} The man brought back must be the beare	#263 #682 #761 #764 #783 #826	p	fa16-002-004348	fa16-002-004318	fa16-002-004391	fa16-002-004360	fa16-002-004332
12	{{602}} If it is a man of his family, an uncle, broth	#602 #682 #783	p	fa16-002-004348	fa16-002-004318	fa16-002-004391	fa16-002-004377	fa16-002-004360
13	The day after the crime, in principle (in fact, as s	#682 #783	p	fa16-002-004348	fa16-002-004318	fa16-002-004391	fa16-002-004385	fa16-002-004377
14	[// [// __Amma__ __wa[unavailable]a__ __say	#682 #783	p	fa16-002-004391	fa16-002-004348	fa16-002-004465	fa16-002-004403	fa16-002-004385
15	that is to say: God, see to it that this (the murde	#682 #783	p	fa16-002-004391	fa16-002-004348	fa16-002-004465	fa16-002-004416	fa16-002-004403
16	The slaughtered animal is then stripped of its hi	#682 #783	p	fa16-002-004391	fa16-002-004348	fa16-002-004465	fa16-002-004422	fa16-002-004416
17	The family of the murderer is still obliged to han	#553 #582 #672 #682 #768	p	fa16-002-004391	fa16-002-004348	fa16-002-004465	fa16-002-004429	fa16-002-004422
18	{{682 672 582}} The terms in which our informan	#582 #672 #682	p	fa16-002-004391	fa16-002-004348	fa16-002-004465	fa16-002-004446	fa16-002-004429

Figure 3.2: *Snippet of the original OAL data file, with no data preparation done yet. The first eight columns and eighteen rows are shown*

were the types either a 'p' (for paragraph), list-item, enote or quote (again data type string). With the frequencies of each of these types being 92,4% paragraphs, 4,7% list-items, 1,7% enotes, and 0,5% quotes. Some of the other 23 columns left (such as author, culture, etc.) could be useful in future research. For this study only the second and third columns were used, an explanation of how the data was prepared for the analyses can be found in section 3.3.3, Data preparation.

The other data file, containing information on suicide events was in another format than the OAL data file. This file consists of twelve columns (semicolon separated) and 245 rows, thus 245 cases. The first column is the ID of the text, other information mentioned are the cultures (including culture code), region and location of each case, a description of the case, whether suicide was present yes (1) or no (0), some comments, information on the creation of the file and the creation ID.
A snippet of the data file can be seen below in Figure 3.3.

	A	B	C	D	E	F	G	H	I	J	K	L
1	id	culture	culture_code	region	location	description	probability	position	comment	created	modified	created_by_id
2	1	Akan	FE12	West Africa	Ghana and Cote d'Ivoire	The Akan are peoples	1	8.0,-2.0	Language: Kwa branch Niger-Congo family	33:03.8	33:03.8	1
3	2	Azande	FO07	Central Africa	Sudan, CAR, Democratic Republic of the Congo	The Azande are a large	1	4.5,21.0		33:04.0	33:04.1	1
4	3	Bagisu	FK13	East Africa	Mount Elgon in eastern Uganda.	The Bagisu or Gisu live	0	1.0,34.5		33:04.1	33:04.1	1
5	4	Banyoro	FK11	East Africa	Uganda	The Banyoro live large	0	1.7,31.0		33:04.2	33:04.2	1
6	5	Bena	FN31	East Africa	Tanzania	The Bena are agricultu	0	-9.25,35.0		33:04.2	33:04.3	1
7	6	Dogon	FA16	West Africa	Mali, Burkina Faso	The Dogon are a grou	1	14.0,-3.0	Language: Mande and Gur branches of the Niger-	33:04.3	33:04.3	1
8	7	Ganda	FK07	East Africa	Uganda	The Ganda, who refer	1	0.5,33.0		33:04.4	33:04.4	1
9	8	Hausa	MS12	West Africa	Nigeria and Niger	The Hausa constitute	1	12.5,8.0	Language: Chadic group Afro-Asiatic family	33:04.4	33:04.4	1
10	9	Igbo	FF26	West Africa	Nigeria	Igbo is the language	0	6.0,7.0	Language: Igbo is a Benue-Congo branch Niger-C	33:04.5	33:04.5	1

Figure 3.3: *Snippet of the original suicide data file. All the columns and the first ten rows are shown.*

OCM codes

The most interesting column in the OAL file is the OCM column. OCM stands for 'outline cultural materials', meaning that every code stands for a description of a variable. A few examples: the code 682 stands for 'Offences against life', all the cases in this file thus contained at least OCM code 682 (Offenses Against Life). Code 847 means 'Abortion and Infanticide', so if a case contained abortion and/or infanticide, then it would be labeled with OCM code 847 (Abortion and Infanticide).
The `suicide.csv` data file was not in the same format as the OAL file, as can be seen in the previous section. The suicide file did contain text fragments as did the OAL file, but the cases in the suicide file did not contain any OCM codes apart from OCM code 762, which stands for 'Suicide'. Therefore, this data file was not used for any further research apart from the first few steps of the CRISP-DM method, business understanding and data understanding.
The different OCM codes did not differ in value, meaning that no OCM code is stronger than another one. Every OCM could also only be mentioned once per case, so no OCM codes were mentioned twice in the same line.

3.3 CRISP-DM

As can be seen in Figure 3.4, the CRISP-DM cycle consists of a process of six steps, where some steps are related to each other.

The subsections below explain each step of the CRISP-DM cycle separately.

3.3.1 Business understanding

To start with the CRISP-DM process a correct business understanding has to be established. Meaning that the researcher has to have an understanding with the client, what they want as an output for example, and what data needs to be analysed.
In section 1, Introduction, the research questions were established. As said previously, the 'business questions' (or domain questions) are the same as the questions asked by the Institute of Governance and Global affairs (ISGA). The questions are based on a focus group session with domain experts. The business understanding is therefore finding answers to these questions domain questions, stated in section 1, Introduction. To do this,

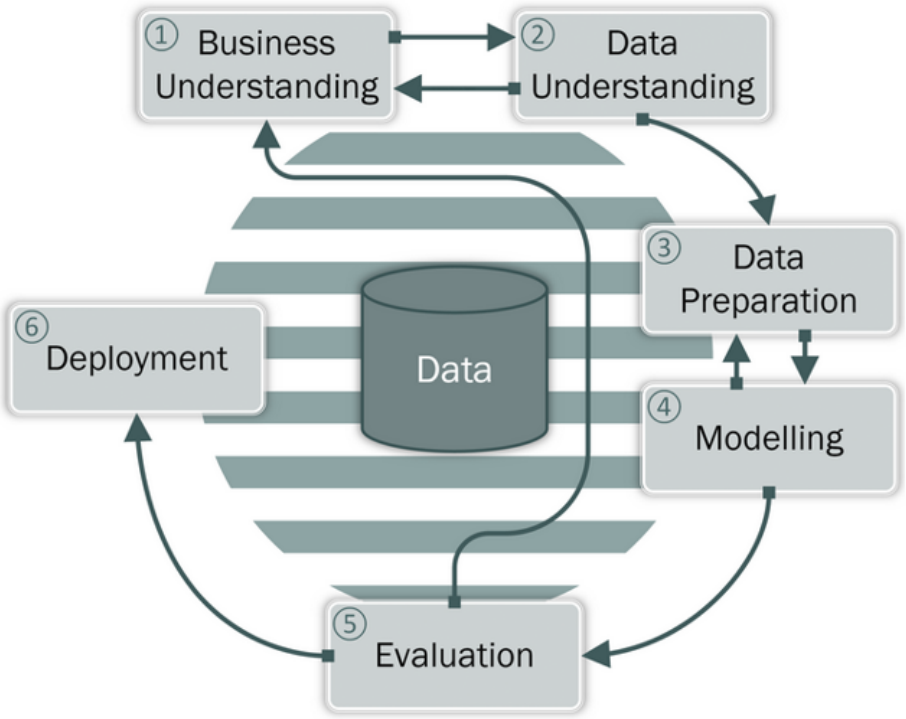

Figure 3.4: *CRISP-DM process with the relationships between each phase[43]*

regression models will be run on thirteen different target variables, the OCM codes, chosen by domain experts. The chosen OCM codes are: 762 (Suicide), 578 (Ingroup Antagonisms), 627 (Informal Ingroup Justice), 628 (Inter-community Relations), 672 (Liability), 681 (Sanctions), 683 (Offenses Against the Person), 684 (Sex and Marital Offenses), 685 (Property Offenses), 728 (Peacemaking), and 754 (Sorcery) for the first file called the 'p' file, and the groups 'War and Peacemaking' and 'Drugs and Alcohol' for the second data file, called the 'combined' file (explained later on). All the OCM codes that were targeted are related to conflicts as either 'bargaining strategies', for instance OCM codes 762 (Suicide), 754 (Sorcery), and the group War and Peacemaking, or types of conflict (OCM codes 681 (Sanctions), 578 (Ingroup Antagonisms), and the OCM codes for different types of offenses: 683, 684, and 685, respectively Offenses Against the Person, Sex and Marital Offenses, and Property Offenses). Other reasons to why these target variables were chosen was because they were related to types of relations (OCM code 628 (Inter-community relations)), or are correlates of violent conflict (such as OCM code 672 (Liability), and the group

Drugs/Alcohol).

3.3.2 Data understanding

Data understanding means that the data needs to be collected, then explored and a detailed description of the data needs to be provided for others to understand the process.
The data provided are two datasets, one is a dataset on offenses against life events, containing various data. This dataset is considered quite big as it consists of 7426 lines. The other dataset is a dataset on suicide, it is smaller than the OAL data file, and consists of a combination of three different files.
As said previously, only the OAL data was used for the analyses. With a Python code (see the Appendix, page 67) were all the suicide cases extracted from the OAL data file to better understand the data. Then all these cases were read and hand-coded, meaning that the age, sex and weapon of the perpetrator and victim were written down, as well as if religion was playing a role in the (attempted) homicide and/or (attempted) suicide. Further data understanding was done by skimming through the suicide data file, to understand the different cultures and beliefs in other countries. The data mining goals of the domain questions are finding the right model for each target variable, trying to make predictions of future events (prediction analysis), and trying to find patterns within the data (pattern recognition). The PyCaret program will be used to create a model to analyse the before mentioned chosen targets, and provide a feature importance graph, which indicates what variables are important when predicting the target variable. In the next section, section 3.3.3 Data preparation, an explanation will be given on what information extracted from the OAL data file was used for this research.

3.3.3 Data preparation

From the OAL dataset all the lines containing suicide were extracted. Then, the newly created file was cleaned up, columns with not-needed data were removed (such as the page numbers in books for instance), and lines in the file that would give an error were fixed. This was all done to do some test runs with the written code and programs used.
When analysing the complete OAL dataset it was also cleaned up, the code was run multiple times to see what lines would return an error and then these lines would be removed or fixed so that the program would run

without any errors.
After running the program with the complete OAL dataset, it was discovered that some authors put their text multiple times in the HRAF database, leading to invalid results. To fix that, further data preparation was done by filtering out all the paragraphs. Every text containing a 'p' for paragraph in the third column would be selected, then other texts which were marked with 'list-item', 'enote' or 'quote' were removed from the data file, resulting in a new data file with 7046 cases left. From now on this file will be called the 'p' file (p stands for 'paragraph'). The list-items, enotes, quotes, etc. were removed because these texts were not relevant for the research on events where other-directed or self-directed harm was involved. When this was done, PyCaret was run with target variable 762 (Suicide) and the Bayesian Ridge model was evaluated, as this model was chosen as the best model to predict the input target variable 762 (Suicide). When creating a feature importance plot it was discovered that only the column containing the OCM codes was useful to predict the target variable. For this see Figure 3.5.

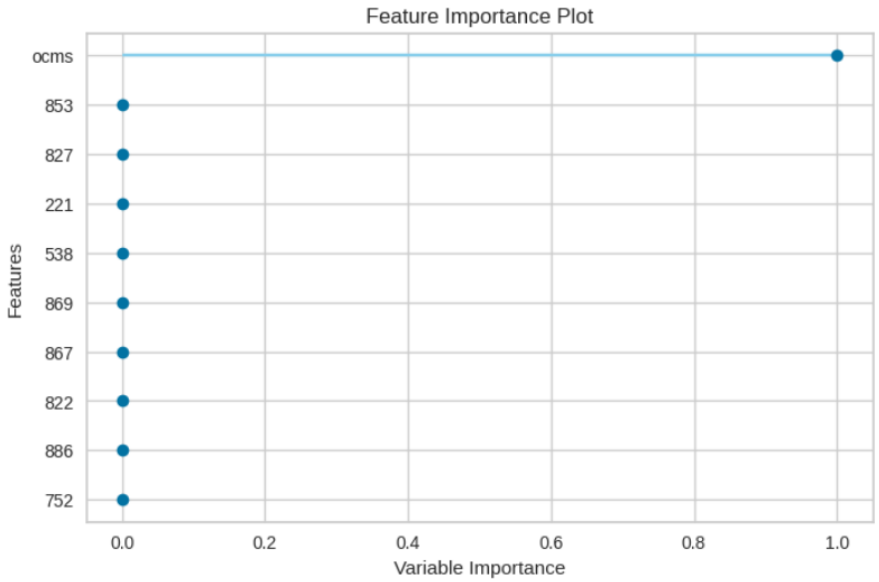

Figure 3.5: *Feature importance plot of target variable 762 (Suicide), input file: the original OAL data file*

After learning that only the OCM codes were useful, all the other available columns could be discarded. To research the OCM codes, the one-hot encoding method was chosen, which will be explained in the next subsec-

tion.
For the second CRISP-DM cycle, some further data preparations were done to run the RapidMiner program and create a decision tree for target 762 (Suicide). The program would not run correctly with the one-hot encoded data file called `homicide_contain_p_encoding.csv` (see the next subsection), as all the variables were numeric. To change this, the data in the dataset was changed from 1's and 0's to respectively true and false. This was done in Excel by using the 'replace' tool and replacing every 1 with true and every 0 with false.

One-hot encoding

To do analyses on the OCM codes of the data, the one-hot encoding method was used. The code to turn the data into a one-hot encoding data file can be found in the Appendix on page 66. The one-hot encoding process was done in R, and it would create a column for every OCM code appearing in the data file. R was used as the code was already written in R, so no further editing of the program was required. If a text would be coded with the following OCM codes such as: `682, 762, 155`, then these columns would be marked with a '1'. All the other columns with codes such as `761` and `113` for example, that would not appear in the text, would then be coded with a '0'. A snippet of the data before and after the encoding is shown in Figures 3.6 and 3.7. Note that both the original data file and the data file used for the analyses can be found in its entirety on GitHub.

	A	B	C	D	E	F	G	H
1	text	ocms	type	pageEid	prevPage	nextPage	sreid	sreprev
2	{{682 684 685 847}} Homicides are condemned tc	#682 #683 #684 #685 #689 #847	p	fa08-003-005132	fa08-003-005067	fa08-003-005176	fa08-003-005164	fa08-003-005157
3	{{682}} Murder brought exclusion from the grou	#425 #682	p	fa16-001-004598	fa16-001-004548	fa16-001-004680	fa16-001-004672	fa16-001-004666
4	{{682}} Murder, as a matter of fact, seems to hav	#152 #157 #682	p	fa16-002-004212	fa16-002-004170	fa16-002-004252	fa16-002-004236	fa16-002-004229
5	{{181 682}} One can best explain the horror that	#181 #682	p	fa16-002-004212	fa16-002-004170	fa16-002-004252	fa16-002-004245	fa16-002-004236
6	The murderer was formerly, we are told, exclude	#181 #682	p	fa16-002-004252	fa16-002-004212	fa16-002-004282	fa16-002-004270	fa16-002-004245
7	Even today, it is very rare that a murderer goes t	#181 #682	p	fa16-002-004252	fa16-002-004212	fa16-002-004282	fa16-002-004276	fa16-002-004270
8	{{682}} If, later on, chance brings the murderer ir	#682	p	fa16-002-004282	fa16-002-004252	fa16-002-004318	fa16-002-004293	fa16-002-004276
9	Sometimes, however, the murderer is permitted	#114 #171 #682 #783	p	fa16-002-004282	fa16-002-004252	fa16-002-004318	fa16-002-004300	fa16-002-004293
10	{{783 682}} This text by itself is very clear. If one	#682 #783 #825	p	fa16-002-004318	fa16-002-004282	fa16-002-004348	fa16-002-004332	fa16-002-004300
11	{{764}} The man brought back must be the beare	#263 #682 #761 #764 #783 #826	p	fa16-002-004348	fa16-002-004318	fa16-002-004391	fa16-002-004360	fa16-002-004332
12	{{602}} If it is a man of his family, an uncle, broth	#602 #682 #783	p	fa16-002-004348	fa16-002-004318	fa16-002-004391	fa16-002-004377	fa16-002-004360
13	The day after the crime, in principle (in fact, as s	#682 #783	p	fa16-002-004348	fa16-002-004318	fa16-002-004391	fa16-002-004385	fa16-002-004377
14	[// [// __Amma__ __wa[unavailable]a__ __say	#682 #783	p	fa16-002-004391	fa16-002-004348	fa16-002-004465	fa16-002-004403	fa16-002-004385
15	that is to say: God, see to it that this (the murde	#682 #783	p	fa16-002-004391	fa16-002-004348	fa16-002-004465	fa16-002-004416	fa16-002-004403
16	The slaughtered animal is then stripped of its hi	#682 #783	p	fa16-002-004391	fa16-002-004348	fa16-002-004465	fa16-002-004422	fa16-002-004416
17	The family of the murderer is still obliged to han	#553 #582 #672 #682 #768	p	fa16-002-004391	fa16-002-004348	fa16-002-004465	fa16-002-004429	fa16-002-004422
18	{{682 672 582}} The terms in which our informan	#582 #672 #682	p	fa16-002-004391	fa16-002-004348	fa16-002-004465	fa16-002-004446	fa16-002-004429

Figure 3.6: *Snippet of the OAL data file before the one-hot encoding method was applied, shown are the first eighteen rows and eight columns*

Figure 3.7 shows a snippet of the `homicide_contain_p_encoding.csv` file, this file was used for the analyses done with the single OCM codes as target variables. The data file thus contained a header row with numeric integers (the OCM codes) and from there every row was binary coded for

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	682	683	684	685	689	847	425	152	157	181	114	171	783	825	263	761	764	826
2	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
5	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
6	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
7	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0
10	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
11	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1
12	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
13	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
14	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
15	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
16	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
17	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 3.7: *Snippet of the homicide_contain_p_encoding.csv file, after the one-hot encoding method was applied. This data was used for the analyses done. Shown are the first eighteen rows and eighteen columns*

every OCM code present in the case.
Some data analysis was also done on grouped OCM codes. The 529 OCM codes were grouped together in 93 different groups with titles such as 'war and peacemaking,' 'machines and tools,' and 'religious offenses'. The groups were created and chosen by Dr. K.L. Syme and M. Lelasseux. OCM codes were grouped together based on categories and subcategories, partially already categorised by the HRAF and partially chosen by the previously mentioned people. Other OCM codes were grouped together based on relationships between the meanings of the codes, for instance OCM codes 462 (Division of Labor by Gender) and 890 (Gender Roles and Issues) were put in the group 'Division of labor by gender' together. The code which was done to do this can be found on GitHub with the title `create_combined_file.py`. The data was then again one-hot encoded, the 7046 original cases were kept the same, only this time instead of coding a single OCM code such as 682 (Offenses Against Life) with a 1, an entire group would be coded with a 1 if a text contained any OCM code of that group. If none of the OCM codes in a group were found in a case, then that group would be coded with a 0.

Exceptions

In the original OAL data file there were 7623 cases filed. Of these 7623 cases, 59 cases contained the 847 OCM-code, standing for Abortion and Infanticide. Abortion is one of the exceptions in the data file, meaning that in several countries abortion is seen as an offence against life, where in other countries abortion is legal, meaning people can have an abortion and

not be seen as a murderer. Because the portion of abortion cases was 59 out of 7623 cases, the cases were not removed from the dataset, because it was believed that this data would not influence the outcome of the analyses and would not have a significant effect.

3.3.4 Modeling

In machine learning modeling multiple types of data are used. We have train data, test data and validation data.
To try out the written code a test data file was created, which contained ten rows and ten columns with fictional (binary encoded) data. The test file was run and a feature importance graph was deployed. From there the `homicide_contain_p_encoding` file was run multiple times, each time with a different target variable. Thirteen different targets were chosen beforehand, eleven were for the `homicide_contain_p.csv` file and two were for the `combined_homicide_data.csv` file. The chosen targets are (as mentioned before): 762 (Suicide), 578 (Ingroup Antagonisms), 627 (Informal Ingroup Justice), 628 (Inter-community Relations), 672 (Liability), 681 (Sanctions), 683 (Offenses Against the Person), 684 (Sex and Marital Offenses), 685 (Property Offenses), 728 (Peacemaking), and 754 (Sorcery) for the 'p' file, and the groups 'War and Peacemaking' and 'Drugs and Alcohol' for the combined file. Note that the input for every target variable was the same, a one-hot encoded file with 529 OCM codes for the 'p' file and 93 different groups for the 'combined' file.
The program in which the code was run was Google Colab, a free cloud-based Python environment. First the data was imported into the program, then PyCaret was called to do regression on the provided data, with the following line:

```
from pycaret.regression import *
NAME = setup(data = data, target = 'X', session_id = Y)
```

(All the code used can be found on GitHub and in Google Colab via the links in the Appendix on page 65, and via the link to Google Colab.)

A regression model is based on a regression formula. There are many different regression formulas, depending on the type of regression and the type of data, such as a linear or non-linear regression, and whether the variables are categorical or continuous for instance. Here, two examples of different formulas are given, one for linear multiple regression[14],

namely:

$$y_i = \alpha + \beta_1 x_{i_1} + \beta_2 x_{i_2} + \ldots + \beta_n x_{i_n} + \varepsilon$$

with:
y_i = dependant variable
x_i = explanatory variables
α = y-intercept (this is a constant term)
β_n = slope coefficients for each explanatory variable
ε = the error term of the model

and one general formula for a non-linear regression model, namely:

$$y = f(x, \beta) + \varepsilon$$

with:
f = a non-linear regression function
x = vector of P predictors
β = vector of k parameters
ε = the error term of the model

After the regression function was called, the `compare_models()` function was called to compare all the available models and to find the best model given the given target. The best model would then be created and tuned, and lastly the `evaluate_model()` function was called to evaluate the created model. From there the feature importance plot was saved, to show what variables are the most important in predicting the given target. The modeling step of the CRISP-DM cycle was mainly done by PyCaret. PyCaret would provide a table with all the regression models known in the program, then evaluate them all at the same time and give as a result a table with the best to worst models to evaluate the data given the input variable. The metrics used in this research to determine the best model are the R^2 value, MSE value, and RMSE value, standing for respectively Mean Squared Error and Root Mean Squared Error. The tables that PyCaret provides contain multiple metrics to evaluate the models, apart from the R^2, MSE, and RMSE values, does PyCaret also provide the MAE (Mean Absolute Error), RMSLE (Root Mean Squared Logarithmic Error), and MAPE (mean absolute percentage error) values. These error values will, however, not be mentioned further in this research. The R^2, MSE and RMSE values will be used in this research as they give a good indication of the fit of the model. The R^2 value explains the extent to which the variance

of one variable explains the variance of the second variable, in this case the variance between the target variable and the predictor variables. The closer the R^2 value is to one, the better. An R^2 value close to zero indicates that the model does not perform better than any other average model. The R^2 value can also go below zero, thus being negative. The range for R^2 is therefore -∞ to 1, but it is typically stated to be from 0 to 1. In general it is said that the lower the MSE and RMSE values are, the smaller the errors are of the model and the better the model will perform. The ranges for the MSE and RMSE values are both from 0 to ∞. The RMSE value was chosen as a used metric for this research as it provides an estimation of how well the model is able to predict the target value, in this case the chosen OCM code. The MSE value was chosen as the error metric to take into account as it is more sensitive to outliers than the MAE value, for example. The results (the table with models for each target variable) can be found both in section 4, Results and in the Appendix. The nineteen different regression models compared at the same time by PyCaret and their characteristics such as the type of model, the advantages and the disadvantages of every type of model can be found in the table on the next page.

3.3.5 Evaluation CRISP-DM cycle 1

PyCaret evaluated 19 different regression models at the same time for each given input variable (the target variable). To better understand the results in the next section, a quick overview of what can be seen in the tables and figures will be given here. In Figure 4.1 (see page 34) a table can be seen with nineteen different models, all compared at the same time by PyCaret using the `compare_models()` command. Now, we are the most interested in the R^2 column, the R-squared value. The R-squared value of a model shows how well the data fits the regression model, the better the fit, the higher the value. The optimal R^2 value is 1. The RMSE (Root Mean Squared Error) value is also of importance. The RMSE value measures the average difference between values predicted by the given model and the actual values. The closer this value is to 0, the better the model. Together with the R^2 value and the MSE (Mean Squared Error) value it forms the main performance indicators for a regression model. The Mean Squared Error value is the average squared error between the target variable and its projected value. Just like with the RMSE value, the lower the value is, the better (see [2]).
The other values left seen in the figures are the MAE (Mean Absolute Error), RMSLE (Root Mean Squared Logaritmic Error) and MAPE (Mean Ab-

Information on the nineteen different regression models			
Type	**Models**	**Advantages**	**Disadvantages**
Boosting	Light Gradient Boosting Machine, AdaBoost Regressor, Gradient Boosting Regressor, Extreme Gradient Boosting	Easy to interpret, resilient	Sensitive to outliers, difficult (almost impossible) to scale up
Ensemble	Light Gradient Boosting Machine, Dummy Regressor, AdaBoost Regressor, Gradient Boosting Regressor, Extreme Gradient Boosting, Random Forest Regressor, Extra Trees Regressor, Decision Tree Regressor	High predictive accuracy, useful when the data is both linear and non-linear, less noisy	Difficult to interpret, sensitive to wrong selection (can easily lead to lower predictive accuracy)
Greedy	Lasso Least Angle Regression, Orthogonal Matching Pursuit, Least Angle Regression	Can avoid overfitting, can do feature selection	No backtracking, may be difficult to interpret
Linear	Bayesian Ridge, Lasso Regression, Lasso Least Angle Regression, Elastic Net, Ridge Regression, Least Angle Regression, Linear Regression	Easy to interpret, can handle multiple independent variables at the same time, flexible and adaptable	May not capture non-linearity or complex patterns, sensitive to outliers
Non-linear	Light Gradient Boosting Machine, AdaBoost Regressor, Huber Regressor, Gradient Boosting Regressor, K Neighbors Regressor, Extreme Gradient Boosting, Random Forest Regressor, Extra Trees Regressor, Decision Tree Regressor, Passive Aggressive Regressor	Can capture complex relationships, high accuracy	Complex, difficult to interpret, prone to overfitting (or underfitting)

Table 3.1: *In this table the nineteen models are mentioned, categorized by their type and for every type the advantages and disadvantages are mentioned*

solute Percentage Error) values. The Mean Absolute Error refers to the magnitude of difference between the prediction of an observation and the true value of that observation. The Root Mean Squared Logarithmic Error adds 1 to both actual and predicted values before taking the natural logarithm. This avoids taking the natural log of possible 0 (zero) values. The Mean Absolute Percentage Error measures the average magnitude of an error produced by a model, or it measures how far off predictions are on average per individual variable.
Back to Figure 4.1, in this figure we see that Bayesian Ridge is the best model to predict OCM code 762 (Suicide), with an R^2 value of 0,0065, MSE value of 0,0158, and RMSE value of 0,1242. Looking at all the different target values (of which the tables can all be found in the Appendix on pages 70 to 78), Bayesian Ridge is not the only best model to fit the data and predict the given target value. Best found models for all the targets combined are: Bayesian Ridge, Light Gradient Boosting Machine, Extreme Gradient Boosting Machine, Gradient Boosting Regressor, Orthogonal Matching Pursuit and K Neighbors Regressor (although this model was not used to determine the feature importance plot of OCM code 728, Peacemaking). For all these models was the R^2 value the highest when comparing all the models at the same time. The models have in common that they are regression models, as the program was set to examine different regression models. The difference in not having the same best model for each target, even though the data setup was all the same for every target, can be explained by the relations between every variable. As some variables might be related to more variables than others, the outcome of the best model to use can change. Bayesian Ridge for instance, is a linear regression model, whereas Light Gradient Boosting Machine is an ensemble learning technique. Orthogonal Matching Pursuit is model that is handy to use when data is sparse, meaning that not many variables contribute to the target variable. This model was used for the 'combined' file. Extreme Gradient Boosting Machine (XGBoost) and Gradient Boosting Regressor are both ensemble learning techniques used for regression tasks. XGBoost is a newer model, known for its speed and high performance. Gradient Boosting Regressor, however, is a simpler model and therefore easier to interpret.
To tune the models and therefore improve the models, the command `tuned_x_y = tune_model(x_y, optimize = 'R2')` was used. This function tunes the hyperparameters of the chosen model. In Figures 4.2 and 4.3 (found in the Results section) the tuning of the Bayesian Ridge model for target 762 (Suicide) can be found, with the chosen value to be optimized being the R^2 value. All the other results can be found in the Appendix

on pages 79-82. R^2, the coefficient of determination, was optimised as R^2 is a measure that provides information about how well the model fits the data. For all the target variables it was chosen to optimize the R^2 value. When talking about regression, the R^2 value is a statistical measure that says how well the regression line approximates the actual data. In Figure 3.8 a sketch is made of a regression line, the blue dots represent data points and the red line is the regression line fitting the data points.

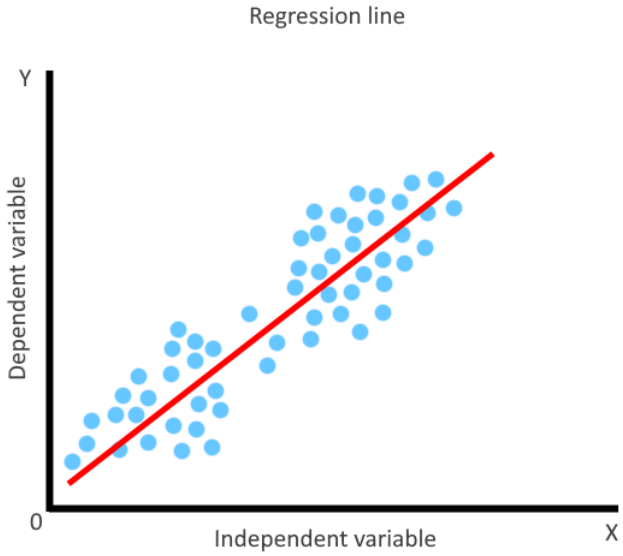

Figure 3.8: *Example of a regression line, graph made by M. Lelasseux*

3.3.6 Evaluation CRISP-DM cycle 2

For the second CRISP-DM cycle, which addresses research questions 2 and 3, RapidMiner was used to study a hierarchy of OCM code 762 (Suicide). Python and Sklearn were used to determine clusters within the OCM codes and the cases. All the Figures created can be found in the next section, section 4 Results.
RapidMiner was run four times after preparing the data, once with pruning and prepruning on, once with only pruning, once with only prepruning and once with no pruning enabled at all. Pruning is the act of cutting branches of the decision tree that are no longer needed, so those that are non-critical to determine the target value.

The general parameters settings for the decision trees found are:
criterion: gain ratio
maximal depth: 10
confidence: 0,1 (when pruning was applied)

And for the prepruning process the settings are:
minimal gain: 0,01
minimal leaf size: 2
minimal size for split: 4
number of prepruning alternatives: 5.

These settings were chosen after running the program multiple times and comparing the different results. The obtained results are now readable and the trees are not bigger than half a page.
Changing the hyperparameters would change the shape of the decision tree, mainly the size. First `gain ratio` was chosen as criterion, gain ratio is a measure that takes both the information gain and the number of outcomes of a feature into account when determining the best feature (in this case an OCM code) to split on. With the `maximal depth` parameter the depth of the decision tree can be chosen, to keep the tree readable `10` was chosen. A smaller tree would result in too much branches being cut off, resulting in no decsion tree at all. Increasing the value from 10 would make the decsion tree bigger and not fit for the analysis of this thesis. However, domain experts could study a bigger hierarchy tree and incorporate more variables in the results. The `minimal gain` parameter was set to `0,01` as the data contained 529 OCM codes and thus a lot of different variables. A higher value of minimal gain results in fewer splits.
For the `leaf size` parameter `2` was chosen, as the cases in the data file were not linked to each other, it was difficult to find examples with which leafs could be created. The minimal leaf size of 2 means that at least two examples need to be found in the data to create a leaf[22]. A value of `4` was chosen for the `minimal size for split` parameter to prevent the tree from becoming extremely big. Last but not least, for the
`number of prepruning alternatives` parameter `5` was chosen after testing the program with different values. Later on the difference between the hierarchy trees will be discussed.

3.3.7 Deployment

The last phase of the CRISP-DM method is deployment, in this case the result, discussion and conclusion sections and this thesis in its entirety will be the deployment of all the research done.

3.4 CRISP-DM cycle 1

This section provides a quick summary of the first CRISP-DM cycle, addressing the main research question and the first domain question.
As said previously, for the first cycle the OCM codes of the OAL data file were analysed. After establishing the business questions and understanding the data by hand coding the provided suicide dataset, a test file was created and run. From there it was discovered that some texts were duplicates in the file, therefore it was chosen to create a new file with only the paragraphs, called the 'p' file. Now the 'p' file was run. After the first test run it was discovered that the feature importance plot would return that the 'OCM' column was the most important variable, leaving all the other data unused (see Figure 3.5). Now, the one-hot encoding technique was applied on all the 529 OCM codes, to create a data file with only the OCM codes. This data file would then be used when doing the data analyses. The one-hot encoding technique created a data file with only the OCM codes per column, in this case called `p_encoding`. From there PyCaret was run to evaluate all the regression models known in the program, the targets of the `homicide_contain_p_encoding.csv` file (which is the full name of the used data file) were OCM codes: 762 (Suicide), 754 (Sorcery), 728 (Peacemaking), 685 (Property Offenses), 684 (Sex and Marital Offenses), 683 (Offenses Against the Person), 681 (Sanctions), 672 (Liability), 628 (Inter-community Relations), 627 (Informal Ingroup Justice), and 578 (Ingroup Antagonisms). After finding the best possible model for the provided target, the model was created and tuned, optimizing the R^2 value. After the models were tuned they were evaluated by PyCaret, providing a feature importance plot for every target. The results, including all the feature importance plots, can be found in section 4, Results. One exception was target 728 (Peacemaking), this target had `K Neighbors Regressor` as best model, however, this model could not provide a feature importance plot. The `K Neighbors Regressor` algorithm is based on the `K-Nearest Neighbors` algorithm, a supervised learning method. The `K Neighbors Regressor` does not provide a feature importance plot because 'Feature importance' is not defined for this

algorithm. Because of this, it was chosen to do further analyses with the Bayesian Ridge model for this target, as Bayesian Ridge came in as the second best model.
Lastly, another data file was created for this CRISP-DM cycle, a data file called 'combined'. This file contained self-chosen categories of groups of OCM codes, mentioned before. The code to create the new groups can be found on GitHub. It was chosen to do this to shorten the runtime and to compare the results of the 'p' file with the results of the 'combined' file. It was hypothesized that a less dense data file would lead to a shorter runtime. Grouping the OCM codes was also done to investigate whether groups of OCM codes would still give the same results as leaving the OCM codes separate. In other words, when a model would indicate that a target variable can be predicted by, say, OCM codes indicating different kinds of offenses, or different group dynamics and machine tools, would these results then still be found when we group all the OCM codes 'offenses' together, and all the group dynamics codes together, etc.? To research this the 'combined' file was created. It it hypothesized that the R^2 values will increase in comparison to the R^2 values of the targets of the 'p' file, because of less noise and conflicting distributions.
The one-hot encoding method was also used on the combined file, creating the definite file `combined_homicide_data_encoding.csv`. The targets of the analyses done on this file were the grouped OCM codes Drugs & Alcohol and War & Peacemaking. Creating and tuning the models was done the same as with the previous targets for the
`homicide_contain_p_encoding.csv` file.

3.5 CRISP-DM cycle 2

In this section the second CRISP-DM cycle will be briefly explained, this cycle addresses the third and fourth domain question of this research. After the first cycle was completed, the choice was made to do another cycle, but this time only focusing on the target Suicide (OCM code 762). This target was chosen as it is the most important object in self-directed harm.
A more in-depth analysis was done by using other programs and techniques to research a possible hierarchy for target 762 (Suicide) and to find clusters within the complete dataset. The program RapidMiner was used to research and create the hierarchy decision tree. RapidMiner was run with an education licence. To find clusters in the
`homicide_contains_p_encoding.csv` file the library 'sklearn' was used, this was done in Google Colab just as done before when creating models

and the feature importance plots. After preparing the data for RapidMiner to work correctly with the program (thus changing every 1 to 'true' and every 0 to 'false') the program was run multiple times with different hyperparameters settings, to create and analyse four different hierarchy trees with either pruning and/or prepruning applied. The hierarchies (decision trees) and the parameters can be found in section 4, Results.
To determine clusters within the OAL dataset, the code was run twice. The first time clusters were determined within the 7046 cases of the OAL data file, both for the 'p' and the 'combined' file. The second time the data was transposed from a row-based to a column-based dataset. This was done to determine clusters within the 529 OCM codes in the 'p' file and within the 93 groups of OCM codes within the 'combined' file. The results can be found in section 4, Results.

Chapter 4

Results

PyCaret was run multiple times during this research, each time with a different chosen target. The four files used were: `homicide_contain_p.csv`, also previously called the 'p' file, this file contains all the different cases which were coded with a 'p' in the original OAL data file. Another file, the `homicide_contain_p_encoding.csv` file, previously called 'p_encoding', is derived from the `homicide_contain_p.csv` file and contains only the 529 OCM codes mentioned (every code is a column) and for every case every column is coded with either a 1 (true) or a 0 (false).
The `combined_homicide_data.csv` file, also called 'combined', is a data file which contains the original information from the OAL data file (with only the paragraphs as cases). After the original information, 93 new columns were created with the new groups of OCM codes and whether one or more OCM codes of that group were being mentioned in every case yes (1) or no (0). Lastly, the `combined_homicide_data_encoding.csv`, also previously called 'combined_encoding' file, contains only the 93 created groups of OCM codes and whether one or more OCM codes per group are mentioned in each case. Coded with either a 1 (true) or a 0 (false).

4.1 Frequency of OCM codes

First a Python script was run to determine the top 10 OCM codes in the OAL data file. The script can be found in the Appendix on page 68.
The top 10 OCM codes (after OCM code 682 (Offenses Against Life), as every text in the OAL file contains OCM code 682) can be found in the table below. A table with all the OCM codes and their frequencies found in the feature importance plots can be found on GitHub.

Frequencies of the 10 most common OCM codes		
OCM code	Definition	Frequency
682	Offenses Against Life	7046
627	Informal Ingroup Justice	886
578	Ingroup Antagonisms	738
628	Inter-community Relations	588
672	Liability	513
754	Sorcery	462
683	Offenses Against the Person	421
695	Trial Procedure	349
613	Lineages	348
674	Crime	319
648	International Relations	317

Table 4.1: *In this table the frequencies of the occurrences of the top 10 (top 11 if OCM 682 is also included) OCM codes can be seen, including the definition of every OCM code*

4.2 Models

PyCaret runs multiple models at the same time, to determine which model works best for the given input. Meaning, the model that gives the most accurate results for the given target variable. Thirteen different targets were chosen beforehand, eleven were for the `homicide_contain_p.csv` file and two were for the `combined_homicide_data.csv` file. The targets being: 762 (Suicide), 578 (Ingroup Antagonisms), 627 (Informal Ingroup Justice), 628 (Inter-community Relations), 672 (Liability), 681 (Sanctions), 683 (Offenses Against the Person), 684 (Sex and Marital Offenses), 685 (Property Offenses), 728 (Peacemaking), and 754 (Sorcery) for the 'p' file, and the groups 'War and Peacemaking' and 'Drugs and Alcohol' for the combined file.

In Figure 4.1 the outcome of the best models for target variable 762 (Suicide) can be found. As can be seen in the table, Bayesian Ridge is the best model to do analyses with and to predict values with in the future. The models Least Angle Regression and Linear Regression are both no fit at all for this target variable. The table shows extremely high values for the MAE, MSE, RMSE, and MAPE values for the Linear Regression model, and an extremely low R^2 value of $-53,14e^20$. For the Least Angle Regression model extremely high values for the MAE, MSE, and RMSE values are seen, even so as an extremely low R^2 value of $-26,12e^19$. The reason to why these values are so high and low is because a linear model is not the correct model to interpret this data. There could be non-linear relationships between the OCM codes or interactions among variables that are not captured by the Least Angle Regression model and Linear Regression model.

The other tables with the nineteen regression models for the remaining targets are not shown in this section, but can all be found in the Appendix on pages 70-78.

4.3 Tuned models

To improve the models, and therefore to get more accurate results, the models were all tuned. Below the 'before and after tuning' tables of target 762 (Suicide) can be found. After running multiple analyses it was found that regression models Gradient Boosting Regressor and Light Gradient

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
br	Bayesian Ridge	0.0321	0.0158	0.1242	0.0065	0.0863	0.9756	0.5670
lightgbm	Light Gradient Boosting Machine	0.0332	0.0159	0.1244	0.0048	0.0872	0.9564	0.5120
lasso	Lasso Regression	0.0319	0.0160	0.1248	-0.0024	0.0867	0.9838	0.0950
dummy	Dummy Regressor	0.0319	0.0160	0.1248	-0.0024	0.0867	0.9838	0.0490
llar	Lasso Least Angle Regression	0.0319	0.0160	0.1248	-0.0024	0.0867	0.9838	0.0910
en	Elastic Net	0.0319	0.0160	0.1248	-0.0024	0.0867	0.9838	0.0910
ada	AdaBoost Regressor	0.0292	0.0160	0.1250	-0.0091	0.0868	0.9726	0.2960
huber	Huber Regressor	0.0163	0.0162	0.1257	-0.0160	0.0871	0.9997	0.6380
gbr	Gradient Boosting Regressor	0.0294	0.0163	0.1259	-0.0252	0.0879	0.9616	1.9930
omp	Orthogonal Matching Pursuit	0.0357	0.0164	0.1265	-0.0334	0.0892	0.9552	0.1750
ridge	Ridge Regression	0.0412	0.0167	0.1276	-0.0539	0.0908	0.9477	0.1360
knn	K Neighbors Regressor	0.0297	0.0185	0.1342	-0.1639	0.0978	0.9569	0.1140
xgboost	Extreme Gradient Boosting	0.0312	0.0188	0.1351	-0.2011	0.0950	0.9482	0.5700
rf	Random Forest Regressor	0.0373	0.0203	0.1405	-0.2941	0.1026	0.9310	12.2560
et	Extra Trees Regressor	0.0356	0.0244	0.1533	-0.5573	0.1097	0.9389	17.4510
dt	Decision Tree Regressor	0.0358	0.0256	0.1576	-0.6284	0.1121	0.9323	0.2560
par	Passive Aggressive Regressor	0.1402	0.0440	0.2077	-1.8598	0.1642	0.8809	0.1510
lar	Least Angle Regression	96824239.6594	57886719047055329280000	1074923503.7451	-2612940019911945748480000	0.5230	0.9372	0.3080
lr	Linear Regression	5085914547.2464	11299929378592554745856.0000	66285934316.5261	-5314149723959194256670720000	2.2555	8538707523.8154	0.2190

Figure 4.1: *Table with the evaluation of the regression models, target variable: 762 (Suicide), input variables: all the 529 OCM codes, file used: homicide_contain_p_encoding.csv*

Boosting Machine could not be improved with the `tune_model()` function of PyCaret. The other regression models did show an improvement after using the tuning function. All the 'before and after' tuning figures of all the target variables can be found in the Appendix on pages 79 to 82.

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.0297	0.0137	0.1170	0.0196	0.0811	0.9720
1	0.0279	0.0119	0.1090	0.0102	0.0759	0.9739
2	0.0364	0.0196	0.1401	0.0121	0.0971	0.9739
3	0.0322	0.0158	0.1257	0.0099	0.0873	0.9759
4	0.0368	0.0199	0.1411	-0.0021	0.0981	0.9791
5	0.0292	0.0119	0.1092	0.0081	0.0762	0.9732
6	0.0248	0.0081	0.0900	-0.0072	0.0633	0.9761
7	0.0353	0.0198	0.1406	0.0055	0.0975	0.9783
8	0.0277	0.0119	0.1091	0.0108	0.0759	0.9740
9	0.0411	0.0257	0.1604	-0.0021	0.1110	0.9793
Mean	0.0321	0.0158	0.1242	0.0065	0.0863	0.9756
Std	0.0049	0.0051	0.0201	0.0077	0.0137	0.0025

Figure 4.2: *The Bayesian Ridge model with target 762 (Suicide) before tuning*

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.0302	0.0134	0.1157	0.0421	0.0801	0.9556
1	0.0285	0.0119	0.1090	0.0095	0.0761	0.9686
2	0.0385	0.0195	0.1397	0.0176	0.0971	0.9636
3	0.0326	0.0157	0.1255	0.0136	0.0873	0.9700
4	0.0378	0.0200	0.1415	-0.0075	0.0986	0.9772
5	0.0300	0.0119	0.1089	0.0127	0.0762	0.9648
6	0.0254	0.0081	0.0901	-0.0080	0.0635	0.9702
7	0.0367	0.0197	0.1402	0.0103	0.0974	0.9710
8	0.0285	0.0119	0.1089	0.0130	0.0761	0.9672
9	0.0423	0.0258	0.1608	-0.0066	0.1113	0.9773
Mean	0.0330	0.0158	0.1240	0.0097	0.0864	0.9686
Std	0.0052	0.0051	0.0202	0.0142	0.0137	0.0061

Figure 4.3: *The Bayesian Ridge model with target 762 (Suicide) after tuning, parameter tuned: 'R^2'*

4.4 Feature importance plots

A feature importance plot is used to determine the effect that a specific variable has on predicting the target variable. It calculates a score for all the input variables in a given model. The higher the score, the larger the effect this variable has on the target variable. The graph thus shows the importance of each variable, hence the name 'feature importance plot'. The scales in the plots differ from a scale from 0-1 to a scale from 0-100. The different formats of the scales were chosen by the program itself and do not indicate a difference in performance or outcome. All the feature importance plots made with the Light Gradient Boosting Machine model were shown on a scale from 0-100, the other models used a 0-1 scale. On the y-axis the ten most predictive variables are shown, on the x-axis the variable importance is shown. The variable importance indicates the relative importance of each variable in a dataset or datafile when building a (predictive) model, the higher the value, the more that variable helps predict the target variable. In this section the top 10 of the feature importance plots of the targets 762 (Suicide), 578 (Ingroup Antagonisms), 627 (Informal Ingroup Justice), 628 (Inter-community Relations), 683 (Offences Against the Person), 684 (Sex and Marital Offenses), 685 (Property Offenses), 754 (Sorcery), and War & Peacemaking will be shown. In the Appendix on pages 83 to 85 the feature importance plots of targets 672 (Liability), 681 (Sanctions), 728 (Peacemaking), and Drugs & Alcohol are shown. The results of the targets Liability and Sanctions will not be further discussed, as the results of these targets were not of any relevance. They are still shown in the Appendix for everyone interested in the results.
First the feature importance plot of target 762, Suicide, is shown. This is also the target variable that will be analysed a second time when doing the second CRISP-DM cycle. From there all the other plots are shown in order of ascending OCM code.
In section 5, Discussion, the findings of this research will be discussed in detail.

Figure 4.4 shows the feature importance plot of target 762, Suicide. Mortality is the most important predictor of suicide, which is instinctive as suicide death is directly linked to mortality. Furthermore, we see that this figure is in line with previously mentioned research[9] where it was stated that sexual conflict is a predictor of suicidal behaviour. In this figure we see several links to sexual conflicts such as Sexual Stimulation, Sexuality, Termination of Marriage, and Conception (i.e. not being able to conceive a child). Other research[26] also stated that suicidal behaviour is mainly

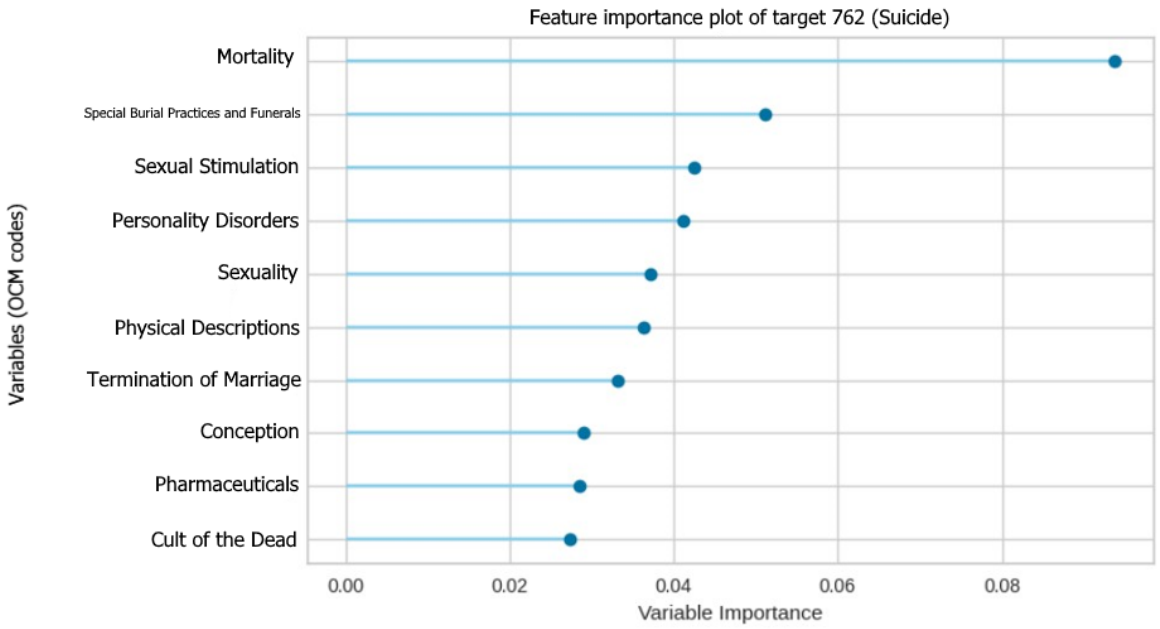

Figure 4.4: *Feature importance plot created with the Bayesian Ridge model, target: 762 (Suicide), file: homicide_contain_p_encoding.csv*

linked to being female and young (which can then be linked again to for example Termination of Marriage and Conception).

Figure 4.5 on the next page shows the main predictors of OCM code 578 (Ingroup Antagonisms). Predictors mentioned such as Informal Ingroup Justice, Community Structure, Offenses Against the Person, Warfare, and Social Relations and Groups are in line with research (see [6, 21, 50]) stating that ingroup tendencies can lead to conflict. Also Alcoholic Beverages (see [7]) are an already known predictor of conflict.

In Figure 4.6, which is shown on the next page, the most important variables to predict OCM code 627 (Informal Ingroup Justice) can be seen. Here it can be seen that, as mentioned before[6], different factors of ingroups are a predictor of Informal Ingroup Justice, being for instance Intercommunity Relations, Kin Relationships, Community Heads, Sibs, and Ingroup Antagonisms.

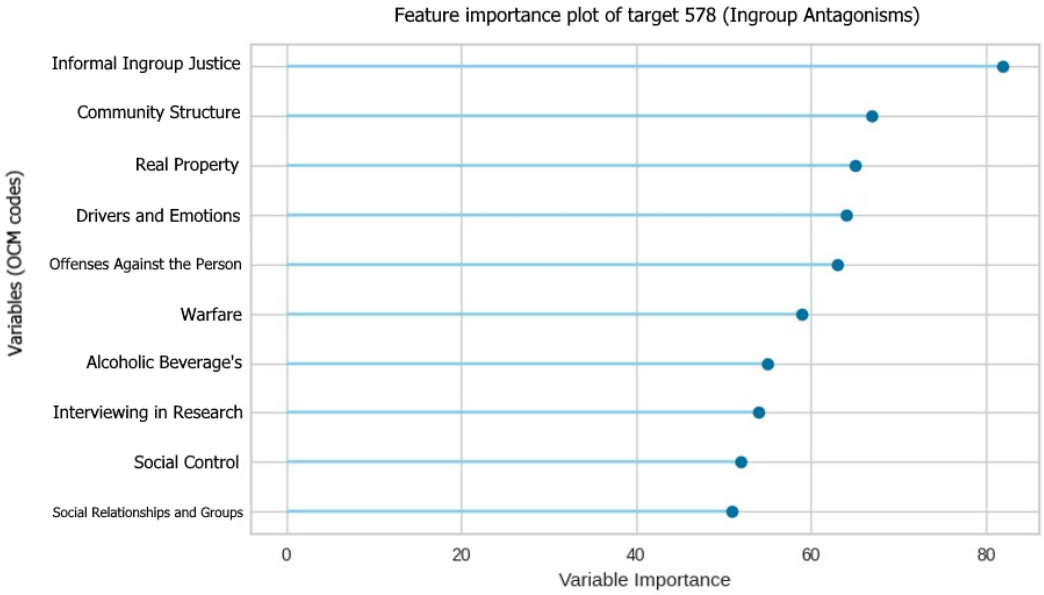

Figure 4.5: *Feature importance plot created with the Light Gradient Boosting Machine model, target: 578 (Ingroup Antagonisms), file: homicide_contain_p_encoding.csv*

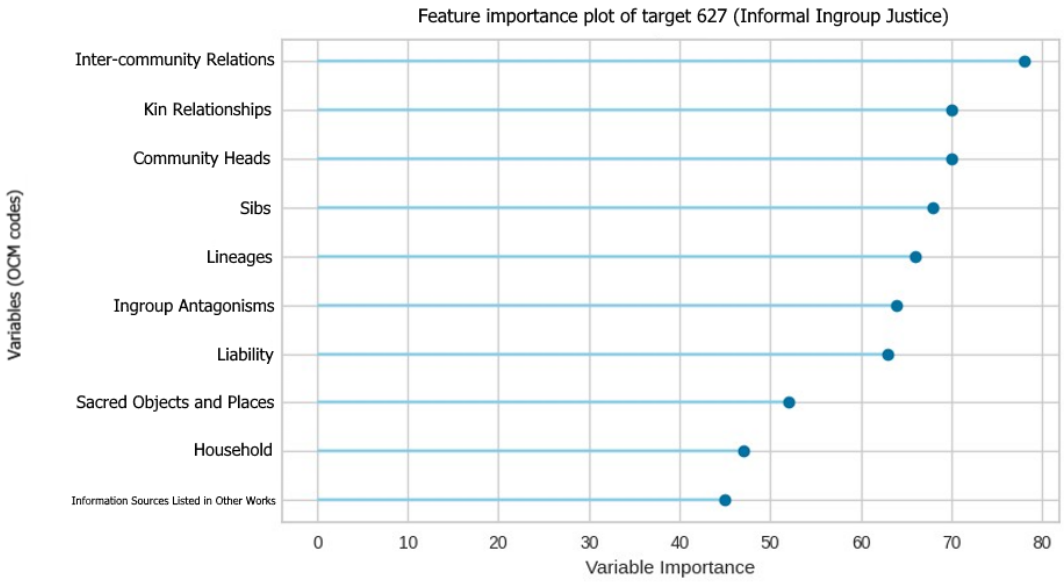

Figure 4.6: *Feature importance plot created with the Light Gradient Boosting Machine model, target: 627 (Informal Ingroup Justice), file: homicide_contain_p_encoding.csv*

Below in Figure 4.7 the feature importance plot of OCM code 628 (Inter-community Relations) is shown. This figure can again be linked to warfare (Weapons, Instigation of War) for instance, a factor which is known as a predictor of conflict between groups[49, 50]. Other predictors that are also in line with previously mentioned research[46] are Arranging a Marriage, and Community Structure, where a marriage arrangement can be seen as a conflict between two parties.

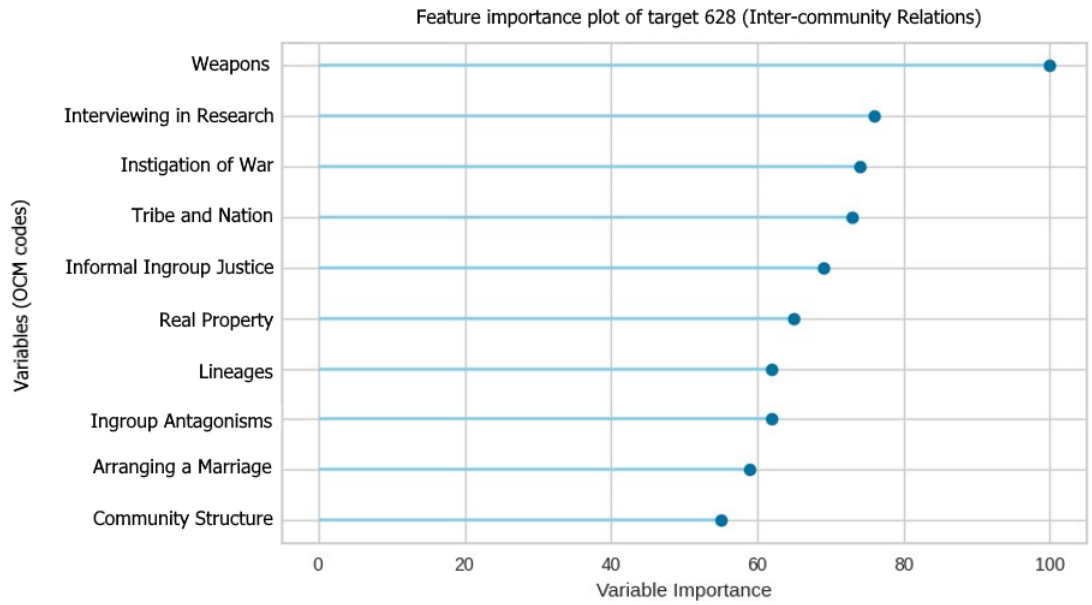

Figure 4.7: *Feature importance plot created with the Light Gradient Boosting Machine model, target: 628 (Inter-community Relations), file: homicide_contain_p_encoding.csv*

Figures 4.8 and 4.9 show the main predictors of different offenses such as Offenses Against the Person and Sex and Marital Offenses. Sorcery, which is the main predictor of an offense against a person, is a well known factor of an offense, still observable all over the world. Now Sex and Marital offenses and Premarital Sex Relations can be linked to conflicts, which can then be linked to offenses. An example of a sexual conflict is rape, which is an extreme conflict[9] between (at least) two people, where one party is being powerlessness[45]. Most of the predictors of Sex and Marital Offenses are intuitive, namely Extramarital Sex Relations, Sexual Intercourse and Secondary Marriages.

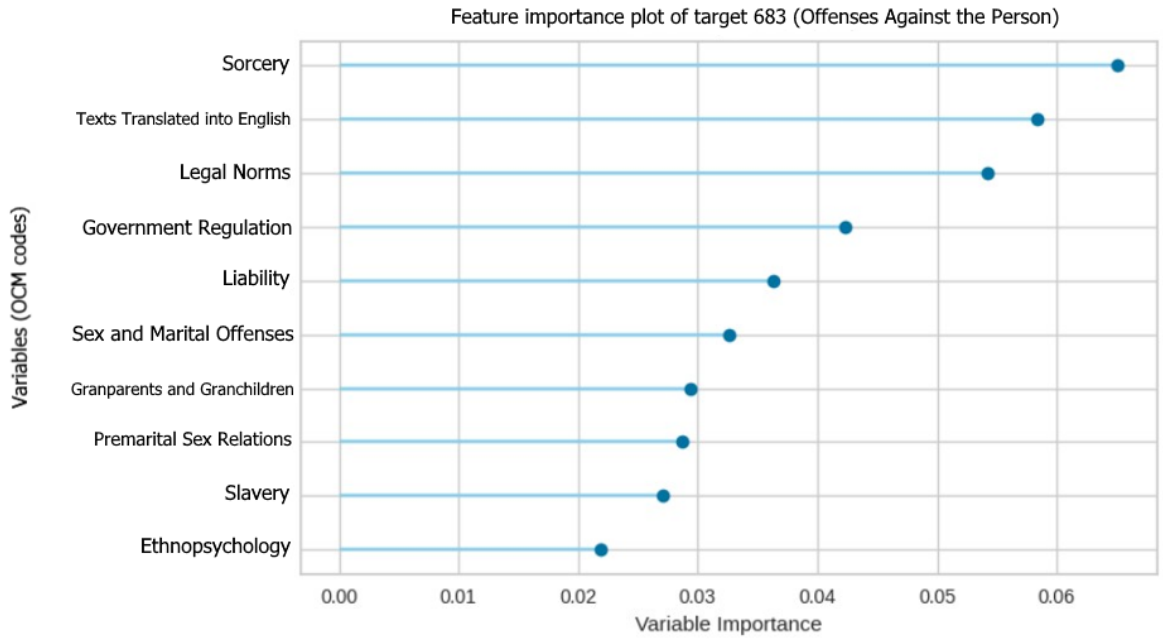

Figure 4.8: *Feature importance plot created with the Gradient Boosting Regressor model, target: 683 (Offenses Against the Person), file: homicide_contain_p_encoding.csv*

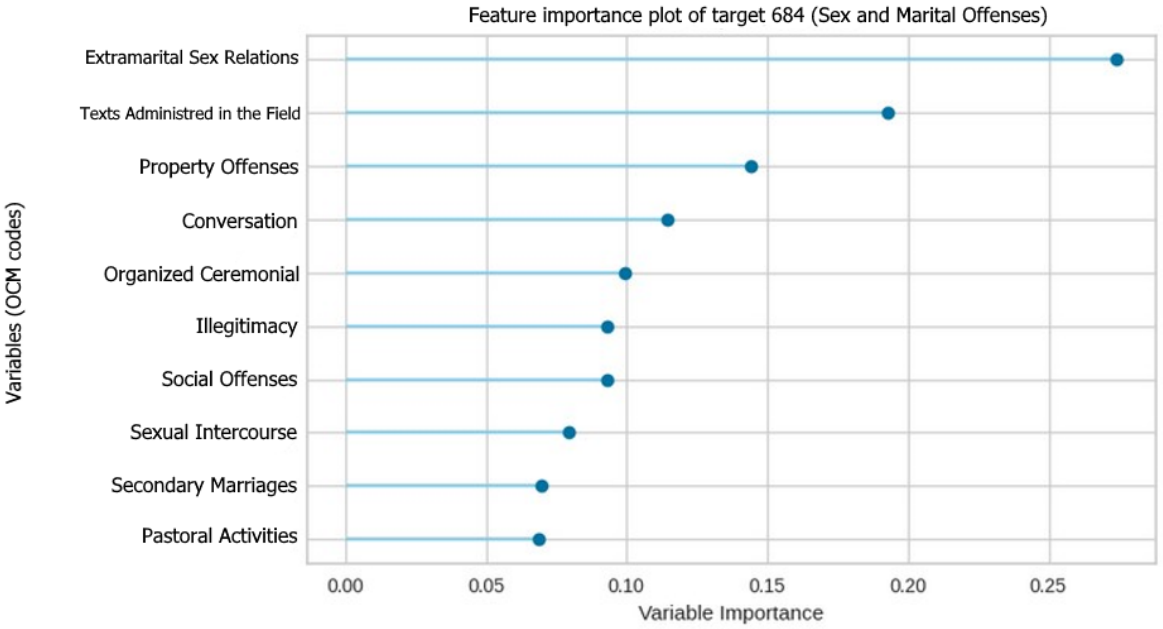

Figure 4.9: *Feature importance plot created with the Bayesian Ridge model, target: 684 (Sex and Marital Affairs), file: homicide_contain_p_encoding.csv*

Now, Figure 4.10 shows the feature importance plot, and therefore the main predictors, of OCM code 685 (Property Offenses). Research[33] has stated that externalizing behaviour is found more in men, which can then be linked to Social Offenses and Sex and Marital Offenses for example, both seen as factors in Figure 4.10. Other variables can also be linked to more externalizing behaviour, such as Wrongs, and Offenses Against the Person.

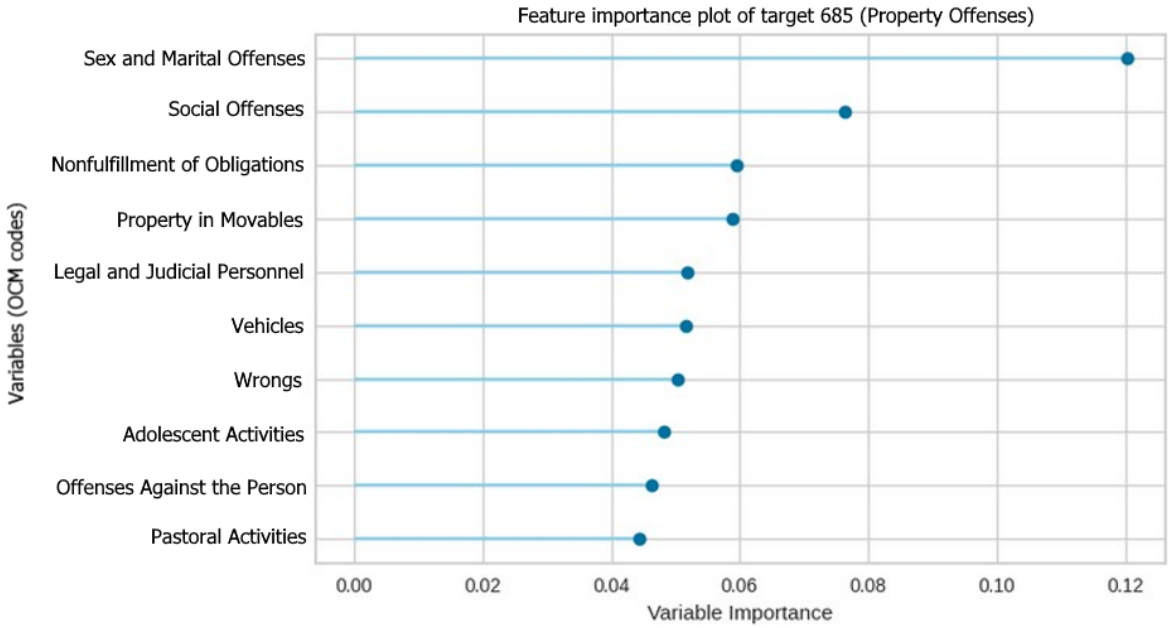

Figure 4.10: *Feature importance plot created with the Bayesian Ridge model, target: 685 (Property Offenses), file: homicide_contain_p_encoding.csv*

In Figure 4.11 on the next page we see mainly variables that have to do with intergroup violence being the main predictors of Sorcery. In this case these are Ingroup Antagonisms, Community Structure, and Informal Ingroup Justice. Previous research[6, 21, 50] mentioned that friction within a group or between groups can lead to interpersonal or intergroup violence. Here, the bargaining strategy used is Sorcery. Shamans and Psychotherapists and Magic are intuitively connected to sorcery.

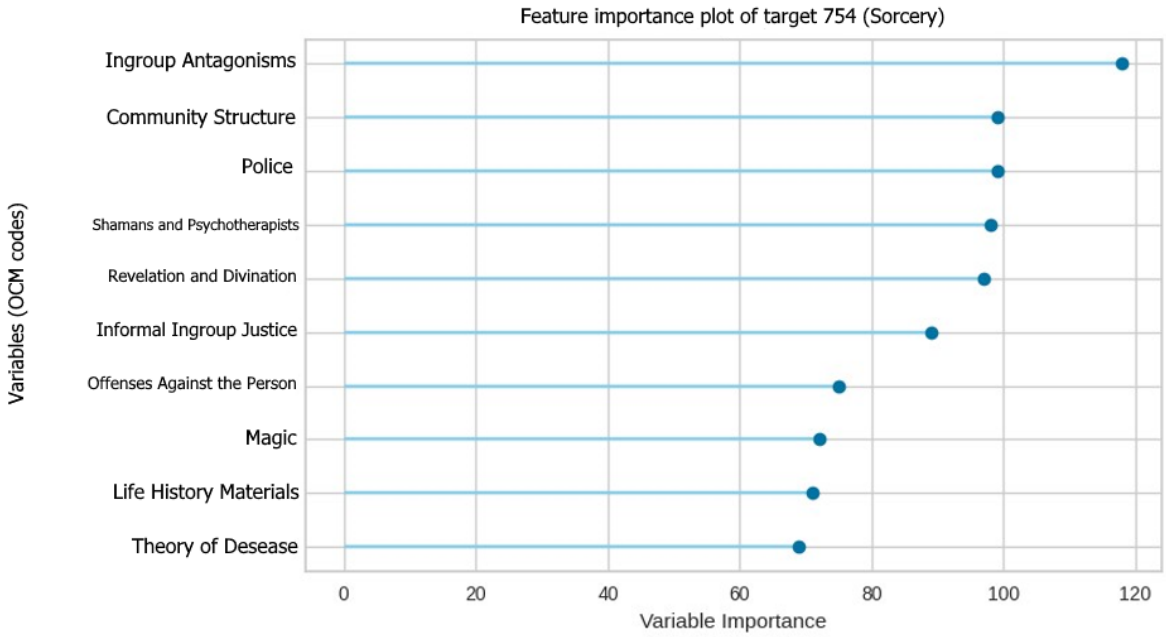

Figure 4.11: *Feature importance plot created with the Light Gradient Boosting Machine model, target: 754 (Sorcery), file: homicide_contain_p_encoding.csv*

Last but not least does Figure 4.12 on the next page show the main predictors of War and Peacemaking. The groups of OCM codes that are important here are (as mentioned in various research[49, 50]): Agriculture, Food consumption, and Food processing, all related to conflicts over land and scarce sources. Another important predictor is Military/armed forces for example, which can be linked to externalizing behaviour and lashing out (see [38] and [33]).

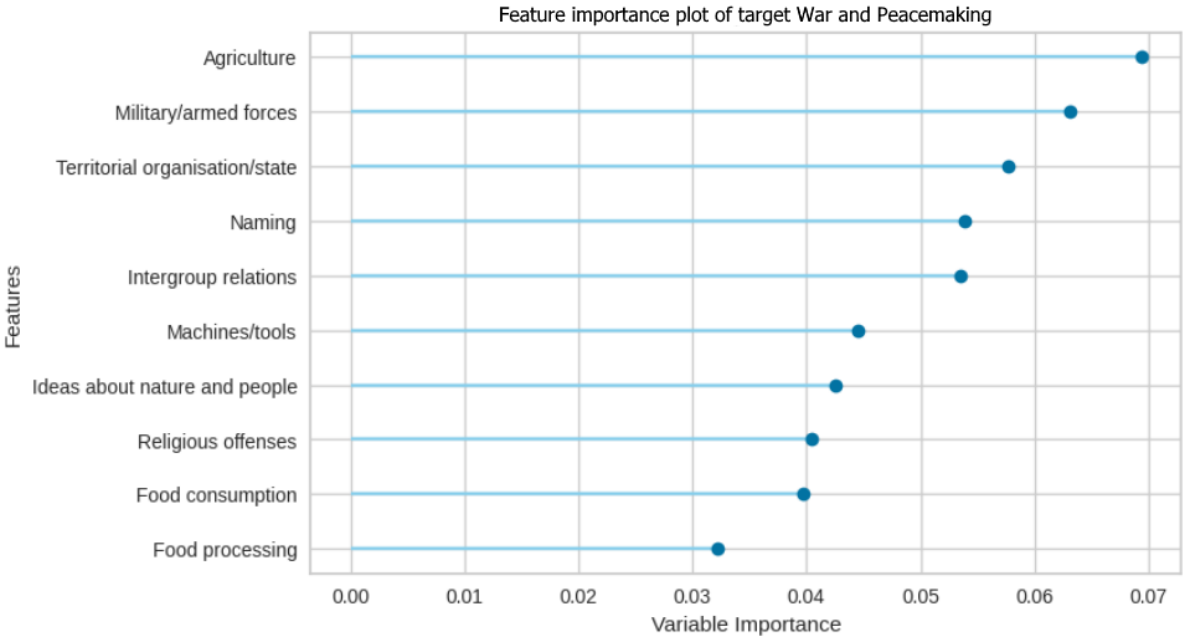

Figure 4.12: *Feature importance plot created with the Bayesian Ridge model, target: War and Peacemaking, file: combined_homicide_data_encoding.csv*

4.5 Hierarchy and clusters

Within data one may find a, or multiple, hierarchies and clusters. A hierarchy is defined as a data model which uses a decision tree for instance as its basic structure. It then organizes data into nested levels of abstraction, such as classes and sub-classes or instances. A cluster is a group of similar objects within a dataset, grouped together. In this section the hierarchy of OCM code 762 (Suicide) can be found, together with the cluster analysis done.

Below the hierarchy of OCM code 762 (Suicide) is shown. The hierarchy was made with the program RapidMiner, using the decision tree making algorithm. Figures 4.13, 4.14, 4.15, and 4.16, show four different decision trees, with in every tree either prepruning, pruning, or both enabled or disabled.

Figures 4.13, 4.14, 4.15 and 4.16 can all be read in the same way. For Figure 4.13 this means that when OCM codes 886 (Senescence), 580 (Marriage), 272 (Nonalcoholic Beverages), 481 (Locomotion), and 165 (Mortality) are all false and OCM codes 831 (Sexuality) and 683 (Offenses Against the Person) are both true, then OCM code 762 (Suicide) will be true.

Version of February 29, 2024– Created February 29, 2024 - 19:58

Settings RapidMiner Figure 4.13	
criterion	gain_ratio
maximal depth	10
apply pruning	yes
confidence	0,1
apply prepruning	yes
minimal gain	0,01
minimal leaf size	2
minimal size for split	4
number or prepruning alternatives	5

Table 4.2: *Settings of the program RapidMinder, used to create the hierarchy of target 762 (Suicide), seen in Figure 4.13. Here, both prepruning and pruning are applied*

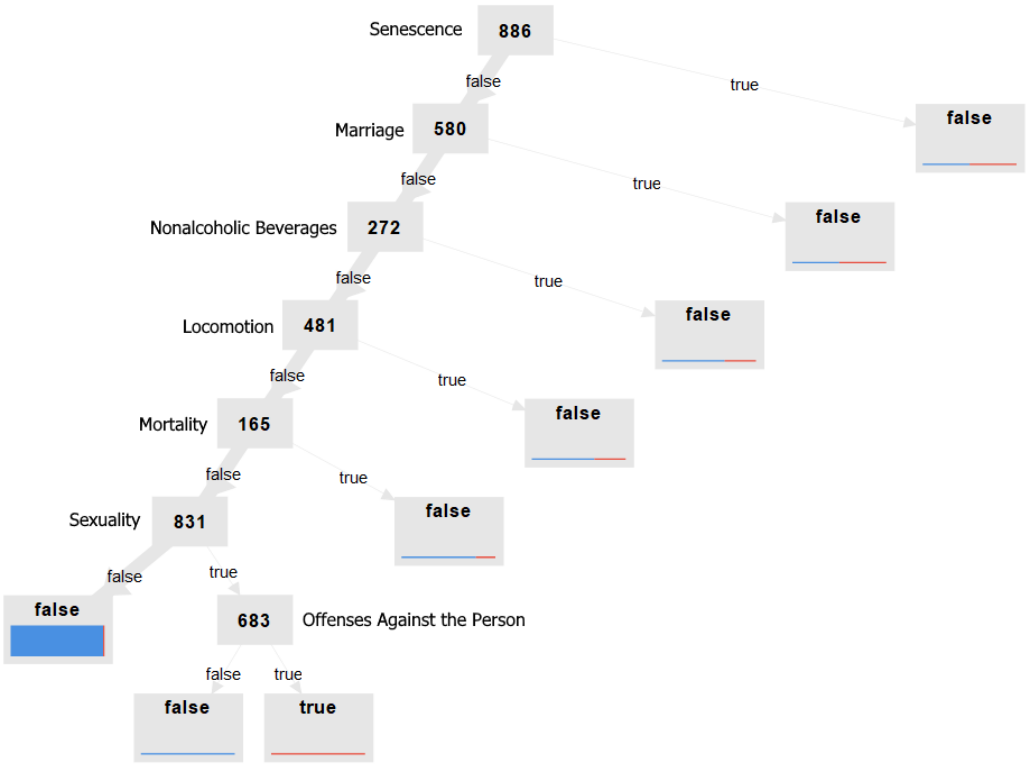

Figure 4.13: *Hierarchy (decision tree) of target 762 (Suicide), pruning and prepruning applied*

Settings RapidMiner Figure 4.14	
criterion	gain_ratio
maximal depth	10
apply pruning	yes
confidence	0,1
apply prepruning	no

Table 4.3: *Settings of the program RapidMinder, used to create the hierarchy of target 762 (Suicide), seen in Figure 4.14. Here, only pruning is applied*

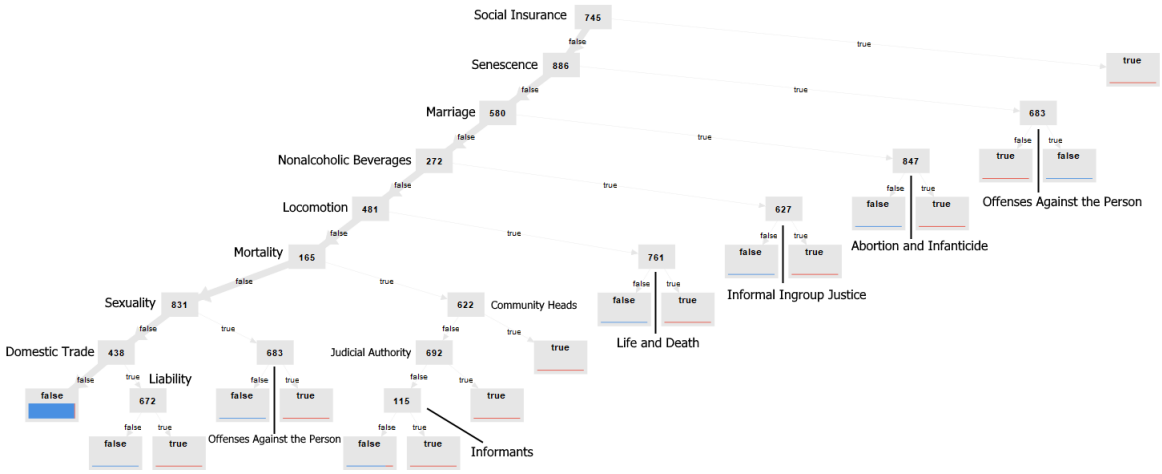

Figure 4.14: *Hierarchy (decision tree) of target 762 (Suicide), only pruning applied*

Settings RapidMiner Figure 4.15	
criterion	gain_ratio
maximal depth	10
apply pruning	no
apply prepruning	yes
minimal gain	0,01
minimal leaf size	2
minimal size for split	4
number or prepruning alternatives	5

Table 4.4: *Settings of the program RapidMinder, used to create the hierarchy of target 762 (Suicide), seen in Figure 4.15. Here, only prepruning is applied*

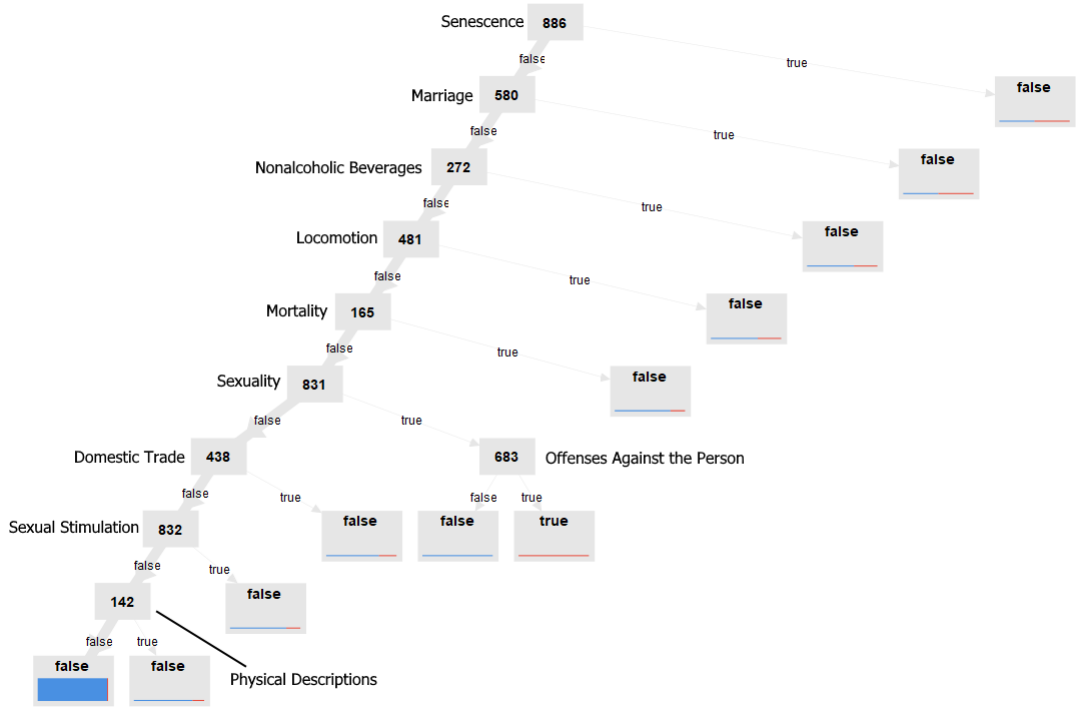

Figure 4.15: *Hierarchy (decision tree) of target 762 (Suicide), only prepruning applied*

Settings RapidMiner Figure 4.16	
criterion	gain_ratio
maximal depth	10
apply pruning	no
apply prepruning	no

Table 4.5: *Settings of the program RapidMinder, used to create the hierarchy of target 762 (Suicide), seen in Figure 4.16. Here, no pruning at all was applied*

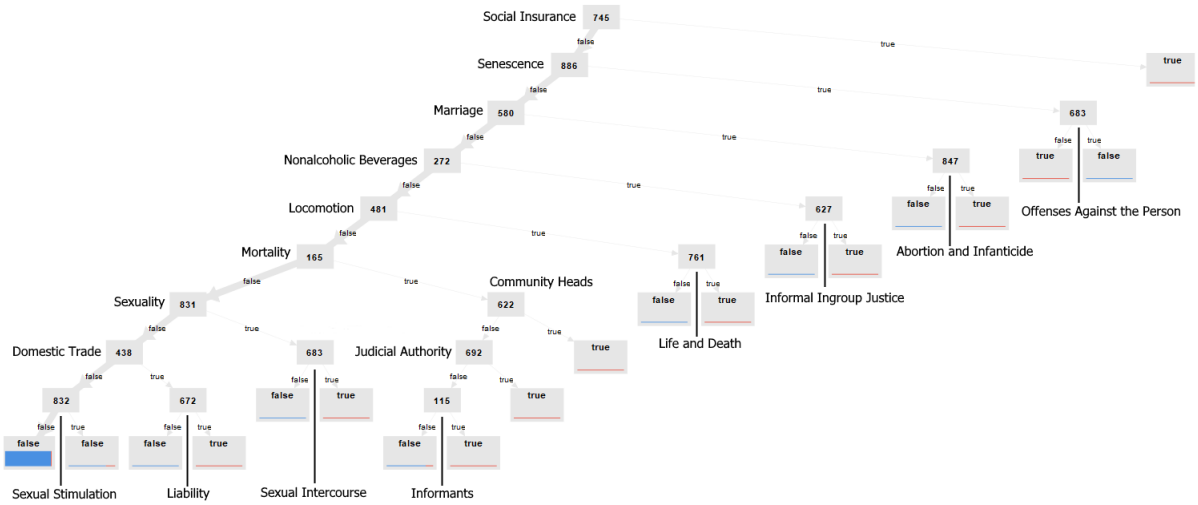

Figure 4.16: *Hierarchy (decision tree) of target 762 (Suicide), no pruning at all applied*

Looking at Figure 4.13, it is contradicting to the findings in Figure 4.4. Figure 4.13 shows that when OCM codes 886 (Senescence), 580 (Marriage), 272 (Nonalcoholic Beverages), 481 (Locomotion), and 165 (Mortality) are false and OCM codes 831 (Sexuality) and 683 (Offenses Against the Person) are true then OCM code 762, Suicide, is true. However, in Figure 4.4 it can be seen that OCM code 165, Mortality, is the most important variable to determine whether Suicide would be coded with a 1 or a 0 in the dataset, meaning that Suicide would be true or false for the given case. OCM code 165 (Mortality) implies that when Mortality is coded with a 1, the chances are higher for Suicide to be coded with a 1 (thus being true). The feature importance plot does not give a 100% certainty, meaning that even though Mortality is coded with a 1, then the same paragraph is not necessarily coded with a 1 for Suicide. The decision tree in Figure 4.13 does give a 100% certainty that when the conditions for all the before mentioned OCM codes are correct, that Suicide would be coded with a 1.
A possible explanation to this contradiction in the findings could be that the path in the decision tree leading to the leaf with 'true' could be all determined from only a few cases (rows) in the given data. To research this a small Python code was written (see the Appendix, page 69), this code was run on the `homicide_contain_p_encoding.csv` file and filtered out all the cases where OCM codes 886 (Senescence), 580 (Marriage), 272 (Nonalcoholic Beverages), 481 (Locomotion), and 165 (Mortality) were a 0 (false) and OCM codes 831 (Sexuality) and 683 (Offenses Against the Person) were a 1 (true). The results showed that this was the case for two paragraphs, the cases in lines 6348 and 6466. Now, to check the credibility of Figure 4.4, the code was re-written to filter out all the cases where OCM codes 165 (Mortality) and 762 (Suicide) were both a 1 (true) or a 0 (false). The results were:
165 = 0: 6992 cases
165 = 1: 54 cases
165 = 0 and 762 = 0: 6887 cases
165 = 1 and 762 = 1: 11 cases

These results show that only eleven cases were found where 165 (Mortality) and 762 (Suicide) were both coded with a 1. This comes down to 11/54, rounded up to 20% of the cases. This now explains why the results in Figure 4.4 and Figure 4.13 differ, the hierarchy is probably based on only two cases within the 7046 cases in the dataset (so a very small percentage), whereas the feature importance plot is based on probably 54 cases where 165 (Mortality) was coded with a 1 (but only with a success rate of 20% to determine Suicide). This cannot be said with 100% certainty as a model

never describes the data fully (so for 100%), but the figures are presumably based on the before mentioned two cases for the hierarchy and 54 cases for the feature importance plot.

To determine possible clusters within the `homicide_contain_p_encoding.csv` file and the `combined_homicide_data_encoding.csv` file, a Python script was written and run in Google Colab (see the link in the Appendix on page 65). The figures plotted can also be found in the Appendix (pages 86-89) and on GitHub. To visualise the difference between the 'p' file and the 'combined' file the plots are both shown next to each other. The first two figures show a heatmap with all the 7049 cases plotted (one for each data file). In the Appendix two other figures can be found that show a K-means clustering plot where Principal Component Analysis (PCA) was applied.

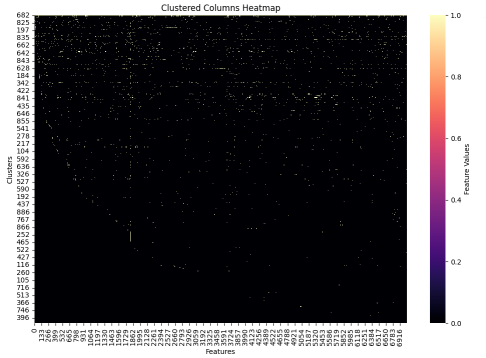

Figure 4.17: *Heatmap with found clusters, data file: homicide_contain_p_encoding.csv*

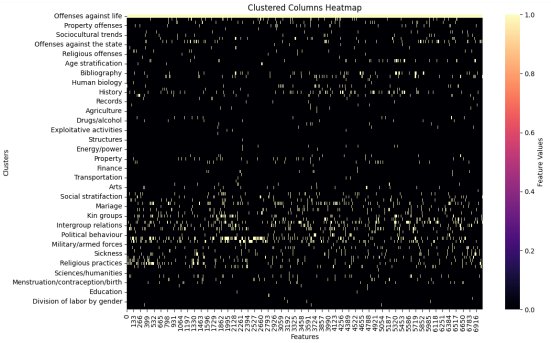

Figure 4.18: *Heatmap with found clusters, data file: combined_homicide_data_encoding.csv*

Chapter 5

Discussion

Optimizing the models

Figure 4.1 shows the nineteen different regression models and the average value for each metric of each model. For target 762, Suicide, we see that the Bayesian Ridge model is the best choice to use as a model to do analyses with. The most notable value in this table is the R^2 value. A higher R^2 value indicates a better fit for the model. The optimal, yet barely achievable, R^2 value is 1,0. The range of the R^2 value is typically between 0 and 1 (but, as said before, R^2 values can also be negative). R^2 values close to zero indicate a weak correlation, or sometimes even no correlation at all between the target variable and the other input variables. The larger the model is, the lower the R^2 value can become, sometimes resulting in a negative R^2 value[23]. In Figure 4.1 we see negative R^2 values for the models 'Lasso Regression' to 'Linear Regression'. The R^2 value for the Linear Regression model is even so low, namely $-53,14 \cdot 10^{22}$, that we can confidently say that the relationships between the variables (in this case the target variable and the other 528 different OCM codes) are not linear. The researchers Snijders & Bosker[42] offer two explanations for a decrease of the R^2 value and/or a negative R^2 value in a larger model. The first explanation is that there is a chance of fluctuation (or sampling variance) which is the most prominent when the sample size is small. Another explanation can be miss-specification of the model, this occurs when the new predictor is redundant in relation to one or more other predictors in the model.
In Figures 4.2 and 4.3 the difference between a tuned and non-tuned model can be seen. The R^2 value was chosen to optimize, as Figures 4.2 and 4.3 show that tuning this model with the `tune_model()` command gives an R^2 improvement of 0,0032; the R^2 value was 0,0065 before (seen in 4.2) and

0,0097 after (seen in 4.3). Possible other solutions to improve the model, and therefore increase the R^2 value, are selecting only relevant variables to do analyses with, and focusing on refining the model's features. This could be done by tuning the hyperparameters by hand for instance. A hyperparameter is an external variable that specifies details of the learning process of the model. An extra experiment was conducted to research whether reducing the amount of variables would improve the model. This experiment can be found in the subsection below.
An explanation to why the model only improved by 0,0032 could be that the hyperparameters were not tuned enough by hand when tuning the model. To research if that is the case, every hyperparameter can be tuned by hand in future research.
Apart from the 0,0032 improvement after tuning the R^2 value, the figures also show that the RMSE value decreased by 0,0002, from 0,1242 to 0,1240. This could be discarded as the change is almost insignificant, but it is still an indication that the model improved after tuning. Because, as stated before, the lower the RMSE value, the better.
Overall do the R^2 value of 0,0097 and RMSE value of 0,1240 indicate that the regression model Bayesian Ridge does not fit the data well for target variable 762 (Suicide). It therefore also indicates that the model's ability to make accurate predictions on the target variable is limited. In the appendix on pages 70 to 78 the other tables with the results from all the chosen target variables can be found. In every table we see that the R^2 value never becomes higher than 0,1263, seen in Figure 7.10. This R^2 value being closer to 0 than to 1, and all the other R^2 values also being closer to 0 than to 1, indicate that regression models are not the best fit for evaluating this data. Other supervised learning and unsupervised learning techniques can be evaluated in future research, to research if other machine learning techniques are a better choice to do analyses with. An example of a machine learning technique that can be used for both supervised learning and unsupervised learning is a neural network. A neural network can be used on large datasets and is specialised in recognizing patterns and solving problems, they are models that are composed of different elements, or units, which combine multiple inputs together and produce a single output as result[18]. Another possible machine learning technique that can be used on this data is association rule mining, which is a technique that searches for relationships among variables[53].
Back to Figure 4.1 and Bayesian Ridge being the best model for analyses given target variable 762 (Suicide). In Table 3.1 an overview is given of the regression model types boosting, ensemble, greedy, linear, and non-linear, and what the advantages and disadvantages are of every type of model.

Now, as can be seen in Table 3.1, Bayesian Ridge is a linear model. This can be seen as contradicting with the assumption that the data is not linear as Figure 4.1 shows that the R^2 value for the Linear Regression model is extremely high ($-53,14 \cdot 10^{22}$). However, multiple possible explanations can be given to why Bayesian Ridge was still stated as the best model by PyCaret. First it is stated that the Bayesian Ridge model may not capture non-linearity or complex patterns, so even if the data would not be linear, this model might not be able to catch that. Second, an experiment will be conducted to research whether or not noise within the data would have tempered with the results, making it possible that reducing the amount of variables could also lead to another model being the best fit for this data with target variable 762, Suicide, making Bayesian Ridge indeed not the best fit for this target. The experiment conducted is explained below.

Reducing variables

To research whether reducing the amount of variables would improve the model, an extra experiment was conducted. In this research the `feature_selection` function was added to the setup. This resulted in the following setup function:

```
from pycaret.regression import *
exp_762 = setup(data = data, target = '762',
          session_id=100, feature_selection = True)
```

This immediately led to a new model being the best model to do analyses with, namely the Gradient Boosting Regressor model, with an R^2 value of 0,0157 and MSE and RMSE values of respectively 0,0159 and 0,1239. Table 3.1 shows that the Gradient Boosting Regressor model can be categorised as a non-linear, ensemble, and boosting method. This model being a non-linear model is an explanation to why it performs better than the Bayesian Ridge model, as it was hypothesised that the data is non-linear (mentioned before). All the results of this experiment can be found in the Appendix on pages 90 and 91. After using the `tune_model()` function where the R^2 value was optimized, the R^2 value even increased from 0,0157 to 0,0204. Lastly, as done before, the feature importance plot of this model with target 762 (Suicide) was obtained. The results of this figure differ with the previously found results, which was expected as the program applied feature selection and thus removed certain variables from the data before doing the analyses and creating the feature importance plot. In conclusion, seeing that the R^2 value improved by 0,0107 (the R^2

value of the Bayesian Ridge model was 0,0097 after tuning and the R^2 value of the Gradient Boosting Regressor was 0,0204 after tuning), we can say that applying feature selection will improve the data analysis. Noise within the data file was thus one of the reasons why the Bayesian Ridge model did not perform that well and had an R^2 value of 0,0097, thus being close to zero.

Feature importance plots

In the subsections below the feature importance plots shown in section 4, Results, will be discussed. The results are discussed in three groups, namely strategies, communities and offenses. It should be mentioned that the data provided, and therefore analysed in this study, was originally already focused on OCM code 682, Offenses Against Life. Meaning that all the results found should be seen from a perspective where we analyse Offenses Against Life data. An example of this is: when looking at Suicide within an Offenses Against Life dataset, one may find different results than when one would look at Suicide within a Sex and Marital Offenses dataset. Or, when one would look at homicides (a form of an offense against life) within a Suicide dataset. That being said, in the subsections below the feature importance graphs obtained in this research will be discussed.

Strategies

Suicide threats, sorcery and warfare can all be seen as bargaining strategies. Therefore targets 762 (Suicide), 754 (Sorcery) and War & Peacemaking will be discussed together in this section.
In Figure 4.4 we see the feature importance plot of target 762, Suicide. The variables Mortality, Special Burial Practices and Funerals, and Cult of the Dead are all related directly to death and therefore suicide and will for that reason not be explained in more detail. The found variables Sexual Stimulation, Sexuality, Termination of Marriage, and Conception are all in line with the theory that states that conflicts such as forced marriages and sexual assault are associated with suicidal behaviour, this is also consistent with the findings of Syme, Garfield, and Hagen (2016)[45] that suicidal behaviours in the HRAF are associated with young people, sexuality, and reproduction. Research (see [9], and [45]) indicates that main predictors of suicidal behaviour are extreme conflict and powerlessness, here, every variable mentioned in the figure can be tied to either extreme conflict or powerlessness. For Conception for instance, it can be tied to both extreme

conflict when a partner wants to divorce their spouse (also Termination of Marriage) when she cannot fall pregnant (powerlessness as the woman may want to get pregnant but is unable to).
In Figure 4.11 it can be seen that Sorcery is associated with Ingroup Antagonisms, Community Structure, Informal Ingroup Justice, and Offenses Against the Person which are all indicators of sorcery being done within a community. In history it is common that sorcery would be used as a weapon (a bargaining strategy) between communities and groups, but research[32] also shows that sorcery is found within a group and thus between ingroup members. A possible explanation for the findings in this study could be that the texts in the HRAF were more focused on relationships within families and kin, than being focused on relationships between groups. The Life History Materials variable could therefore have been influenced by one text in the data file.
The variables in Figure 4.12 make between themselves sense as Agriculture, Territorial organisations/state, Machines/tools, Food consumption, and Food processing, are all directly linked to each other. Agriculture is a known conflict in both history and nowadays, groups have been fighting for land for centuries[35] and archaeologist link warfare and agriculture directly to each other. War is also a concept between groups, which explains why the variable Intergoup Relations is among the highest variables to predict War and Peacemaking events. Warfare is thus a bargaining strategy between groups for, among other things, land (agriculture).

Communities

Targets 578 (Ingroup Antagonisms), 627 (Informal Ingroup Justice), and 628 (Inter-community Relations) are all targets that have something to do with communities, whether it is within a community/group or between groups. Both within groups and between groups does conflict occur, variables Informal Ingroup Justice, Community Structure, Kin Relationships, Community Heads, Sibs, Lineages, Ingroup Antagonisms, Household, and Tribe and Nation are all variables linked to conflicts within groups. The variables Warfare, Social Control, Social Relationships and Groups, Intercommunity Relations, Weapons, and Instigation of War can be linked to conflict between groups. A known bargaining strategy is arranging a marriage[9], Figure 4.7 supports this as Arranging a Marriage is found as one of the predictable variables of Inter-community Relations. An arranged marriage is often seen as a bargaining strategy between groups, but as a conflict between kin (when the arranged marriage is a forced marriage). Everything linked to war such as Weapons and Instigation of War,

etc. can be directly linked to conflict, which is then linked to other-directed harm (and sometimes self-harm) between people.

Offenses

Offenses Against the Person, Sex and Marital Offenses, and Property Offenses, respectively targets 683, 684, and 685 are all different types of offenses. In Figures 4.8, 4.9 and 4.10 we see that there are a lot of similarities between the outcome of the plots. This can be linked to at least two things, 1) the types of offenses are fairly similar with all three types being cases of other-directed harm, and 2) because the OCM codes are similar and ascending it is possible that many texts contain multiple offense codes as the author can easily flag multiple similar OCM codes at the same time. Table 4.1 (see page 32) also shows that OCM codes 682 (Offenses Against Life) and 683 (Offenses Against the Person) are in the top 10 most frequently used OCM codes in the OAL data file. It is a limitation of this study that same authors, books, articles and paragraphs were not separately analysed during the analyses.
In Figure 4.8 the highest predictable variable is Sorcery, which can be explained as sorcery is one of the most common and oldest forms of aggression throughout history. Other striking variables in this Figure are Liability, Grandparents and Grandchildren, and Slavery. Liability can be explained as this variable could indicate who the perpetrators are and how they were held accountable for their actions. Slavery is still legal in several countries in the world, which could indicate why it is in this top 10. An offense that has to do with slavery could be a slave running away, this would be seen as an offense against his or her owner.
Figure 4.9 shows the variables that are commonly used to predict events with Sex and Marital Offenses, here the variables Extramarital Sex Relations, Organized Ceremonial (like a marriage), Sexual Intercourse, and Secondary Marriages (which could be linked to divorce) speak for themselves. The other variables such as Property Offenses (linked as an offense to the other offenses), Social Offenses, and Pastoral Activities are all variables that indicate a relation between two people, which is the same for Sex and Marital Offenses. In this figure it is mostly striking that Organized Ceremonial (which could also indicate an arranged marriage) is not ranked higher, as the theory indicates that an arranged marriage can be seen as a conflict and a conflict could lead to Sex and Marital Offenses.
In the figure of Property Offenses (Figure 4.10) different offenses are the highest ranked predictable variables. Other variables such as Property in Movables, and Vehicles indicate things (properties) of a person. Most

variables in this figure indicate punishments, which is in line with for example the Legal and Judicial Personnel variable, which can be seen as a third party involved to solve the conflict.
Overall, it can be said that the variables Texts Translated into English, Texts Administered in the Field, Conversation, and Nonfulfillment of Obligations are all variables that could have been mentioned by one or several authors using these variables in various, similar texts, often enough to be showed in these figures, but not often enough to be significant in the overall study. This is in line with the findings of the frequencies of the OCM codes (which can be found on GitHub).

'p' vs 'combined' file

When comparing the 'p' file results with the 'combined' file results we see that mostly the runtime is decreased when using the 'combined'. The runtime to obtain the Bayesian Ridge model for target variable 685 (Property Offenses) for example, is 0,5370 seconds (see Figure 7.8 in the Appendix on page 75). The runtime to obtain the Bayesian Ridge model for target group 'War and Peacemaking' is 0,0410 seconds (see Figure 7.12 in the Appendix on page 78). This results in a difference of 0,496 seconds between obtaining both Bayesian Ridge models, while for both target variables the same model is compared. Using the 'combined' file thus decreases the runtime. Another model, the Orthogonal Matching Pursuit model, was chosen as best model to analyse the Drugs and Alcohol target with. The decrease in runtime and newly introduced model can be explained by the decrease in data volume when grouping several OCM codes together, also resulting in less noise in the data. Another explanation can be that there are less variables in total to compare with each other as there were 93 groups in total instead of 529 separate variables (the 529 OCM codes found in the data). The runtime did not decrease, however, because of the use of other functions. The code and functions used for the analyses of the 'p' file and the 'combined' file were both the same. The conditions for both files were thus the same, except for the OCM codes which were all separate in the 'p' file and combined in the 'combined' file.

Hierarchy and clusters

In section 4.5, Hierarchy and clusters, the decision tree of OCM code 762 (Suicide) is shown. The hierarchy tree shows that sex and social offenses (which in this case could be perhaps infidelity) predict that OCM code 762

(Suicide) will be a 1, in other words: will be true. This is consistent with various research (see [45] and [46]) that state that suicidal behaviour in this data is more often associated with sexual conflict and transgressions, than with other factors such as old age.
Figures 4.13, 4.14, 4.15, and 4.16 show four different decision trees, all for target variable 762 (Suicide). Every tree is made with either both pruning and prepruning applied, only pruning or prepruning applied or no pruning and prepruning applied at all. The trees in Figures 4.13 and 4.15 show similarities, in both these trees at least prepruning was applied. The difference in pruning and prepruning is that with pruning the program waits until the whole tree is finished before cutting branches, whereas with prepruning the program stops creating branches before it has completed classifying the entire training set. Now, in Figures 4.13 and 4.15 we see that applying prepruning creates a smaller tree with less branches and leaves (the variables, in this case OCM codes), compared to the trees in Figures 4.14 and 4.16 where only pruning was applied or no (pre)pruning at all was applied. Applying prepruning has in this case resulted in the 'Social Insurance' branch being cut off, as the trees in Figures 4.14 and 4.16 start with OCM code 745 (Social Insurance), and the trees in Figures 4.13 and 4.15 start with OCM code 886 (Senescence).

Last but not least was there some research done on clusters within the data files. Cluster analysis is typically done when there is no assumption made about likely relationships within the data. It provides information about if and where associations and patterns in the provided data exist, but not what those might be or what they mean. Because no relationship between the 7046 cases and 529 OCM codes were suspected, cluster analysis was done.
Figures 4.17 and 4.18 show no cluster forming. This is as expected, as the parameters were set to find clusters withing the 7046 separate cases. This now concludes that the cases in the OAL file are all separate and not linked to each other.
Figures 7.33 and 7.34, are shown in the Appendix (pages 88 and 89) and do show cluster forming. The analysis done in these figures is K-means clustering with PCA, the Principle Component Analysis is done before applying the K-means clustering algorithm. As stated before is PCA a dimensionality reduction technique. It transforms the original features into a new set of uncorrelated features, called principal components. In the figures these components are called Principle Component 1 and Principle Component 2. When adding the parameter `k = 5` with `k` being the number of clusters, we can find cluster forming in the 'p' file and the 'combined' file. When choosing $k > 5$ no distinguishable clusters could be found,

therefore $k = 5$ was chosen as this showed five distinguished clusters, with 5 being a personal preference. These figures were plotted after transposing the data from a row-based dataset to a column-based dataset to determine clusters within the 529 OCM codes (or in the case of the combined file, 93 groups). The red 'x' shown on the figures shows the centroid of the found and said cluster, not to be mistaken with the *center* of the clusters. The centroid of a cluster is determined after summing up the position of all individual data points of a single cluster, and then dividing that number by the number of data points. Interesting to see in Figures 7.33 and 7.34 is that the centroid is not always in the center. The combined data file shows more clusters close to each other, with three clusters grouped together and two clusters grouped together. Unlike the clusters of the p-file where only two clusters are grouped together. In this case, as the input data was only binary encoded OCM codes and binary encoded groups of OCM codes, it could be possible to identify and label the five found clusters. To identify the dimensions of the PCA components, more research needs to be done. PCA tries to put as much information as possible in the first component, then in the second, and so on, until a graph can be obtained. In this case domain experts expect PCA1 to be linked to 'Sanctions', but then PCA2 could anything, from 'Kinship' to 'Religion', for instance. It is impossible to determine the five clusters and PCA's, without knowing what OCM codes are loaded into the Principle Components.

Limitations

This research does have a limitation, which is the source of the individual cases in the data. When doing the analyses, the source of each case was not taken into account. This means that every case was treated individually, so in this research it is not known if an author is being mentioned several times in the same dataset. Same goes for articles, books, etc. being mentioned several times. The disadvantage of this is that we do not know if one particular author for instance would use a certain OCM code more often than others. Personal preferences of authors, books, articles, etc. could have affected the results.

Future research

When wanting to continue this research, some modifications can be done to improve future research and results. First of all, one could look at co-

variation between the OCM codes within this dataset. Or a larger dataset could be obtained, for instance the whole HRAF. In this case analyses can be done on all the data known within the HRAF, making the models and results more reliable.
When doing future research one should also look at the geographical specifics of the OCM codes. Meaning, we see different types of conflict in different places on earth. In some geographical locations warfare is more focused on land for example, where in other locations the cause of war is a conflict over scarce sources. Other examples that are geographically focused are sorcery and suicide, research[52] has stated that seniors aged 75 and older have the highest suicide rates of all age groups in most industrialized countries, and sorcery is still used in various countries nowadays.
Last but not least, future research can also be done on cluster analyses. This time the K-means algorithm was used, which is a centroid model. Other possibilities to do cluster analysis with are density models[15] such as DBSCAN (Density Based Spatial Clustering of Applications with Noise) and OPTICS (Ordering Points To Identify the Clustering Structure), or a subspace model[31] such as bi-clustering. These models could be used to do analyses with and are relevant in future research as density-based clustering models (such as DBSCAN) can find clusters of arbitrary shape, and determine what information should be classified as noise or outliers. Subspace clustering could be useful in future research as it can be applied on high dimensional data.

Version of February 29, 2024– Created February 29, 2024 - 19:58

Chapter 6

Conclusion

The main research question of this research was: What machine learning techniques support the analysis of offenses against life data extracted from the HRAF, and can help predict future events where other-directed harm and self-harm could occur?

Possible main machine learning techniques that can be used to do analyses with on the data extracted from the HRAF are supervised regression models. However, the metrics used to investigate the models in this research, namely the R^2, MSE and RMSE values, indicate that regression models are not the best fit to research this data. In Figure 4.1 it can be seen that the R^2 value for the Bayesian Ridge model with target variable 762 (Suicide), is 0,0065. After tuning the model this value increased to 0,0097, but this is still closer to 0 than to 1. The MSE and RMSE values are respectively 0,0158 and 0,1242. The experiment conducted afterwards supports the hypothesis that there was a lot of noise in the data and that variable selection could help with improving the models. Figure 7.35 in the Appendix shows that after using the `feature_selection` function the R^2 value can be increased to 0,0157, and after tuning the new chosen model (the Gradient Boosting Regressor model) that value can even be increased to 0,0204. But, as this value is still closer to 0 than to 1 as before, it can be concluded that regression models are not the best fit for this data. To improve the models of this study more hyperparameters need to be added to the models to call the results reliable. This does not mean that the results are entirely useless, but models that fit the data better need to be created and analysed before concluding real findings. Other possible techniques to analyse the data are neural networks and association rule mining. Lastly, other machine learning techniques that support the analysis

of this data and can help predict future other-directed harm and self-harm events are thus unsupervised cluster analysis and Principle Component Analysis, seen in Figures 4.17 and 4.18 on page 49 and in Figures 7.33 and 7.34 on pages 88 and 89 in the Appendix.

The domain questions of this research were:
2) What correlations, relations and patterns can be found in the provided data, focusing on suicide events and different dimensions of aggression, using machine learning models?
3) What information and relationships such as clusters can be deducted from the OCM codes given in the provided offenses against life data file?
4) What variables are important when predicting specific events such as a suicide event or other cases important to other-directed and self-directed harm?

Overall it can be said that there are many regression models possible to do analyses with on one-hot encoded data, specifically the ethnographic data used in this research. The possible regression models that can be used for this data are: Bayesian Ridge, Light Gradient Boosting Machine, Extreme Gradient Boosting, Gradient Boosting Regressor, Orthogonal Matching Pursuit, and K Neighbors Regressor. The feature importance plots in section 4, Results, show that there are correlations to be found between OCM codes, as some codes help predict other codes, in this case the chosen target value. Furthermore, do the decision trees shown in Figures 4.13, 4.14, 4.15, and 4.16 show that a hierarchy can be found for target variable Suicide. The path that predicts OCM code 762 (Suicide) being a 1 (thus being true) in Figure 4.13 is OCM codes 886 (Senescence), 580 (Marriage), 272 (Nonalcoholic Beverages), 481 (Locomotion), and 165 (Mortality) being false, and OCM codes 831 (Sexuality) and 683 (Offenses Against the Person) being true. This results in two found cases for the `homicide_contain_p_encoding.csv` file. The findings of this hierarchy are consistent with the theory mentioned and findings in previous research[45, 46], as sex and social offenses are known conflicts for suicidal behaviours. The suicide events mentioned in the found cases of this hierarchy can thus be linked to being potential bargaining strategies.

The heatmap figures shown in section 4, Results, show that there is no cluster forming to be found between the cases of the OAL data file. A relationship between OCM codes can be deducted from Figures 4.13, 4.14, 4.15 and 4.16 that show the hierarchy found for OCM code 762 (Suicide). Figure 4.13 and Figure 4.15 show that when prepruning is applied, OCM

code 745 (Social Insurance) is cut off, as can be seen in Figures 4.14 and 4.16 where OCM code 745 (Social Insurance) is at the top of the tree. The decision trees then start at OCM code 886 (Senescence).

As can be seen in Figure 4.4 the 10 most important variables to predict OCM code 762 (Suicide) are thus Mortality, Special Burial Practices and Funerals, Sexual Stimulation, Personality Disorders, Sexuality, Physical Descriptions, Termination of Marriage, Conception, Pharmaceuticals, and Cult of the dead. This is in line with previously mentioned theory which states (in addition to other things) that forced sexual acts and arranged marriages are types of conflict, and therefore predict suicidal behaviour in mostly people with less power, such as women and young people. However, after conducting an experiment where feature selection was applied to reduce noise, it was found that the top 10 most important variables to predict OCM code 762 (Suicide) are Mortality, Special Burial Practices and Funerals, Liability, Legal Norms, Trial Procedure, Eschatology, Termination of Marriage, Execution of Justice, Judicial Authority, and Personality Disorders. The variables Mortality, Special Burial Practices and Funerals, Personality Disorders, and Termination of Marriage are found in both feature importance graphs 4.4 and 7.38. The differences between the results of Figures 4.4 and 7.38 can be explained by the use of the `feature_selection()` function, used in the regression model when obtaining Figure 7.38.
Other variables that are important to predict cases important to other-directed harm can be seen in Figure 4.8, which shows the feature importance graph of target variable Offenses Against the Person. The most important variable to predict OCM code 683 (Offenses Against the Person) is in this data thus Sorcery, in line with research[35] that shows that sorcery is a known ingroup offense between people.

In conclusion, regression models, K-means clustering and PCA are useful techniques to study the ethnographic data extracted from the HRAF in a machine learning way. However, to obtain more reliable results one should first investigate other machine learning techniques such as neural networks and association rule mining. Furthermore, no correlations were found in the offenses against life data file. Relations between OCM codes were found when predicting certain OCM codes and looking at the feature importance plots. The hierarchy in the form of a decision tree found when predicting OCM code 762 (Suicide) also indicates a relationship between variables.

Chapter 7

Appendix

GitHub

The GitHub link to all the code used and the original and modified data files:
https://github.com/maxine-mxl/Thesis-Maxine

Colab

On GitHub the entire Google Colab Notebook can be found. The link below is a direct link to the Colab Notebook:
https://colab.research.google.com/drive/Thesis-Maxine

Code

Below you can find the code to apply the one-hot encoding method on your dataset, written in R by Dr. K.L. Syme, who was kind enough to share her code for this research.

```
#one-hot encoding

hs <- read_xlsx("/Users/Documents/location/name_file.xlsx")

cc <- str_split(hs$ocms, ",")

unique_ocms2 <- unique(unlist(cc))

binary_matrix2 <- matrix(0, nrow = nrow(hs), ncol = length(unique_ocms2))

colnames(binary_matrix2) <- unique_ocms2

for (i in 1:nrow(hs)) {
  ocm_values <- unlist(cc[i])
  binary_matrix2[i, ocm_values] <- 1
}

hs_combined <- cbind(hs, binary_matrix2)

write.csv(hs_combined, "/Users/location/name_file.csv", row.names = FALSE)
```

Here the code to extract the suicide cases from the offenses against life dataset can be found.

```
import pandas as pd

homicide_data = pd.read_excel('./data/homicide.xlsx')
suicide_data = pd.read_csv('./data/suicide.csv', sep=',')

homicide_data['text'] = homicide_data['text'].fillna('')

homicide_data['contains_ocm762'] = homicide_data.text.apply(lambda x: '762' in x)

#homicide_data['contains_word_suicide'] = homicide_data.text.apply(
   # lambda x: 'suicide' in x or 'Suicide' in x
   # or 'killed themselves' in x or 'Killed themselves' in x
   # or 'killed herself' in x or 'Killed herself' in x
   # or 'killed himself' in x or 'Killed himself' in x
   # or 'self-murder' in x or 'Self-murder' in x
   # or 'self murder' in x or 'Self murder' in x
  #  or 'self-slaughter' in x or 'Self-slaughter' in x
#)

homicide_check_data = homicide_data.loc[homicide_data['contains_ocm762'] == True]

print(homicide_check_data)

homicide_check_data.to_excel('./data/762_check.xlsx')
```

The code to determine the frequencies of the (top 10) OCM codes can be seen below.

```
import csv

input_csv_file = 'homicide_contain_p_encoding.csv'
output_csv_file = 'output_p_all_frequencies.csv'

#creating a list to store the count of '1' for each column
column_counts = []

with open(input_csv_file, 'r', newline='') as file:
    reader = csv.reader(file)

    header = next(reader, None)
    column_count_dict = {}

    #iterate through each column and initialize the count to 0
    for column in header:
        column_count_dict[column] = 0

    #count the numbers of '1'
    for row in reader:
        for i, value in enumerate(row):
            if value == '1':
                column_count_dict[header[i]] += 1

#append the counts to the column_counts list
column_counts = [(column, count) for column,
                 count in column_count_dict.items()]

with open(output_csv_file, 'w', newline='') as output_file:
    writer = csv.writer(output_file)

    #headers for the output csv
    writer.writerow(["OCM code", "Frequency"])

    for column, count in column_counts:
        writer.writerow([column, count])

print(f"Answers saved to {output_csv_file}")
```

The code used to determine all the cases where OCM codes 886, 580, 272, 481, and 165 were false and codes 831 and 683 were true can be seen below. The code was written in Python and the output were two cases within the dataset.

```
import pandas as pd

#import data
data = pd.read_csv('./data/homicide_contain_p_encoding.csv')

#886, 580, 272, 481, and 165 are false
#831 and 683 are true
filtered_data = data[(data['886'] == 0) & (data['580'] == 0) & (data['272'] == 0) &
(data['481'] == 0) & (data['165'] == 0) &
(data['831'] == 1) & (data['683'] == 1) ]

#display the filtered data
print(filtered_data)
```

The code used to determine all the cases with OCM codes 165 and 762 can be seen below. The code was written in Python and the output were eleven cases where both OCM code 165 and 762 were true (a 1).

```
import pandas as pd

#import data
data = pd.read_csv('./data/homicide_contain_p_encoding.csv')

filtered_data = data[(data['165'] == 1) & (data['762'] == 1)]

#display the filtered data
print(filtered_data)
```

Tables

Outcome of the best models for target '578' (which is the OCM code for Ingroup Antagonisms). As can be seen in the table, Light Gradient Boosting Machine is the best model to do analyses with and to predict values with in the future.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.1791	0.0887	0.2975	0.0565	0.2091	0.8219	0.9680
br	Bayesian Ridge	0.1809	0.0889	0.2978	0.0544	0.2083	0.8436	0.5620
gbr	Gradient Boosting Regressor	0.1793	0.0902	0.3001	0.0392	0.2096	0.8545	2.7120
omp	Orthogonal Matching Pursuit	0.1821	0.0907	0.3009	0.0347	0.2113	0.8381	0.2040
xgboost	Extreme Gradient Boosting	0.1735	0.0920	0.3030	0.0205	0.2121	0.8034	0.6970
ridge	Ridge Regression	0.1892	0.0934	0.3054	0.0051	0.2141	0.8170	0.1220
llar	Lasso Least Angle Regression	0.1883	0.0942	0.3066	-0.0016	0.2143	0.8948	0.1900
dummy	Dummy Regressor	0.1883	0.0942	0.3066	-0.0016	0.2143	0.8948	0.0700
en	Elastic Net	0.1883	0.0942	0.3066	-0.0016	0.2143	0.8948	0.0880
lasso	Lasso Regression	0.1883	0.0942	0.3066	-0.0016	0.2143	0.8948	0.0910
rf	Random Forest Regressor	0.1801	0.1007	0.3169	-0.0712	0.2294	0.6990	23.6870
ada	AdaBoost Regressor	0.2313	0.1030	0.3196	-0.0887	0.2331	0.8101	0.6190
huber	Huber Regressor	0.1053	0.1052	0.3239	-0.1174	0.2246	0.9998	1.7080
knn	K Neighbors Regressor	0.2055	0.1056	0.3244	-0.1228	0.2398	0.7088	0.1520
et	Extra Trees Regressor	0.1700	0.1226	0.3496	-0.3045	0.2477	0.6734	31.1410
dt	Decision Tree Regressor	0.1687	0.1296	0.3594	-0.3781	0.2529	0.6677	0.6120
par	Passive Aggressive Regressor	0.3924	0.2744	0.5146	-1.9674	0.3357	0.9759	0.1390
lr	Linear Regression	12823823065.3597	384895417316961943552000.0000	150043870537.6072	-39879119896528215133388 8.0000	2.5397	16588873323.5282	0.2980
lar	Least Angle Regression	5634386409.7109	391273273359278800896000.0000	62551840369.6414	-46249673094397105589452 8.0000	0.4655	15096480978.4699	0.4210

Figure 7.1: *Table with the evaluation of the regression models, target variable: 578 (Ingroup Antagonisms), input variables: all the 529 OCM codes, file used: homicide_contain_p_encoding.csv*

Outcome of the best models for target '627' (which is the OCM code for Informal Ingroup Justice). As can be seen in the table, Light Gradient Boosting Machine is the best model to do analyses with and to predict values with in the future.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.1981	0.0975	0.3115	0.0861	0.2191	0.7816	0.8750
br	Bayesian Ridge	0.2028	0.0996	0.3150	0.0650	0.2203	0.8227	0.8160
gbr	Gradient Boosting Regressor	0.2006	0.1001	0.3157	0.0610	0.2206	0.8227	3.0680
omp	Orthogonal Matching Pursuit	0.2052	0.1011	0.3174	0.0508	0.2229	0.8139	0.1430
xgboost	Extreme Gradient Boosting	0.1966	0.1022	0.3190	0.0403	0.2226	0.7811	0.7810
ridge	Ridge Regression	0.2120	0.1040	0.3220	0.0222	0.2260	0.7940	0.2430
llar	Lasso Least Angle Regression	0.2135	0.1068	0.3262	-0.0027	0.2283	0.8785	0.1430
dummy	Dummy Regressor	0.2135	0.1068	0.3262	-0.0027	0.2283	0.8785	0.0780
en	Elastic Net	0.2135	0.1068	0.3262	-0.0027	0.2283	0.8785	0.1600
lasso	Lasso Regression	0.2135	0.1068	0.3262	-0.0027	0.2283	0.8785	0.1700
ada	AdaBoost Regressor	0.2186	0.1068	0.3262	-0.0044	0.2313	0.8277	0.5350
rf	Random Forest Regressor	0.1996	0.1091	0.3297	-0.0258	0.2383	0.6822	24.0080
knn	K Neighbors Regressor	0.2017	0.1113	0.3329	-0.0463	0.2433	0.6941	0.1810
huber	Huber Regressor	0.1215	0.1214	0.3477	-0.1384	0.2411	0.9999	1.9880
et	Extra Trees Regressor	0.1893	0.1321	0.3630	-0.2451	0.2575	0.6731	30.8920
dt	Decision Tree Regressor	0.1889	0.1383	0.3715	-0.3043	0.2622	0.6745	0.4250
par	Passive Aggressive Regressor	0.4127	0.3020	0.5395	-1.8604	0.3518	0.9403	0.2380
lar	Least Angle Regression	2386779542.1563	330432300167081911 9104.0000	29072723411.5503	-339553231537722884 09600.0000	0.7964	0.7916	0.4000
lr	Linear Regression	21668158939.3471	140073753438934049751040.0000	258208403345.1348	-129233152735519455995 4944.0000	2.9389	8102821157.5122	0.8820

Figure 7.2: *Table with the evaluation of the regression models, target variable: 627 (Informal Ingroup Justice), input variables: all the 529 OCM codes, file used: homicide_contain_p_encoding.csv*

Outcome of the best models for target '628' (which is the OCM code for Inter-community Relations). As can be seen in the table, Light Gradient Boosting Machine is the best model to do analyses with and to predict values with in the future.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.1458	0.0707	0.2658	0.0788	0.1869	0.8253	0.6740
gbr	Gradient Boosting Regressor	0.1431	0.0713	0.2667	0.0730	0.1861	0.8510	2.3990
br	Bayesian Ridge	0.1485	0.0720	0.2682	0.0629	0.1873	0.8607	0.8280
xgboost	Extreme Gradient Boosting	0.1380	0.0720	0.2681	0.0625	0.1879	0.7986	1.0560
omp	Orthogonal Matching Pursuit	0.1510	0.0735	0.2708	0.0433	0.1902	0.8517	0.1310
ridge	Ridge Regression	0.1583	0.0752	0.2739	0.0218	0.1931	0.8301	0.1380
llar	Lasso Least Angle Regression	0.1541	0.0771	0.2773	-0.0013	0.1936	0.9159	0.1310
dummy	Dummy Regressor	0.1541	0.0771	0.2773	-0.0013	0.1936	0.9159	0.0660
en	Elastic Net	0.1541	0.0771	0.2773	-0.0013	0.1936	0.9159	0.0990
lasso	Lasso Regression	0.1541	0.0771	0.2773	-0.0013	0.1936	0.9159	0.1460
ada	AdaBoost Regressor	0.1601	0.0782	0.2795	-0.0237	0.1974	0.8752	0.4650
rf	Random Forest Regressor	0.1492	0.0795	0.2818	-0.0381	0.2043	0.7219	21.2110
knn	K Neighbors Regressor	0.1425	0.0802	0.2830	-0.0459	0.2060	0.7677	0.1650
huber	Huber Regressor	0.0842	0.0841	0.2896	-0.0913	0.2007	0.9997	1.7980
et	Extra Trees Regressor	0.1436	0.0982	0.3128	-0.2786	0.2225	0.7086	25.9760
dt	Decision Tree Regressor	0.1432	0.1039	0.3218	-0.3533	0.2273	0.7138	0.3670
par	Passive Aggressive Regressor	0.3165	0.1912	0.4336	-1.4935	0.2971	0.9082	0.1370
lar	Least Angle Regression	2692114304.5459	35136535118500562206720.0000	59774636479.3106	-450846914487442441502720.0000	1.3010	32.3737	0.3840
lr	Linear Regression	11790682427.3389	603457625054072044425728.0000	147799325819.4135	-7588176916703703119953920.0000	2.3306	3798166801.8313	1.0080

Figure 7.3: *Table with the evaluation of the regression models, target variable: 628 (Inter-community Relations), input variables: all the 529 OCM codes, file used: homicide_contain_p_encoding.csv*

Outcome of the best models for target '672' (which is the OCM code for Liability). As can be seen in the table, Extreme Gradient Boosting is the best model to do analyses with and to predict values with in the future.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
xgboost	Extreme Gradient Boosting	0.1239	0.0606	0.2458	0.1011	0.1725	0.7750	0.6880
lightgbm	Light Gradient Boosting Machine	0.1316	0.0612	0.2470	0.0933	0.1733	0.8218	0.9210
gbr	Gradient Boosting Regressor	0.1255	0.0621	0.2486	0.0846	0.1733	0.8502	2.6480
br	Bayesian Ridge	0.1353	0.0625	0.2495	0.0777	0.1744	0.8567	0.4970
ridge	Ridge Regression	0.1418	0.0628	0.2503	0.0701	0.1771	0.8093	0.1360
omp	Orthogonal Matching Pursuit	0.1360	0.0630	0.2504	0.0700	0.1761	0.8420	0.1280
rf	Random Forest Regressor	0.1269	0.0638	0.2524	0.0433	0.1837	0.6563	16.9610
knn	K Neighbors Regressor	0.1316	0.0671	0.2589	-0.0013	0.1894	0.7224	0.2580
llar	Lasso Least Angle Regression	0.1360	0.0680	0.2602	-0.0023	0.1816	0.9266	0.1210
en	Elastic Net	0.1360	0.0680	0.2602	-0.0023	0.1816	0.9266	0.1550
lasso	Lasso Regression	0.1360	0.0680	0.2602	-0.0023	0.1816	0.9266	0.0950
dummy	Dummy Regressor	0.1360	0.0680	0.2602	-0.0023	0.1816	0.9266	0.0650
huber	Huber Regressor	0.0734	0.0733	0.2700	-0.0780	0.1871	0.9994	1.9820
et	Extra Trees Regressor	0.1166	0.0731	0.2702	-0.1007	0.1930	0.6305	21.3170
ada	AdaBoost Regressor	0.1677	0.0755	0.2731	-0.1325	0.1988	0.8255	0.5390
dt	Decision Tree Regressor	0.1157	0.0753	0.2742	-0.1377	0.1951	0.6324	0.3470
par	Passive Aggressive Regressor	0.3255	0.1892	0.4312	-1.8834	0.3046	0.8730	0.1520
lar	Least Angle Regression	34846977.2614	5986556609138533376.0000	773728420.2506	-11983417866039463116 8.0000	0.5068	2.1648	0.5320
lr	Linear Regression	24888925911.9548	24702788208044337881088.0000	319944803339.1642	-32214197436189862589890 56.0000	2.4848	30241273554.9101	0.3080

Figure 7.4: *Table with the evaluation of the regression models, target variable: 672 (Liability), input variables: all the 529 OCM codes, file used: homicide_contain_p_encoding.csv*

Outcome of the best models for target '681' (which is the OCM code for Sanctions). As can be seen in the table, Bayesian Ridge is the best model to do analyses with and to predict values with in the future.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
br	Bayesian Ridge	0.0702	0.0340	0.1838	0.0337	0.1282	0.9275	0.6660
lightgbm	Light Gradient Boosting Machine	0.0707	0.0341	0.1839	0.0324	0.1287	0.9112	0.7900
gbr	Gradient Boosting Regressor	0.0665	0.0346	0.1852	0.0180	0.1294	0.9123	2.6720
omp	Orthogonal Matching Pursuit	0.0732	0.0347	0.1856	0.0137	0.1306	0.9146	0.1310
ada	AdaBoost Regressor	0.0697	0.0349	0.1859	0.0124	0.1295	0.9373	0.4570
llar	Lasso Least Angle Regression	0.0707	0.0354	0.1872	-0.0019	0.1303	0.9633	0.1300
dummy	Dummy Regressor	0.0707	0.0354	0.1872	-0.0019	0.1303	0.9633	0.0650
en	Elastic Net	0.0707	0.0354	0.1872	-0.0019	0.1303	0.9633	0.1470
lasso	Lasso Regression	0.0707	0.0354	0.1872	-0.0019	0.1303	0.9633	0.0940
ridge	Ridge Regression	0.0802	0.0356	0.1880	-0.0115	0.1330	0.9077	0.2130
huber	Huber Regressor	0.0367	0.0367	0.1906	-0.0380	0.1322	0.9999	1.8520
xgboost	Extreme Gradient Boosting	0.0659	0.0367	0.1911	-0.0488	0.1347	0.8875	0.6910
knn	K Neighbors Regressor	0.0599	0.0388	0.1963	-0.1078	0.1419	0.9162	0.1750
rf	Random Forest Regressor	0.0722	0.0403	0.2001	-0.1521	0.1450	0.8575	16.6370
et	Extra Trees Regressor	0.0706	0.0526	0.2285	-0.5017	0.1613	0.8472	21.8770
dt	Decision Tree Regressor	0.0701	0.0546	0.2330	-0.5627	0.1635	0.8359	0.3240
par	Passive Aggressive Regressor	0.2049	0.0840	0.2862	-1.3680	0.2171	0.8806	0.2350
lar	Least Angle Regression	1022580405.6705	1238099652606732009472.0000	11587946887.6208	-44852554071583399870464.0000	0.4766	8815641574.9619	0.4650
lr	Linear Regression	9838654433.3432	2399325993869459115 2128.0000	98397441346.8397	-81257939438378800696524 8.0000	2.2081	369667012 81.8849	0.3110

Figure 7.5: *Table with the evaluation of the regression models, target variable: 681 (Sanctions), input variables: all the 529 OCM codes, file used: homicide_contain_p_encoding.csv*

Outcome of the best models for target '683' (which is the OCM code for Offenses Against the Person). As can be seen in the table, Gradient Boosting Regressor is the best model to do analyses with and to predict values with in the future.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
gbr	Gradient Boosting Regressor	0.1056	0.0535	0.2311	0.0482	0.1611	0.8850	2.3610
br	Bayesian Ridge	0.1098	0.0542	0.2324	0.0382	0.1620	0.9041	0.7940
lightgbm	Light Gradient Boosting Machine	0.1120	0.0542	0.2325	0.0369	0.1633	0.8851	0.5860
omp	Orthogonal Matching Pursuit	0.1128	0.0549	0.2341	0.0233	0.1644	0.8837	0.1280
xgboost	Extreme Gradient Boosting	0.1028	0.0553	0.2349	0.0157	0.1645	0.8363	1.0150
ridge	Ridge Regression	0.1199	0.0556	0.2356	0.0109	0.1662	0.8683	0.1290
llar	Lasso Least Angle Regression	0.1128	0.0564	0.2372	-0.0012	0.1653	0.9400	0.1250
dummy	Dummy Regressor	0.1128	0.0564	0.2372	-0.0012	0.1653	0.9400	0.0690
en	Elastic Net	0.1128	0.0564	0.2372	-0.0012	0.1653	0.9400	0.0870
lasso	Lasso Regression	0.1128	0.0564	0.2372	-0.0012	0.1653	0.9400	0.1100
ada	AdaBoost Regressor	0.1115	0.0572	0.2387	-0.0134	0.1666	0.9238	0.4670
knn	K Neighbors Regressor	0.0963	0.0582	0.2408	-0.0364	0.1735	0.8311	0.1530
huber	Huber Regressor	0.0601	0.0600	0.2445	-0.0636	0.1695	0.9998	1.7210
rf	Random Forest Regressor	0.1130	0.0621	0.2489	-0.1071	0.1811	0.7377	18.3450
et	Extra Trees Regressor	0.1092	0.0771	0.2774	-0.3746	0.1977	0.7169	23.9200
dt	Decision Tree Regressor	0.1083	0.0813	0.2847	-0.4455	0.2014	0.7117	0.3370
par	Passive Aggressive Regressor	0.2888	0.1456	0.3791	-1.5997	0.2768	0.8398	0.1360
lr	Linear Regression	11470968500.6287	2879263737987493569495O4.0000	210439815046.7644	-489365793508779878252544O.0000	2.2689	123973420400.0298	0.4530
lar	Least Angle Regression	40245036486587.4922	205839378516292464984768852787 2.0000	453712073356609.6875	-427598761594910084941543018332 16.0000	1.0131	218645522523224.4375	0.3600

Figure 7.6: *Table with the evaluation of the regression models, target variable: 683 (Offenses Against the Person), input variables: all the 529 OCM codes, file used: homicide_contain_p_encoding.csv*

Outcome of the best models for target '684' (which is the OCM code for Sex and Marital Offenses). As can be seen in the table, Bayesian Ridge is the best model to do analyses with and to predict values with in the future.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
br	Bayesian Ridge	0.0821	0.0392	0.1965	0.0501	0.1377	0.8909	0.6240
lightgbm	Light Gradient Boosting Machine	0.0817	0.0397	0.1977	0.0399	0.1388	0.8823	1.1760
omp	Orthogonal Matching Pursuit	0.0823	0.0395	0.1973	0.0398	0.1393	0.8650	0.1310
gbr	Gradient Boosting Regressor	0.0762	0.0396	0.1974	0.0393	0.1379	0.8793	2.4230
ada	AdaBoost Regressor	0.0768	0.0408	0.2001	0.0183	0.1398	0.8990	0.4890
ridge	Ridge Regression	0.0900	0.0407	0.2004	0.0037	0.1419	0.8616	0.2060
llar	Lasso Least Angle Regression	0.0842	0.0421	0.2030	-0.0063	0.1415	0.9560	0.1270
dummy	Dummy Regressor	0.0842	0.0421	0.2030	-0.0063	0.1415	0.9560	0.0670
en	Elastic Net	0.0842	0.0421	0.2030	-0.0063	0.1415	0.9560	0.1470
lasso	Lasso Regression	0.0842	0.0421	0.2030	-0.0063	0.1415	0.9560	0.1410
xgboost	Extreme Gradient Boosting	0.0737	0.0415	0.2022	-0.0096	0.1417	0.8385	0.8990
knn	K Neighbors Regressor	0.0635	0.0427	0.2049	-0.0335	0.1460	0.8916	0.1540
huber	Huber Regressor	0.0441	0.0439	0.2072	-0.0445	0.1436	0.9992	1.8060
rf	Random Forest Regressor	0.0787	0.0445	0.2100	-0.1041	0.1515	0.7887	19.2730
et	Extra Trees Regressor	0.0750	0.0538	0.2309	-0.3483	0.1635	0.7526	26.9860
dt	Decision Tree Regressor	0.0753	0.0580	0.2397	-0.4530	0.1683	0.7609	0.3780
par	Passive Aggressive Regressor	0.2127	0.0990	0.3116	-1.4757	0.2232	0.9486	0.2090
lr	Linear Regression	7127040641.4406	22878493637410241380352.0000	81171487132.1332	-4955067534752672749977 60.0000	2.2628	44747174105.0394	0.3440
lar	Least Angle Regression	275951747068 70.3047	9656651425129543343886806548 48.0000	310830814685514.3750	-3075261341481553312535928425 6768.0000	0.8161	3.2680	0.3600

Figure 7.7: *Table with the evaluation of the regression models, target variable: 684 (Sex and Marital Offenses), input variables: all the 529 OCM codes, file used: homicide_contain_p_encoding.csv*

Outcome of the best models for target '685' (which is the OCM code for Property Offenses). As can be seen in the table, Bayesian Ridge is the best model to do analyses with and to predict values with in the future.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
br	Bayesian Ridge	0.0719	0.0354	0.1872	0.0276	0.1308	0.9278	0.5370
lightgbm	Light Gradient Boosting Machine	0.0722	0.0359	0.1887	0.0112	0.1325	0.9232	0.8750
gbr	Gradient Boosting Regressor	0.0689	0.0363	0.1894	0.0056	0.1324	0.9250	2.6820
llar	Lasso Least Angle Regression	0.0730	0.0365	0.1901	-0.0021	0.1323	0.9621	0.2140
dummy	Dummy Regressor	0.0730	0.0365	0.1901	-0.0021	0.1323	0.9621	0.0720
en	Elastic Net	0.0730	0.0365	0.1901	-0.0021	0.1323	0.9621	0.0990
lasso	Lasso Regression	0.0730	0.0365	0.1901	-0.0021	0.1323	0.9621	0.1000
omp	Orthogonal Matching Pursuit	0.0743	0.0365	0.1902	-0.0037	0.1340	0.9187	0.1810
ada	AdaBoost Regressor	0.0688	0.0375	0.1927	-0.0288	0.1339	0.9475	0.4570
huber	Huber Regressor	0.0379	0.0379	0.1937	-0.0393	0.1343	0.9999	1.7830
ridge	Ridge Regression	0.0822	0.0380	0.1941	-0.0456	0.1374	0.9191	0.1360
xgboost	Extreme Gradient Boosting	0.0681	0.0395	0.1977	-0.0849	0.1396	0.9159	0.7160
knn	K Neighbors Regressor	0.0906	0.0446	0.2107	-0.2410	0.1572	0.9133	0.1960
rf	Random Forest Regressor	0.0778	0.0456	0.2129	-0.2675	0.1551	0.8929	21.3090
et	Extra Trees Regressor	0.0766	0.0570	0.2377	-0.5805	0.1692	0.8745	30.0160
dt	Decision Tree Regressor	0.0768	0.0613	0.2466	-0.7066	0.1738	0.8716	0.6930
par	Passive Aggressive Regressor	0.1799	0.0805	0.2811	-1.2361	0.1986	0.9930	0.1310
lr	Linear Regression	10080635894.2186	395943296945892160960256.0000	108383845517.2589	-13092392973882493270753328.0000	2.1612	66792256310.5362	0.3080
lar	Least Angle Regression	1465224621097747.0000	272488784577383967249955355308851 2.0000	16507236257280992.0000	-98759508802189505829673740091260928.0000	1.5009	2322714344882775 2.0000	0.4980

Figure 7.8: *Table with the evaluation of the regression models, target variable: 685 (Property Offenses), input variables: all the 529 OCM codes, file used: homicide_contain_p_encoding.csv*

Outcome of the best models for target '728' (which is the OCM code for Peacemaking). As can be seen in the table, K Neighbours Regressor is the best model to do analyses with and to predict values with in the future.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
knn	K Neighbors Regressor	0.0081	0.0057	0.0728	0.0108	0.0519	nan	0.1510
br	Bayesian Ridge	0.0122	0.0056	0.0711	0.0060	0.0497	nan	0.6940
llar	Lasso Least Angle Regression	0.0113	0.0056	0.0710	-0.0007	0.0494	nan	0.1340
dummy	Dummy Regressor	0.0113	0.0056	0.0710	-0.0007	0.0494	nan	0.0640
en	Elastic Net	0.0113	0.0056	0.0710	-0.0007	0.0494	nan	0.0920
lasso	Lasso Regression	0.0113	0.0056	0.0710	-0.0007	0.0494	nan	0.0910
huber	Huber Regressor	0.0057	0.0057	0.0706	-0.0057	0.0489	nan	1.7330
lightgbm	Light Gradient Boosting Machine	0.0128	0.0057	0.0723	-0.0064	0.0512	nan	0.7360
ada	AdaBoost Regressor	0.0105	0.0059	0.0733	-0.0439	0.0516	nan	0.5150
ridge	Ridge Regression	0.0196	0.0060	0.0748	-0.0518	0.0542	nan	0.1340
omp	Orthogonal Matching Pursuit	0.0168	0.0060	0.0746	-0.0629	0.0536	nan	0.2080
gbr	Gradient Boosting Regressor	0.0105	0.0061	0.0746	-0.0697	0.0529	nan	2.6840
rf	Random Forest Regressor	0.0126	0.0070	0.0820	-0.1826	0.0608	nan	9.8190
xgboost	Extreme Gradient Boosting	0.0118	0.0076	0.0845	-0.2640	0.0607	nan	0.6910
et	Extra Trees Regressor	0.0116	0.0085	0.0898	-0.4080	0.0647	nan	14.1740
dt	Decision Tree Regressor	0.0123	0.0098	0.0967	-0.6387	0.0688	nan	0.2330
par	Passive Aggressive Regressor	0.0719	0.0138	0.1169	-1.2748	0.0945	nan	0.1150
lar	Least Angle Regression	1072732230.2926	73785448305383440384 0.0000	11909260450.0797	-28114571278227204997120.0000	0.5390	nan	0.3700
lr	Linear Regression	2464769083.8397	24269856035894816931 84.0000	26002659463.1979	-30389840631074137099468 8.0000	2.1644	nan	0.4050

Figure 7.9: *Table with the evaluation of the regression models, target variable: 728 (Peacemaking), input variables: all the 529 OCM codes, file used: homicide_contain_p_encoding.csv*

Outcome of the best models for target '754' (which is the OCM code for Sorcery). As can be seen in the table, Light Gradient Boosting Machine is the best model to do analyses with and to predict values with in the future.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.1061	0.0505	0.2244	0.1263	0.1574	0.8106	0.7270
xgboost	Extreme Gradient Boosting	0.0974	0.0511	0.2258	0.1142	0.1585	0.7539	0.6830
gbr	Gradient Boosting Regressor	0.1027	0.0519	0.2276	0.1011	0.1581	0.8395	2.5920
br	Bayesian Ridge	0.1119	0.0520	0.2279	0.0993	0.1593	0.8462	0.3480
omp	Orthogonal Matching Pursuit	0.1104	0.0524	0.2288	0.0927	0.1607	0.8270	0.0910
ridge	Ridge Regression	0.1200	0.0543	0.2328	0.0594	0.1640	0.8158	0.0890
ada	AdaBoost Regressor	0.1072	0.0555	0.2353	0.0406	0.1636	0.8994	0.2750
rf	Random Forest Regressor	0.1065	0.0571	0.2385	0.0089	0.1743	0.6287	17.1970
llar	Lasso Least Angle Regression	0.1157	0.0578	0.2403	-0.0006	0.1675	0.9384	0.1400
en	Elastic Net	0.1157	0.0578	0.2403	-0.0006	0.1675	0.9384	0.0660
lasso	Lasso Regression	0.1157	0.0578	0.2403	-0.0006	0.1675	0.9384	0.0640
dummy	Dummy Regressor	0.1157	0.0578	0.2403	-0.0006	0.1675	0.9384	0.0510
huber	Huber Regressor	0.0617	0.0615	0.2477	-0.0626	0.1716	0.9985	0.6830
et	Extra Trees Regressor	0.0996	0.0692	0.2626	-0.1999	0.1871	0.6083	24.3190
dt	Decision Tree Regressor	0.0999	0.0742	0.2720	-0.2872	0.1924	0.6039	0.3450
knn	K Neighbors Regressor	0.1456	0.0811	0.2842	-0.4019	0.2138	0.6379	0.1720
par	Passive Aggressive Regressor	0.2428	0.1180	0.3433	-1.0502	0.2500	0.8118	0.1010
lar	Least Angle Regression	6222492139.1699	477216702785733986877744.0000	69080873223.4531	-8098522719967176493629440000	0.9071	34.1668	0.3450
lr	Linear Regression	26362688432.4526	4028932031731018036674560000	282759180331.8274	-7216369175251226984448000.0000	2.3864	27405504465.5487	0.2050

Figure 7.10: *Table with the evaluation of the regression models, target variable: 754 (Sorcery), input variables: all the 529 OCM codes, file used: homicide_contain_p_encoding.csv*

Outcome of the best models for target 'Drugs and Alcohol'. As can be seen in the table, Orthogonal Matching Pursuit is the best model to do analyses with and to predict values with in the future.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
omp	Orthogonal Matching Pursuit	0.0364	0.0179	0.1332	0.0084	0.0930	0.9619	0.0500
br	Bayesian Ridge	0.0374	0.0179	0.1333	0.0064	0.0930	0.9650	0.0780
llar	Lasso Least Angle Regression	0.0362	0.0181	0.1339	-0.0010	0.0930	0.9815	0.0460
dummy	Dummy Regressor	0.0362	0.0181	0.1339	-0.0010	0.0930	0.9815	0.0470
en	Elastic Net	0.0362	0.0181	0.1339	-0.0010	0.0930	0.9815	0.0430
lasso	Lasso Regression	0.0362	0.0181	0.1339	-0.0010	0.0930	0.9815	0.0410
lightgbm	Light Gradient Boosting Machine	0.0393	0.0181	0.1341	-0.0051	0.0941	0.9518	0.7180
ridge	Ridge Regression	0.0404	0.0184	0.1349	-0.0183	0.0948	0.9592	0.0430
huber	Huber Regressor	0.0185	0.0184	0.1351	-0.0186	0.0937	0.9999	0.6540
gbr	Gradient Boosting Regressor	0.0352	0.0186	0.1356	-0.0285	0.0948	0.9489	0.6420
lar	Least Angle Regression	0.0409	0.0185	0.1356	-0.0292	0.0953	0.9588	0.0550
ada	AdaBoost Regressor	0.0362	0.0191	0.1374	-0.0630	0.0961	0.9710	0.1480
knn	K Neighbors Regressor	0.0277	0.0195	0.1392	-0.0886	0.0996	0.9602	0.1280
xgboost	Extreme Gradient Boosting	0.0375	0.0205	0.1424	-0.1378	0.1005	0.8914	0.3850
rf	Random Forest Regressor	0.0416	0.0224	0.1492	-0.2571	0.1096	0.8942	2.5870
et	Extra Trees Regressor	0.0402	0.0285	0.1685	-0.6254	0.1207	0.8822	3.7910
dt	Decision Tree Regressor	0.0405	0.0306	0.1745	-0.7446	0.1241	0.8839	0.1300
par	Passive Aggressive Regressor	0.1274	0.0438	0.2056	-1.4484	0.1521	1.0116	0.0620
lr	Linear Regression	109499492.7375	59111384781090037760.0000	2431283298.7349	-42230931686240912670720.0000	0.2082	0.9589	0.0650

Figure 7.11: *Table with the evaluation of the regression models, target variable: Drugs/Alcohol, input variables: the 93 self-chosen groups of OCM codes, file used: combined_homicide_data_encoding.csv*

Outcome of the best models for target 'War and Peacemaking'. As can be seen in the table, Bayesian Ridge is the best model to do analyses with and to predict values with in the future.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
br	Bayesian Ridge	0.0825	0.0410	0.2016	0.0270	0.1408	0.9243	0.0410
omp	Orthogonal Matching Pursuit	0.0823	0.0413	0.2025	0.0189	0.1416	0.9278	0.0250
ridge	Ridge Regression	0.0846	0.0414	0.2026	0.0178	0.1423	0.9126	0.0250
gbr	Gradient Boosting Regressor	0.0809	0.0413	0.2025	0.0169	0.1417	0.9179	0.4650
lightgbm	Light Gradient Boosting Machine	0.0842	0.0415	0.2029	0.0137	0.1426	0.9101	0.4390
lar	Least Angle Regression	0.0851	0.0417	0.2034	0.0096	0.1429	0.9137	0.0340
llar	Lasso Least Angle Regression	0.0845	0.0423	0.2047	-0.0022	0.1426	0.9558	0.0250
dummy	Dummy Regressor	0.0845	0.0423	0.2047	-0.0022	0.1426	0.9558	0.0270
en	Elastic Net	0.0845	0.0423	0.2047	-0.0022	0.1426	0.9558	0.0240
lasso	Lasso Regression	0.0845	0.0423	0.2047	-0.0022	0.1426	0.9558	0.0270
huber	Huber Regressor	0.0443	0.0442	0.2092	-0.0459	0.1450	0.9998	0.1190
xgboost	Extreme Gradient Boosting	0.0848	0.0467	0.2158	-0.1232	0.1513	0.8980	0.1570
knn	K Neighbors Regressor	0.0854	0.0470	0.2162	-0.1249	0.1580	0.8975	0.0400
rf	Random Forest Regressor	0.0852	0.0477	0.2180	-0.1483	0.1575	0.8636	1.5980
ada	AdaBoost Regressor	0.1034	0.0471	0.2149	-0.1524	0.1539	0.9139	0.1570
et	Extra Trees Regressor	0.0843	0.0614	0.2474	-0.4884	0.1754	0.8710	2.0670
dt	Decision Tree Regressor	0.0836	0.0657	0.2558	-0.6007	0.1795	0.8751	0.0430
par	Passive Aggressive Regressor	0.2424	0.1096	0.3285	-1.7469	0.2455	0.8074	0.0310
lr	Linear Regression	4021988366.7513	797496047784263271055360000	89302634215.7693	-215223869551318432061849600	0.2678	0.9138	0.4330

Figure 7.12: *Table with the evaluation of the regression models, target variable: War/Peacemaking, input variables: the self-chosen 93 groups of OCM codes, file used: combined_homicide_data_encoding.csv*

Tuning models

Below the results of all the models that were tuned can be found. As said before, Gradient Boosting Regressor and Light Gradient Boosting Machine did not provide any improvement when tuning the model, so these results are not shown.

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.1201	0.0565	0.2376	0.2046	0.1674	0.6941
1	0.1169	0.0528	0.2299	0.1014	0.1640	0.7248
2	0.1277	0.0652	0.2553	0.1053	0.1775	0.7955
3	0.1160	0.0520	0.2280	0.1677	0.1605	0.7566
4	0.1348	0.0692	0.2630	0.1492	0.1811	0.7816
5	0.1318	0.0706	0.2658	0.1310	0.1841	0.7835
6	0.1236	0.0623	0.2495	0.0801	0.1748	0.7877
7	0.1240	0.0639	0.2528	0.1017	0.1758	0.8342
8	0.1223	0.0597	0.2443	0.0445	0.1716	0.8412
9	0.1224	0.0537	0.2317	-0.0745	0.1683	0.7511
Mean	0.1239	0.0606	0.2458	0.1011	0.1725	0.7750
Std	0.0057	0.0064	0.0130	0.0727	0.0071	0.0431

Figure 7.13: *The Extreme Gradient Boosting model with target 672 (Liability) before tuning*

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.1506	0.0604	0.2457	0.1495	0.1776	0.7321
1	0.1482	0.0550	0.2344	0.0656	0.1725	0.7406
2	0.1504	0.0662	0.2572	0.0916	0.1805	0.8041
3	0.1437	0.0551	0.2346	0.1185	0.1700	0.7613
4	0.1611	0.0730	0.2702	0.1018	0.1908	0.8000
5	0.1614	0.0708	0.2662	0.1284	0.1886	0.7790
6	0.1487	0.0605	0.2460	0.1057	0.1766	0.7823
7	0.1609	0.0680	0.2608	0.0442	0.1875	0.8070
8	0.1555	0.0628	0.2507	-0.0060	0.1831	0.8095
9	0.1523	0.0549	0.2343	-0.0986	0.1757	0.7832
Mean	0.1533	0.0627	0.2500	0.0701	0.1803	0.7799
Std	0.0059	0.0063	0.0126	0.0705	0.0067	0.0261

Figure 7.14: *The Extreme Gradient Boosting model with target 672 (Liability) after tuning, parameter tuned: 'R^2'*

For target 728, Peacemaking, only the used model for the feature importance plot (Bayesian Rigde) will be shown, as the other model (KNN) was not used in the analyses.

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.0629	0.0275	0.1658	0.0014	0.1163	0.9411
1	0.0749	0.0392	0.1980	0.0367	0.1376	0.9307
2	0.0802	0.0448	0.2116	0.0335	0.1464	0.9331
3	0.0703	0.0325	0.1804	0.0229	0.1265	0.9278
4	0.0653	0.0266	0.1630	0.0367	0.1150	0.9140
5	0.0602	0.0255	0.1595	0.0087	0.1127	0.9361
6	0.0733	0.0386	0.1965	0.0532	0.1359	0.9264
7	0.0705	0.0321	0.1793	0.0346	0.1257	0.9195
8	0.0748	0.0384	0.1960	0.0582	0.1356	0.9213
9	0.0700	0.0351	0.1875	0.0516	0.1301	0.9246
Mean	0.0702	0.0340	0.1838	0.0337	0.1282	0.9275
Std	0.0057	0.0060	0.0163	0.0176	0.0105	0.0077

Figure 7.15: *The Bayesian Ridge model with target 681 (Sanctions) before tuning*

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.0626	0.0275	0.1658	0.0014	0.1162	0.9415
1	0.0746	0.0392	0.1980	0.0364	0.1376	0.9310
2	0.0800	0.0448	0.2116	0.0332	0.1464	0.9334
3	0.0700	0.0325	0.1804	0.0229	0.1265	0.9281
4	0.0650	0.0266	0.1630	0.0370	0.1149	0.9142
5	0.0599	0.0255	0.1595	0.0086	0.1126	0.9366
6	0.0730	0.0386	0.1965	0.0533	0.1359	0.9265
7	0.0703	0.0321	0.1793	0.0347	0.1257	0.9196
8	0.0745	0.0384	0.1959	0.0587	0.1355	0.9213
9	0.0697	0.0351	0.1875	0.0516	0.1301	0.9248
Mean	0.0699	0.0340	0.1838	0.0338	0.1281	0.9277
Std	0.0058	0.0060	0.0164	0.0177	0.0105	0.0078

Figure 7.16: *The Bayesian Ridge model with target 681 (Sanctions) after tuning, parameter tuned: 'R^2'*

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.0931	0.0519	0.2277	0.0910	0.1572	0.8809
1	0.0900	0.0445	0.2110	0.0369	0.1485	0.8971
2	0.0667	0.0216	0.1468	-0.0844	0.1083	0.9082
3	0.0868	0.0458	0.2141	0.1446	0.1464	0.8629
4	0.0714	0.0312	0.1767	0.0052	0.1251	0.9122
5	0.0803	0.0384	0.1960	0.0132	0.1368	0.9163
6	0.0751	0.0315	0.1775	0.0533	0.1257	0.8829
7	0.0971	0.0522	0.2285	0.0862	0.1579	0.8831
8	0.0829	0.0429	0.2071	0.0742	0.1436	0.9017
9	0.0781	0.0323	0.1798	0.0809	0.1272	0.8638
Mean	0.0821	0.0392	0.1965	0.0501	0.1377	0.8909
Std	0.0092	0.0095	0.0248	0.0592	0.0152	0.0181

Figure 7.17: *The Bayesian Ridge model with target 684 (Sex and Marital Offenses) before tuning*

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.0930	0.0519	0.2277	0.0909	0.1572	0.8811
1	0.0898	0.0445	0.2110	0.0370	0.1485	0.8972
2	0.0666	0.0215	0.1468	-0.0841	0.1082	0.9085
3	0.0866	0.0458	0.2141	0.1443	0.1464	0.8632
4	0.0713	0.0312	0.1767	0.0052	0.1251	0.9124
5	0.0801	0.0384	0.1960	0.0133	0.1368	0.9165
6	0.0749	0.0315	0.1775	0.0534	0.1257	0.8831
7	0.0970	0.0522	0.2285	0.0861	0.1579	0.8833
8	0.0827	0.0429	0.2070	0.0743	0.1436	0.9017
9	0.0780	0.0323	0.1798	0.0809	0.1272	0.8641
Mean	0.0820	0.0392	0.1965	0.0501	0.1377	0.8911
Std	0.0092	0.0095	0.0248	0.0591	0.0152	0.0181

Figure 7.18: *The Bayesian Ridge model with target 684 (Sex and Marital Offenses) after tuning, parameter tuned: 'R^2'*

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.0771	0.0393	0.1984	0.0333	0.1383	0.9221
1	0.0849	0.0485	0.2203	0.0264	0.1530	0.9293
2	0.0699	0.0354	0.1881	-0.0055	0.1319	0.9519
3	0.0767	0.0404	0.2011	0.0518	0.1392	0.9233
4	0.0619	0.0269	0.1642	0.0233	0.1147	0.9296
5	0.0693	0.0333	0.1826	0.0522	0.1272	0.9175
6	0.0662	0.0287	0.1693	0.0279	0.1190	0.9251
7	0.0819	0.0437	0.2091	0.0169	0.1458	0.9333
8	0.0664	0.0284	0.1685	0.0370	0.1192	0.9112
9	0.0648	0.0291	0.1707	0.0125	0.1198	0.9352
Mean	0.0719	0.0354	0.1872	0.0276	0.1308	0.9278
Std	0.0074	0.0070	0.0184	0.0166	0.0123	0.0106

Figure 7.19: *The Bayesian Ridge model with target 685 (Property Offenses) before tuning*

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.0769	0.0393	0.1984	0.0333	0.1383	0.9223
1	0.0847	0.0485	0.2203	0.0264	0.1529	0.9294
2	0.0697	0.0354	0.1881	-0.0057	0.1319	0.9522
3	0.0765	0.0404	0.2011	0.0518	0.1392	0.9235
4	0.0617	0.0269	0.1641	0.0236	0.1146	0.9297
5	0.0691	0.0333	0.1826	0.0524	0.1272	0.9176
6	0.0661	0.0287	0.1693	0.0280	0.1190	0.9253
7	0.0817	0.0437	0.2091	0.0173	0.1458	0.9333
8	0.0662	0.0284	0.1685	0.0372	0.1191	0.9113
9	0.0646	0.0291	0.1707	0.0127	0.1198	0.9353
Mean	0.0717	0.0354	0.1872	0.0277	0.1308	0.9280
Std	0.0074	0.0070	0.0184	0.0167	0.0123	0.0106

Figure 7.20: *The Bayesian Ridge model with target 685 (Property Offenses) after tuning, parameter tuned: 'R^2'*

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.0105	0.0040	0.0635	-0.0006	0.0443	0.9877
1	0.0105	0.0040	0.0630	0.0164	0.0437	0.9796
2	0.0153	0.0099	0.0994	0.0148	0.0687	0.9852
3	0.0125	0.0060	0.0778	0.0004	0.0541	0.9897
4	0.0074	0.0001	0.0104	0.0000	0.0103	nan
5	0.0127	0.0061	0.0778	-0.0009	0.0542	0.9891
6	0.0145	0.0079	0.0892	0.0124	0.0618	0.9844
7	0.0113	0.0040	0.0635	0.0017	0.0444	0.9830
8	0.0130	0.0060	0.0772	0.0149	0.0536	0.9794
9	0.0146	0.0080	0.0897	0.0004	0.0623	0.9904
Mean	0.0122	0.0056	0.0711	0.0060	0.0497	nan
Std	0.0023	0.0026	0.0234	0.0072	0.0154	nan

Figure 7.21: *The Bayesian Ridge model with target 728 (Peacemaking) before tuning*

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.0106	0.0040	0.0635	-0.0010	0.0443	0.9874
1	0.0106	0.0040	0.0629	0.0178	0.0437	0.9782
2	0.0155	0.0099	0.0993	0.0180	0.0685	0.9832
3	0.0126	0.0060	0.0778	0.0001	0.0541	0.9894
4	0.0075	0.0001	0.0106	0.0000	0.0105	nan
5	0.0127	0.0061	0.0778	-0.0012	0.0542	0.9891
6	0.0147	0.0079	0.0891	0.0129	0.0617	0.9837
7	0.0114	0.0040	0.0635	0.0012	0.0444	0.9827
8	0.0131	0.0060	0.0772	0.0147	0.0536	0.9792
9	0.0147	0.0080	0.0897	0.0000	0.0623	0.9904
Mean	0.0123	0.0056	0.0711	0.0062	0.0497	nan
Std	0.0023	0.0026	0.0233	0.0080	0.0153	nan

Figure 7.22: *The Bayesian Ridge model with target 728 (Peacemaking) after tuning, parameter tuned: 'R^2'*

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.0347	0.0149	0.1219	0.0673	0.0846	0.9303
1	0.0392	0.0210	0.1451	0.0336	0.1006	0.9539
2	0.0329	0.0138	0.1176	0.0126	0.0824	0.9579
3	0.0340	0.0165	0.1284	-0.0334	0.0904	0.9803
4	0.0337	0.0143	0.1196	-0.0223	0.0845	0.9695
5	0.0398	0.0233	0.1527	0.0186	0.1057	0.9649
6	0.0328	0.0143	0.1198	-0.0247	0.0843	0.9762
7	0.0366	0.0176	0.1328	0.0161	0.0926	0.9585
8	0.0392	0.0208	0.1444	0.0447	0.0997	0.9483
9	0.0414	0.0224	0.1498	-0.0291	0.1052	0.9789
Mean	0.0364	0.0179	0.1332	0.0084	0.0930	0.9619
Std	0.0031	0.0035	0.0130	0.0329	0.0087	0.0148

Figure 7.23: *The Orthogonal Matching Pursuit model with target Drugs/Alcohol before tuning*

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.0338	0.0149	0.1220	0.0652	0.0845	0.9347
1	0.0386	0.0210	0.1448	0.0375	0.1003	0.9535
2	0.0327	0.0139	0.1178	0.0083	0.0826	0.9607
3	0.0334	0.0163	0.1278	-0.0237	0.0898	0.9785
4	0.0332	0.0144	0.1198	-0.0257	0.0846	0.9733
5	0.0402	0.0235	0.1534	0.0089	0.1064	0.9697
6	0.0327	0.0142	0.1191	-0.0133	0.0837	0.9719
7	0.0367	0.0177	0.1329	0.0142	0.0927	0.9609
8	0.0388	0.0208	0.1443	0.0455	0.0997	0.9485
9	0.0397	0.0221	0.1488	-0.0149	0.1040	0.9776
Mean	0.0360	0.0179	0.1331	0.0102	0.0928	0.9629
Std	0.0030	0.0035	0.0129	0.0294	0.0086	0.0134

Figure 7.24: *The Orthogonal Matching Pursuit model with target Drugs/Alcohol after tuning, parameter tuned: 'R^2'*

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.0892	0.0500	0.2236	0.0323	0.1548	0.9283
1	0.0833	0.0378	0.1944	0.0270	0.1370	0.9138
2	0.0732	0.0289	0.1701	0.0196	0.1207	0.9124
3	0.0815	0.0385	0.1962	0.0559	0.1369	0.9082
4	0.0783	0.0373	0.1932	-0.0074	0.1360	0.9403
5	0.0712	0.0328	0.1812	0.0136	0.1273	0.9354
6	0.0886	0.0477	0.2185	0.0443	0.1513	0.9213
7	0.0794	0.0379	0.1947	0.0257	0.1362	0.9254
8	0.0904	0.0502	0.2241	0.0295	0.1554	0.9300
9	0.0899	0.0485	0.2202	0.0295	0.1527	0.9278
Mean	0.0825	0.0410	0.2016	0.0270	0.1408	0.9243
Std	0.0067	0.0072	0.0180	0.0161	0.0115	0.0098

Figure 7.25: *The Bayesian Ridge model with target War/Peacemaking before tuning*

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.0890	0.0500	0.2236	0.0321	0.1548	0.9286
1	0.0830	0.0378	0.1944	0.0272	0.1370	0.9142
2	0.0730	0.0289	0.1701	0.0195	0.1207	0.9130
3	0.0812	0.0385	0.1962	0.0558	0.1369	0.9087
4	0.0781	0.0373	0.1932	-0.0072	0.1360	0.9405
5	0.0709	0.0328	0.1812	0.0136	0.1273	0.9358
6	0.0883	0.0477	0.2185	0.0444	0.1513	0.9216
7	0.0792	0.0379	0.1947	0.0262	0.1362	0.9255
8	0.0902	0.0502	0.2241	0.0295	0.1554	0.9303
9	0.0897	0.0485	0.2202	0.0297	0.1526	0.9280
Mean	0.0823	0.0410	0.2016	0.0271	0.1408	0.9246
Std	0.0067	0.0072	0.0180	0.0161	0.0115	0.0097

Figure 7.26: *The Bayesian Ridge model with target War/Peacemaking after tuning, parameter tuned: 'R^2'*

Feature importance plots

The feature importance plots not shown in the main section can be found below.

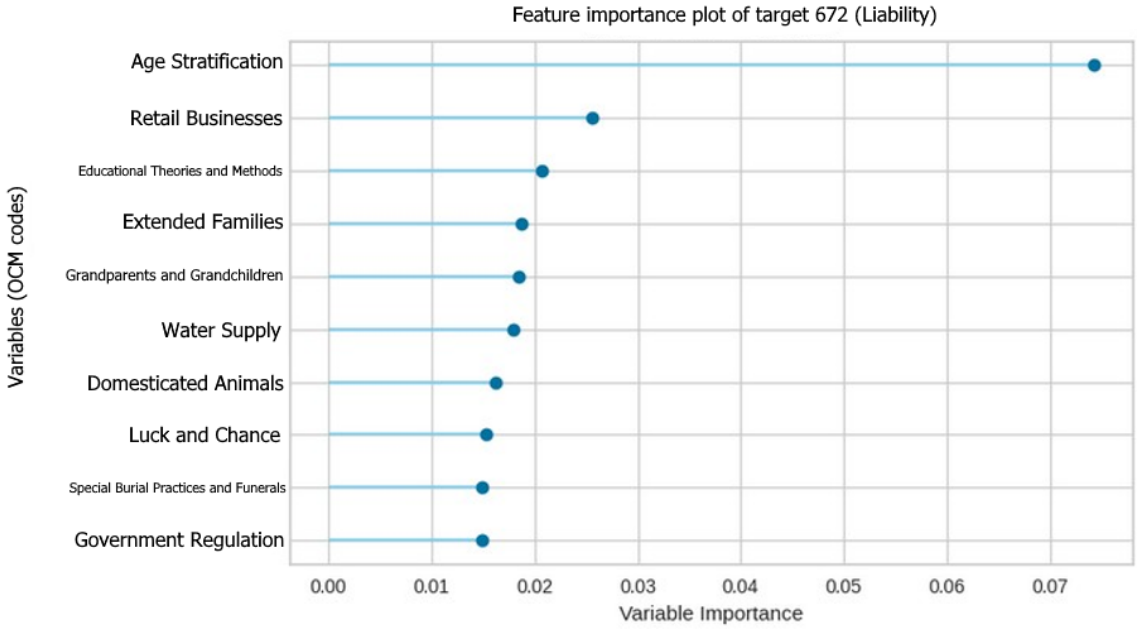

Figure 7.27: *Feature importance plot created with the Extreme Gradient Boosting model, target: 672 (Liability), file: homicide_contain_p_encoding.csv*

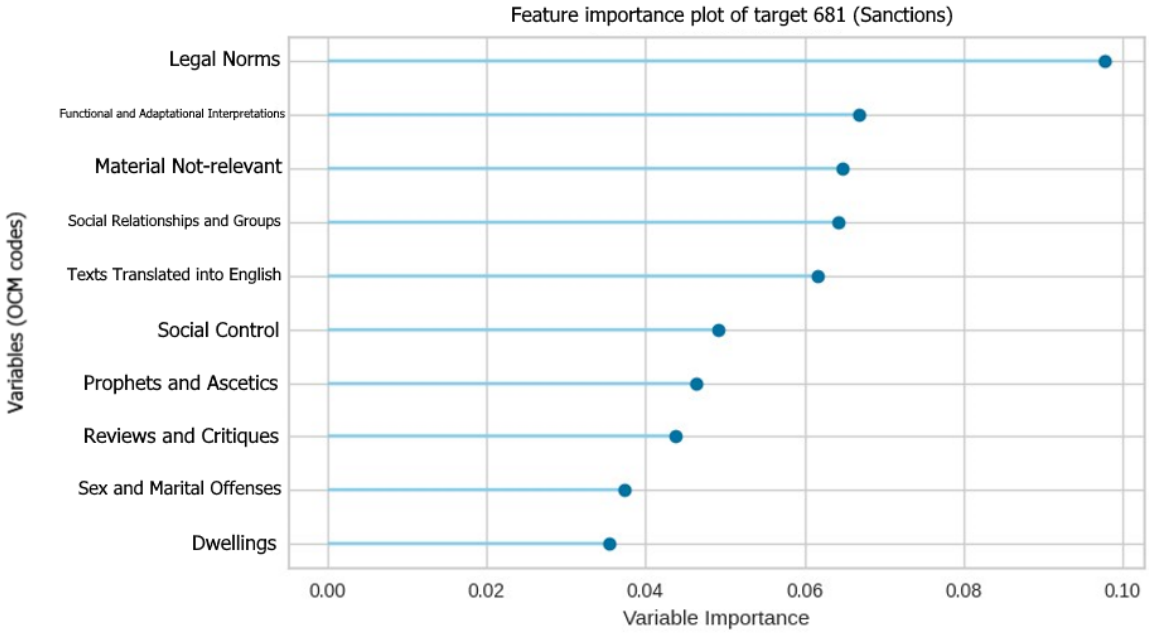

Figure 7.28: *Feature importance plot created with the Bayesian Ridge model, target: 681 (Sanctions), file: homicide_contain_p_encoding.csv*

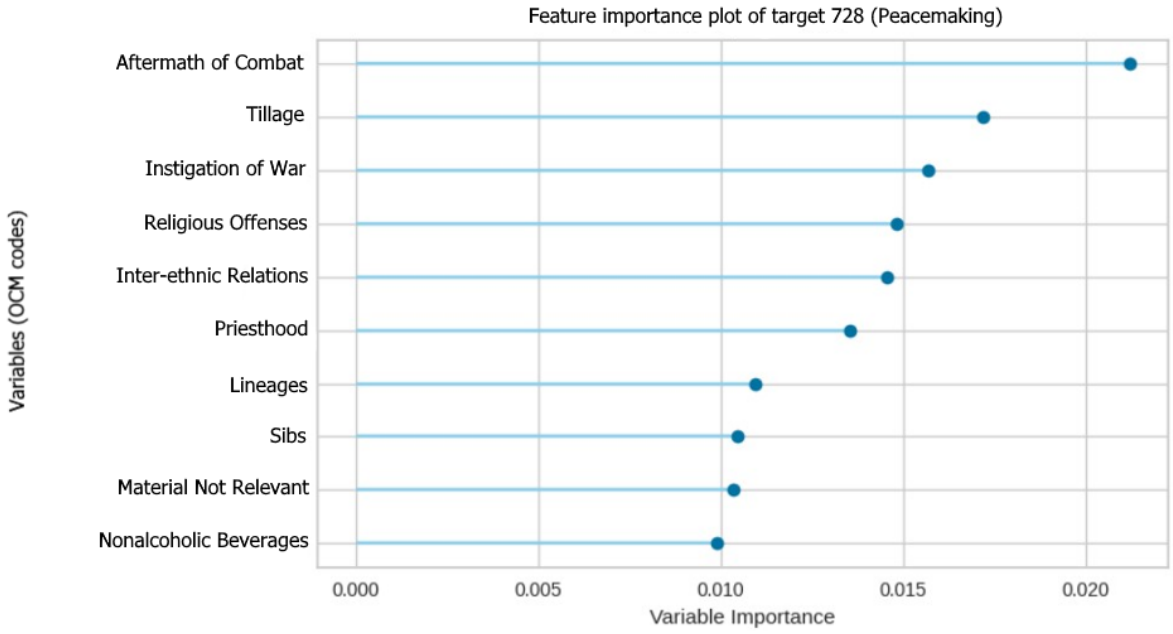

Figure 7.29: *Feature importance plot created with the Bayesian Ridge model, target: 728 (Peacemaking), file: homicide_contain_p_encoding.csv*

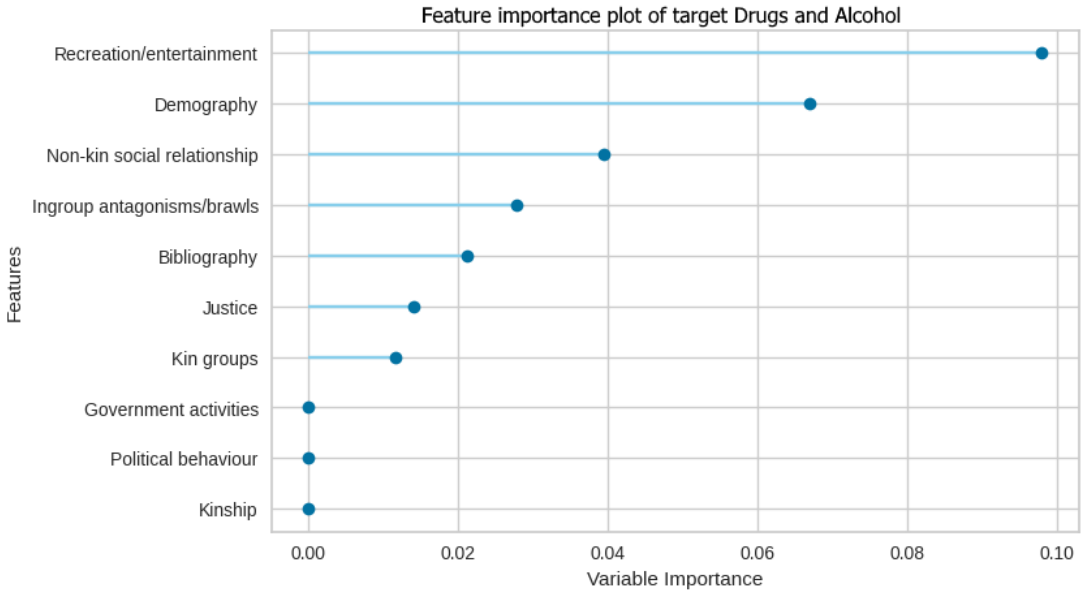

Figure 7.30: *Feature importance plot created with the Orthogonal Matching Pursuit model, target: Drugs and alcohol, file: combined_homicide_data_encoding.csv*

Pictures results

In the figures below the cluster maps found are shown. The pictures can also be found on GitHub.

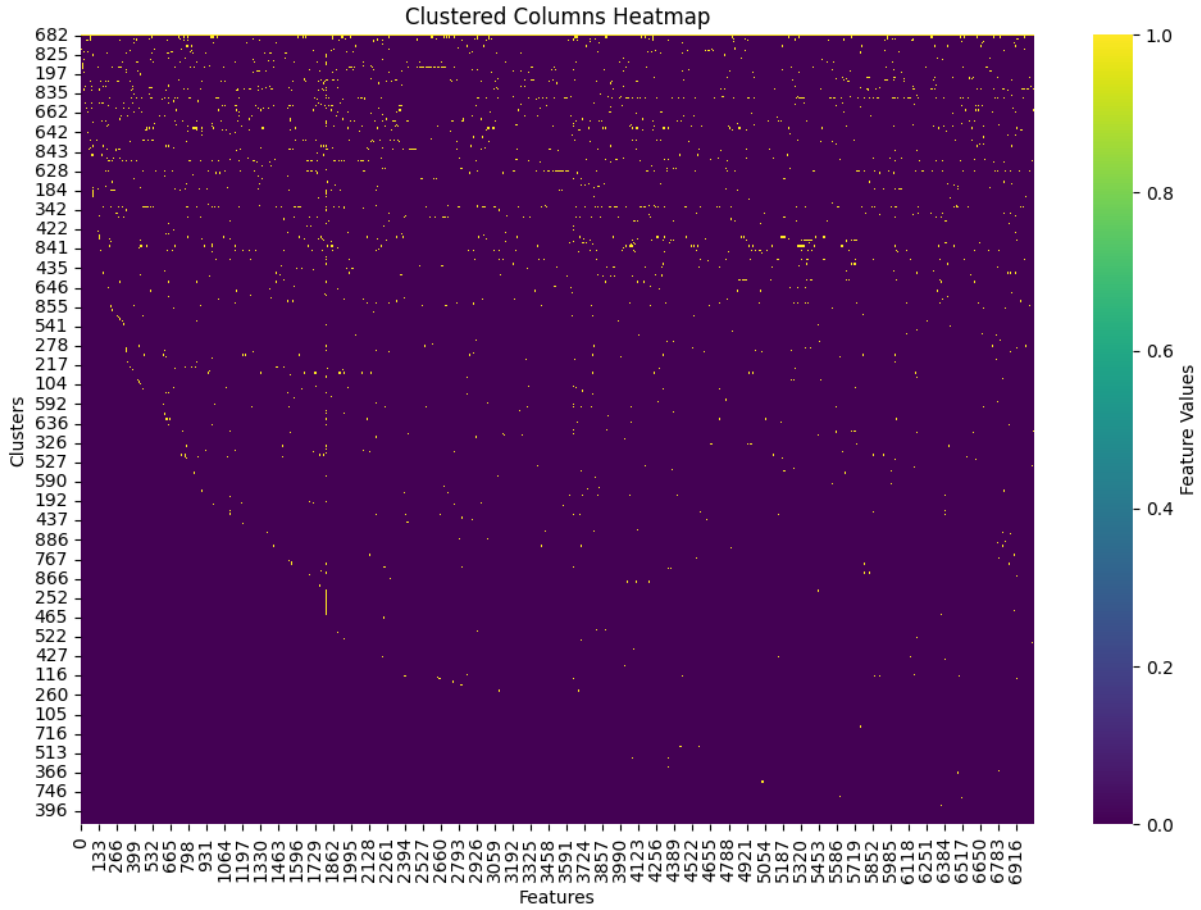

Figure 7.31: *Heatmap with found clusters, data file: homicide_contain_p_encoding.csv*

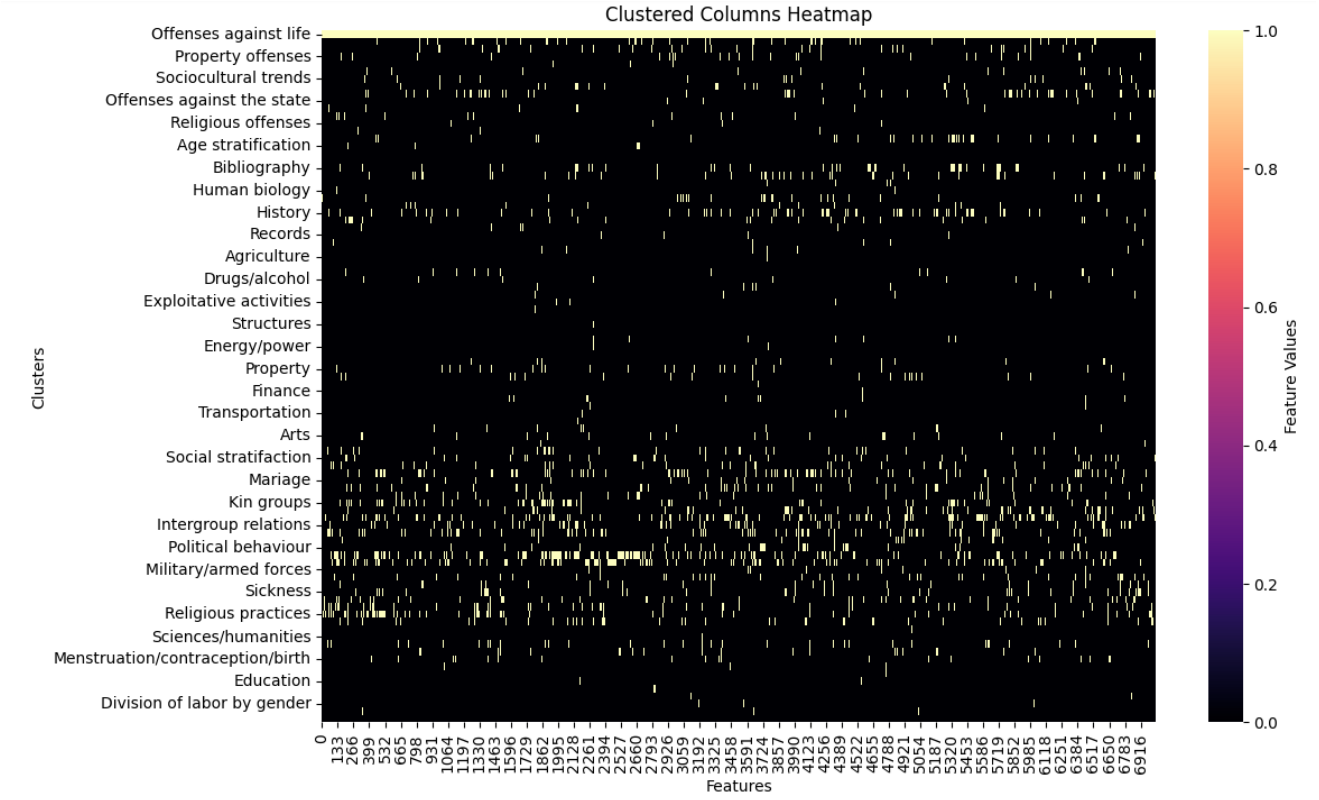

Figure 7.32: *Heatmap with found clusters, data file: combined_homicide_data_encoding.csv*

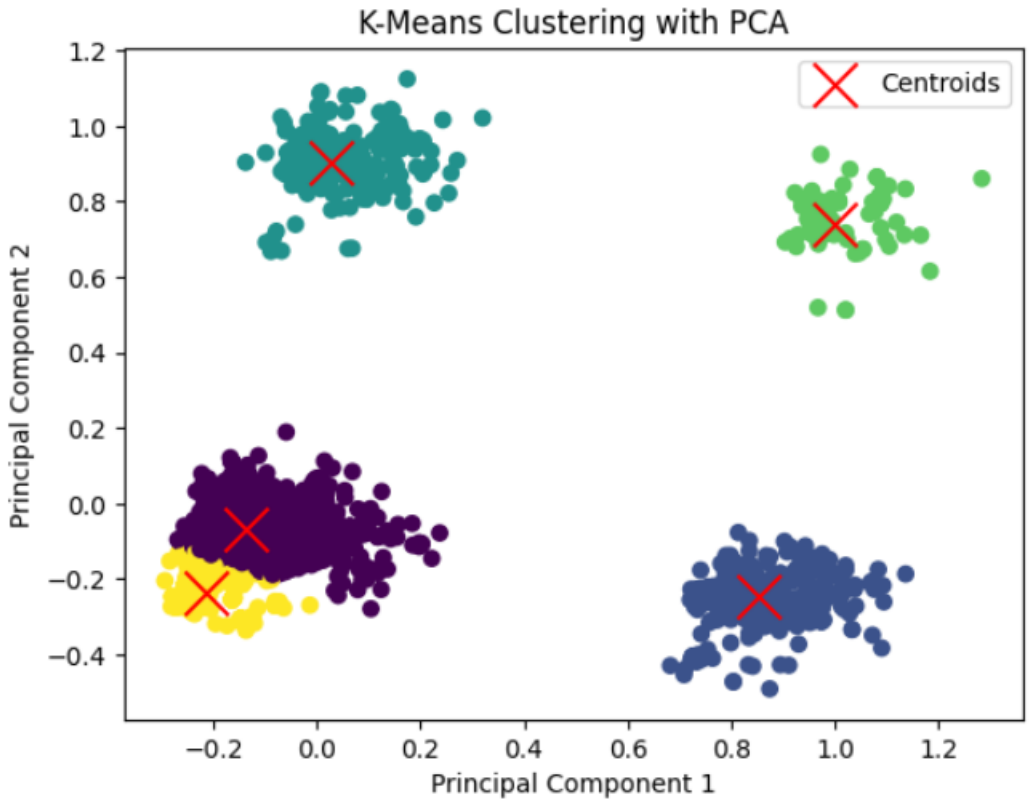

Figure 7.33: *K-means clustering plot with found clusters, data file: homicide_contain_p_encoding.csv*

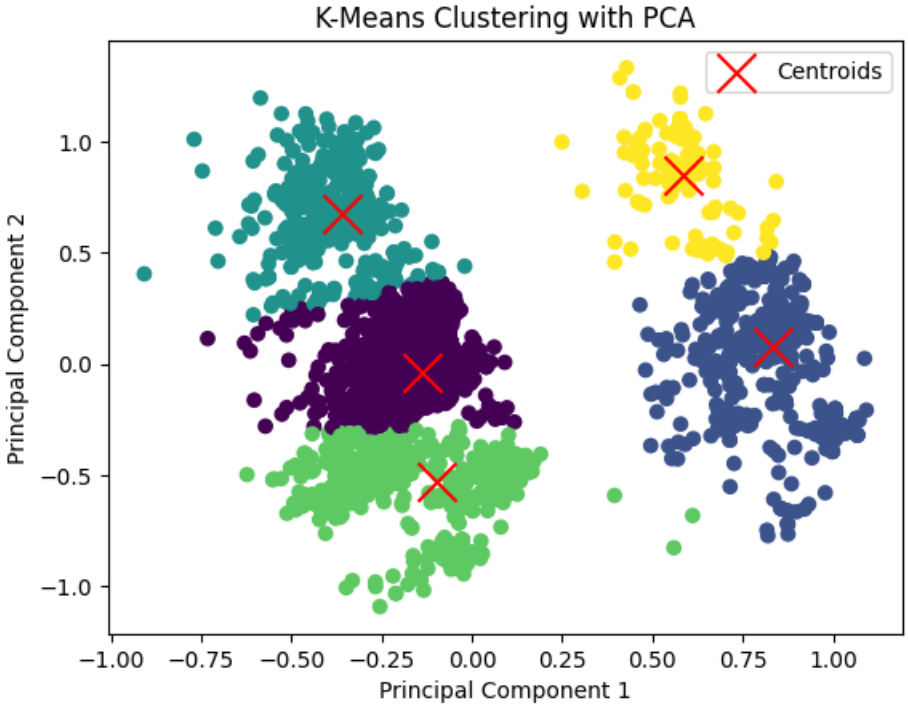

Figure 7.34: *K-means clustering plot with found clusters, data file: combined_homicide_data_encoding.csv*

Experiment

The results of the extra experiment conducted to research the noise within the data can be found in this section.

Outcome of the best models in this experiment with target variable '762' (which is the OCM code for Suicide). As can be seen in the table, Gradient Boosting Regressor is now the best model with an R^2 value of 0,0157.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
gbr	Gradient Boosting Regressor	0.0304	0.0159	0.1239	0.0157	0.0866	0.9496	0.9170
omp	Orthogonal Matching Pursuit	0.0314	0.0159	0.1242	0.0086	0.0868	0.9621	0.2520
br	Bayesian Ridge	0.0323	0.0158	0.1242	0.0074	0.0864	0.9731	0.6480
lightgbm	Light Gradient Boosting Machine	0.0332	0.0159	0.1244	0.0047	0.0873	0.9558	0.7290
ridge	Ridge Regression	0.0360	0.0159	0.1246	0.0019	0.0875	0.9550	0.3450
lr	Linear Regression	0.0363	0.0160	0.1247	0.0005	0.0876	0.9541	0.9940
lar	Least Angle Regression	0.0363	0.0160	0.1247	0.0005	0.0876	0.9541	0.2520
llar	Lasso Least Angle Regression	0.0319	0.0160	0.1248	-0.0024	0.0867	0.9838	0.2630
lasso	Lasso Regression	0.0319	0.0160	0.1248	-0.0024	0.0867	0.9838	0.2300
en	Elastic Net	0.0319	0.0160	0.1248	-0.0024	0.0867	0.9838	0.2990
dummy	Dummy Regressor	0.0319	0.0160	0.1248	-0.0024	0.0867	0.9838	0.2420
huber	Huber Regressor	0.0163	0.0162	0.1257	-0.0159	0.0871	0.9997	0.6760
ada	AdaBoost Regressor	0.0298	0.0162	0.1258	-0.0253	0.0877	0.9718	0.3240
knn	K Neighbors Regressor	0.0265	0.0177	0.1311	-0.1077	0.0944	0.9614	0.3140
rf	Random Forest Regressor	0.0341	0.0191	0.1355	-0.1876	0.0981	0.9178	3.1830
xgboost	Extreme Gradient Boosting	0.0331	0.0190	0.1354	-0.1905	0.0958	0.9175	0.4300
et	Extra Trees Regressor	0.0331	0.0223	0.1465	-0.3913	0.1044	0.8996	5.6160
dt	Decision Tree Regressor	0.0336	0.0238	0.1509	-0.4712	0.1070	0.9121	0.2930
par	Passive Aggressive Regressor	0.1544	0.0522	0.2257	-2.3095	0.1768	0.9538	0.2580

Figure 7.35: *Table with the evaluation of the regression models, target variable: 762 (Suicide), feature selection applied, file used: homicide_contain_p_encoding.csv*

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.0259	0.0120	0.1094	0.1431	0.0753	0.8893
1	0.0242	0.0105	0.1022	0.1288	0.0709	0.8919
2	0.0397	0.0218	0.1477	-0.0978	0.1050	0.9774
3	0.0296	0.0161	0.1268	-0.0067	0.0884	0.9731
4	0.0365	0.0213	0.1459	-0.0709	0.1028	0.9833
5	0.0262	0.0114	0.1069	0.0498	0.0743	0.9397
6	0.0226	0.0080	0.0895	0.0042	0.0631	0.9518
7	0.0334	0.0193	0.1388	0.0311	0.0963	0.9550
8	0.0257	0.0120	0.1095	0.0027	0.0766	0.9568
9	0.0407	0.0264	0.1624	-0.0275	0.1135	0.9775
Mean	0.0304	0.0159	0.1239	0.0157	0.0866	0.9496
Std	0.0063	0.0057	0.0226	0.0731	0.0161	0.0322

Figure 7.36: *The Gradient Boosting Regressor model with target 762 (Suicide) before tuning, results experiment*

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.0289	0.0131	0.1145	0.0620	0.0789	0.9485
1	0.0269	0.0115	0.1073	0.0408	0.0743	0.9593
2	0.0367	0.0195	0.1397	0.0179	0.0970	0.9642
3	0.0311	0.0158	0.1255	0.0128	0.0872	0.9727
4	0.0359	0.0202	0.1420	-0.0148	0.0990	0.9836
5	0.0279	0.0118	0.1087	0.0175	0.0757	0.9686
6	0.0236	0.0077	0.0879	0.0409	0.0612	0.9526
7	0.0348	0.0193	0.1390	0.0275	0.0962	0.9648
8	0.0274	0.0120	0.1096	0.0016	0.0764	0.9762
9	0.0408	0.0257	0.1604	-0.0022	0.1112	0.9784
Mean	0.0314	0.0157	0.1235	0.0204	0.0857	0.9669
Std	0.0051	0.0051	0.0206	0.0219	0.0143	0.0107

Figure 7.37: *The Gradient Boosting Regressor with target 762 (Suicide) after tuning, parameter tuned: 'R^2', results experiment*

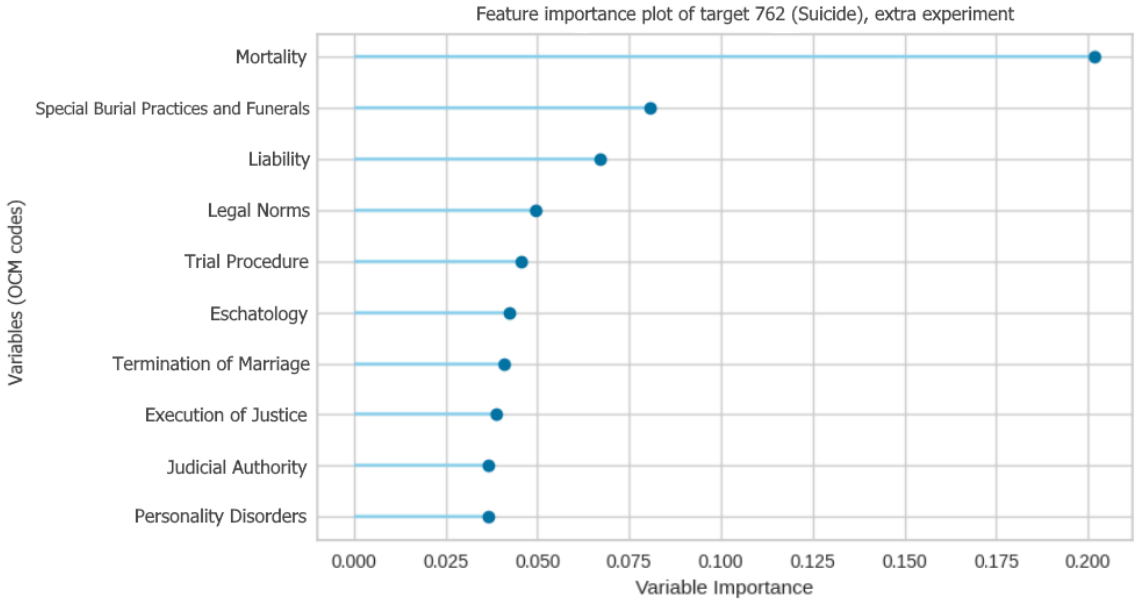

Figure 7.38: *Feature importance plot created with the Gradient Boosting Regressor model, target: 762 (Suicide), feature selection applied, file: homicide_contain_p_encoding.csv*

Acknowledgements

First and foremost I would like to thank my friends and family who all helped me and supported me throughout writing my thesis. Thank you all for helping me with my programs when they stopped working, with helping me fix my code when I could not find the bugs, for checking my spelling when I could not read myself anymore and mainly for your love and support.

Next, I would like to thank my daily supervisor Kristen Syme. She was there to answer all my questions, to help me with finding the right targets and she was the one who had enough confidence in me to help me go to a conference in April 2024. She was also the one who proofread my thesis multiple times and gave me much feedback throughout the whole process for which I am very thankful.

Lastly, I would like to thank my supervisors Marco Spruit and Marieke Liem. After contacting Dr. Liem with an idea for my thesis, she answered me directly and was very supportive and enthusiastic. She then helped me find my second supervisor Marco Spruit, who I already knew from a previous project. Both Dr. Liem and Dr. Spruit had a lot of confidence in me, for that I would like to thank them, as well as thanking them for their guidance and support during my thesis and research.

Bibliography

[1] Ali, M. (2020). *PyCaret: An open source, low-code machine learning library in Python*. PyCaret version 1.0.

[2] Amr, T. (2020). *Hands-On Machine Learning with scikit-learn and Scientific Python Toolkits: A practical guide to implementing supervised and unsupervised machine learning algorithms in Python*. Packt Publishing, Birmingham, 1st edition edition.

[3] Ayers, J. D., Krems, J. A., Hess, N., and Aktipis, A. (2022). Mother-in-law daughter-in-law conflict: an evolutionary perspective and report of empirical data from the usa. *Evolutionary psychological science*, 8(1):56–71.

[4] Bridge, J. A., Ruch, D. A., Sheftall, A. H., Hahm, H. C., O'Keefe, V. M., Fontanella, C. A., Brock, G., Campo, J. V., and Horowitz, L. M. (2023). Youth suicide during the first year of the covid-19 pandemic. *Pediatrics (Evanston)*, 151(3):1–.

[5] Cant, M. A. and Young, A. J. (2013). Resolving social conflict among females without overt aggression. *Philosophical transactions of the Royal Society of London. Series B. Biological sciences*, 368(1631):20130076–20130076.

[6] Droogendyk, L. and Wright, S. C. (2014). Perceptions of interpersonal versus intergroup violence: the case of sexual assault. *PloS one*, 9(11):e112365–e112365.

[7] Erskine-Shaw, M., Monk, R. L., Qureshi, A. W., and Heim, D. (2017). The influence of groups and alcohol consumption on individual risk-taking. *Drug and alcohol dependence*, 179:341–346.

[8] Files, H. R. A. (2023). *eHRAF World Cultures*.

[9] Gaffney, M. R., Adams, K. H., Syme, K. L., and Hagen, E. H. (2022). Depression and suicidality as evolved credible signals of need in social conflicts. *Evolution and human behavior*, 43(3):242–256.

[10] Hagen, E. H., Watson, P. J., and Hammerstein, P. (2008). Gestures of despair and hope: A view on deliberate self-harm from economics and evolutionary biology. *Biological theory*, 3(2):123–138.

[11] Haig, D. (2019). Cooperation and conflict in human pregnancy. *Current biology*, 29(11):R455–R458.

[12] Hamill, M. E., Hernandez, M. C., Bailey, K. R., Cutherell, C. L., Zielinski, M. D., Jenkins, D. H., Naylor, D. F., Matos, M. A., Collier, B. R., and Schiller, H. J. (2023). Legal firearm sales at state level and rates of violent crime, property crime, and homicides. *The Journal of surgical research*, 281:143–154.

[13] Jacoby, T. (2008). *Understanding conflict and violence : theoretical and interdisciplinary approaches*. Routledge, London [etc.

[14] James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). *Linear Regression*, pages 69–134. Springer International Publishing, Cham.

[15] Kanagala, H. K. and Jaya Rama Krishnaiah, V. (2016). A comparative study of k-means, dbscan and optics. In *2016 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–6.

[16] Kennan, J. and Wilson, R. (1993). Bargaining with private information. *Journal of economic literature*, 31(1):45–104.

[17] Kramer, M. D., Krueger, R. F., and Hicks, B. M. (2008). The role of internalizing and externalizing liability factors in accounting for gender differences in the prevalence of common psychopathological syndromes. *Psychological medicine*, 38(1):51–61.

[18] Kriegeskorte, N. and Golan, T. (2019). Neural network models and deep learning - a primer for biologists.

[19] Mathar, R. (2020). *Fundamentals of Data Analytics: With a View to Machine Learning (1st edition)*. Springer Nature, Berlin.

[20] Matsumoto, R., Motomura, E., Shiroyama, T., and Okada, M. (2023). Impact of the japanese government's 'general principles of suicide prevention policy' on youth suicide from 2007 to 2022. *BJPsych open*, 10(1):e16–e16.

[21] McDonald, M. M., Navarrete, C. D., and Van Vugt, M. (2012). Evolution and the psychology of intergroup conflict: the male warrior hypothesis. *Philosophical transactions of the Royal Society of London. Series B. Biological sciences*, 367(1589):670–679.

[22] Mierswa, I., K. R. (2018). *Data science, machine learning, predictive analytics*. RapidMiner Studio (9.1).

[23] Nakagawa, S., Schielzeth, H., and O'Hara, R. B. (2013). A general and simple method for obtaining r2 from generalized linear mixedâeffects models. *Methods in ecology and evolution*, 4(2):133–142.

[24] Nezhad, M. A. S., Khodapanahi, M. K., Yekta, M., Mahmoodikahriz, B., and Ostadghafour, S. (2011). Defense styles in internalizing and externalizing disorders. *Procedia, social and behavioral sciences*, 30:236–241.

[25] Nikstat, A. and Riemann, R. (2020). On the etiology of internalizing and externalizing problem behavior: A twin-family study. *PloS one*, 15(3):e0230626–e0230626.

[26] Nock, M. K., Borges, G., Bromet, E. J., Alonso, J., Angermeyer, M., Beautrais, A., Bruffaerts, R., Chiu, W. T., de Girolamo, G., Gluzman, S., de Graaf, R., Gureje, O., Haro, J. M., Huang, Y., Karam, E., Kessler, R. C., Lepine, J. P., Levinson, D., Medina-Mora, M. E., Ono, Y., Posada-Villa, J., and Williams, D. (2008). Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. *British journal of psychiatry*, 192(2):98–105.

[27] Nunez, E., Steyerberg, E. W., and Nunez, J. (2011). Regression modeling strategies. *Revista Espanola de Cardiologia (English Edition)*, 64(6):501–507.

[28] Ou, C. H. and Hall, W. A. (2018). Anger in the context of postnatal depression: An integrative review. *Birth (Berkeley, Calif.)*, 45(4):336–346.

[29] Parker, G. (2006). Sexual conflict over mating and fertilization: an overview. *Philosophical transactions of the Royal Society of London. Series B. Biological sciences*, 361(1466):235–259.

[30] Parker, G. A., Royle, N. J., and Hartley, I. R. (2002). Intrafamilial conflict and parental investment: a synthesis. *Philosophical transactions of the Royal Society of London. Series B. Biological sciences*, 357(1419):295–307.

[31] Parsons, L., Haque, E., and Liu, H. (2004). Subspace clustering for high dimensional data: a review. *SIGKDD explorations*, 6(1):90–105.

[32] Peacey, S., Campbell, O. L. K., and Mace, R. (2022). Same-sex competition and sexual conflict expressed through witchcraft accusations. *Scientific reports*, 12(1):6655–6655.

[33] Pulay, A. J., Dawson, D. A., Hasin, D. S., Goldstein, R. B., Ruan, J., Pickering, R. P., Huang, B., Chou, P., and Grant, B. F. (2008). Violent behavior and dsm-iv psychiatric disorders: Results from the national epidemiologic survey on alcohol and related conditions. *The journal of clinical psychiatry*, 69(1):12–22.

[34] Rokach, L. and Maimon, O. (2005). *Clustering Methods*, pages 321–352. Springer US, Boston, MA.

[35] Rutar, T. (2023). The prehistory of violence and war: Moving beyond the hobbes-rousseau quagmire. *Journal of Peace Research*, 60(4):720–726.

[36] Sell, A., Bryant, G. A., Cosmides, L., Tooby, J., Sznycer, D., von Rueden, C., Krauss, A., and Gurven, M. (2010). Adaptations in humans for assessing physical strength from the voice. *Proceedings of the Royal Society. B, Biological sciences*, 277(1699):3509–3518.

[37] Sell, A., Cosmides, L., Tooby, J., Sznycer, D., von Rueden, C., and Gurven, M. (2009). Human adaptations for the visual assessment of strength and fighting ability from the body and face. *Proceedings of the Royal Society. B, Biological sciences*, 276(1656):575–584.

[38] Sell, A., Eisner, M., and Ribeaud, D. (2016). Bargaining power and adolescent aggression: the role of fighting ability, coalitional strength, and mate value. *Evolution and human behavior*, 37(2):105–116.

[39] Sell, A., Hone, L. S. E., and Pound, N. (2012). The importance of physical strength to human males. *Human nature (Hawthorne, N.Y.)*, 23(1):30–44.

[40] Senapati, R. E., Jena, S., Parida, J., Panda, A., Patra, P. K., Pati, S., Kaur, H., and Acharya, S. K. (2024). The patterns, trends and major risk factors of suicide among indian adolescents - a scoping review. *BMC psychiatry*, 24(1):35–35.

[41] Skegg, K. (2005). Self-harm. *Lancet*, 366(9495):1471–1483.

[42] Snijders, T. A. B. and Bosker, R. J. R. J. (2012). *Multilevel analysis : an introduction to basic and advanced multilevel modeling*. Sage, Los Angeles, 2nd edition. edition.

[43] Spruit, M., Kais, M., and Menger, V. (2021). Automated business goal extraction from e-mail repositories to bootstrap business understanding. *Future Internet*, 13(10).

[44] Spruit, M., Vroon, R., and Batenburg, R. (2014). Towards healthcare business intelligence in long-term care: An explorative case study in the netherlands. *Computers in Human Behavior*, 30:698–707.

[45] Syme, K. L., Garfield, Z. H., and Hagen, E. H. (2016). Testing the bargaining vs. inclusive fitness models of suicidal behavior against the ethnographic record. *Evolution and Human Behavior*, 37(3):179–192.

[46] Syme, K. L. and Hagen, E. H. (2023). Bargaining and interdependence: Common parent-offspring conflict resolution strategies among chon chuuk and their implications for suicidal behavior. *American anthropologist*, 125(2):262–282.

[47] Trivers, R. L. (1974). Parent-offspring conflict. *American zoologist*, 14(1):249–264.

[48] Urdinez, F. and Cruz, A. (2021). *R for Political Data Science : A Practical Guide*. Chapman Hall/CRC the R Series. Chapman and Hall/CRC, Boca Raton, FL, 1st. edition.

[49] Van Vugt, M. (2009). Sex differences in intergroup competition, aggression, and warfare: the male warrior hypothesis. *Annals of the New York Academy of Sciences*, 1167(1):124–134.

[50] Van Vugt, M., De Cremer, D., and Janssen, D. P. (2007). Gender differences in cooperation and competition: The male-warrior hypothesis. *Psychological science*, 18(1):19–23.

[51] Verhagen-Braspennincx, A., Beijers, G., Janssen, J., and Claes, B. (2024). Incarcerated fathers and their children in the netherlands: Demographic and detention factors that affect the father-child relationship: A secondary analysis. *The Prison journal (Philadelphia, Pa.)*, 104(1):110–130.

[52] Waern, M., Rubenowitz, E., and Wilhelmson, K. (2003). Predictors of suicide in the old elderly. *Gerontology (Basel)*, 49(5):328–334.

[53] Zhang, C. and Zhang, S., editors (2002). *Association Rule*, pages 25–46. Springer Berlin Heidelberg, Berlin, Heidelberg.

[54] Zhang, W., Finy, M. S., Bresin, K., and Verona, E. (2015). Specific patterns of family aggression and adolescents' self- and other-directed harm: The moderating role of personality. *Journal of family violence*, 30(2):161–170.