# Predicting Factors Influencing the Suicides in Sri Lanka

**A dissertation submitted for the Degree of Master of Computer Science**

**D. Samarakkody**
**University of Colombo School of Computing**
**2021**

# DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis. This thesis has also not been submitted for any degree in any university previously.

Student Name: Dhanusha Samarakkody

Registration Number: 2016/MCS/097

Index Number : 16440971

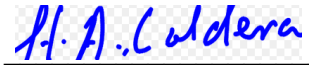
 29/11/2021

Signature of the Student & Date

This is to certify that this thesis is based on the work of Mr. /~~Ms~~. Dhanusha Samarakkody under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by,

Supervisor Name:

 29-11-2021

Signature of the Supervisor & Date

# ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge my supervisor, Dr. H. A. Caldera for his continuous guidance and invaluable support during my research. In addition to that, I am grateful to my beloved wife, Shehari for her unwavering support, encouragement, and understanding throughout the research process. This accomplishment would not have been possible without her. Finally, I would like to extend my gratitude to my family for their support and encouragement.

# ABSTRACT

Suicide is a long-term social issue and a common cause of unnatural death. An individual's suicide risk is usually determined by mental health, but it is also influenced by their background. This research focuses on identifying the factors influencing suicide by scanning the civil, educational, and professional backgrounds of Sri Lankans who have committed suicide from 2014 to 2019. The factors considered in the study are Age Group, Gender, Civil Status, Education Level, Nature of the Occupation, and Reason for Suicide. Initially, the data set is clustered using the k-mode algorithm and identified five clusters that are centered on five different reasons of suicide. Next, the Apriori algorithm is used to identify the associations between the attributes which could lead someone to suicide due to a particular reason. The algorithm is applied for both the entire data set and each cluster. The rules mainly generated around five reasons such as mental disorders, addiction to narcotic drugs, chronic diseases, problems caused with the elders, and harassment by the husband and family disputes. The final evaluation showed that the identified rules are correct with 74%.

**Keywords:** Apriori**,** k-mode clustering, Suicide

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ABBREVIATIONS

| | |
|---|---|
| BIC | Bayesian Information Criterion |
| DM | Data Mining |
| ML | Machine Learning |

# CHAPTER 1: INTRODUCTION

## 1.1 Introduction

Suicide is ranked as one of the most common causes of death around the world. Hence, suicide should be prevented and can be prevented through proper investigations and studies. The research is conducted to identify the factors influencing suicide in Sri Lanka. This chapter gives an overall introduction to the research by describing the motivation, research problem, aim and objectives, and scope of the research.

## 1.2 Motivation

Suicide is the act that intentionally taking one's own life. It is a significant global public health issue that is among the top twenty leading causes of death worldwide. Suicide costs many lives even than malaria, breast cancer, war, or homicide. Every suicide is a tragedy that not only costs millions of lives but also affects many more people through the loss of their loved ones (World Health Organization, 2019).

Suicide is caused or characterized by extreme distress which is an outcome of complex human, socio-economic, and societal factors. It is a sign of serious depression. The World Health Organization (WHO) says that around 800 000 people die every year due to suicide, and every 40 seconds, a person is suicided in the world. It also stated that the third leading cause of death is suicide among 15-19 year old people (WHO, 2019).

Suicide is a serious issue in low- and middle-income countries. Being a lower-middle-income country, Sri Lanka is noted as a higher suicidal country for many decades. The World Population Review website reveals that Sri Lanka has the 29th highest suicide rate which is 14.6 suicides per 100 000 people in 2019 (“Suicide Rate by Country 2020,” 2020). Figure 1 shows the reported suicide incidences per 100000 persons in Sri Lanka from the year 1880 to 2015 (“The Reality of Suicide in Sri Lanka,” 2017).

Figure 1 demonstrates that the overall suicide rate has increased from 2.3 suicides per 100000 persons to 21.2 suicides per 100000 persons from the year 1880 to 1974. After that, there is a considerable increase in the overall suicide rate, 47 suicides per 100000 persons, until the year 1995. Thereafter the suicide rate is started to decrease and in 2006 it is reduced up to 24 suicides per 100000 persons (Thalagala, 2011).

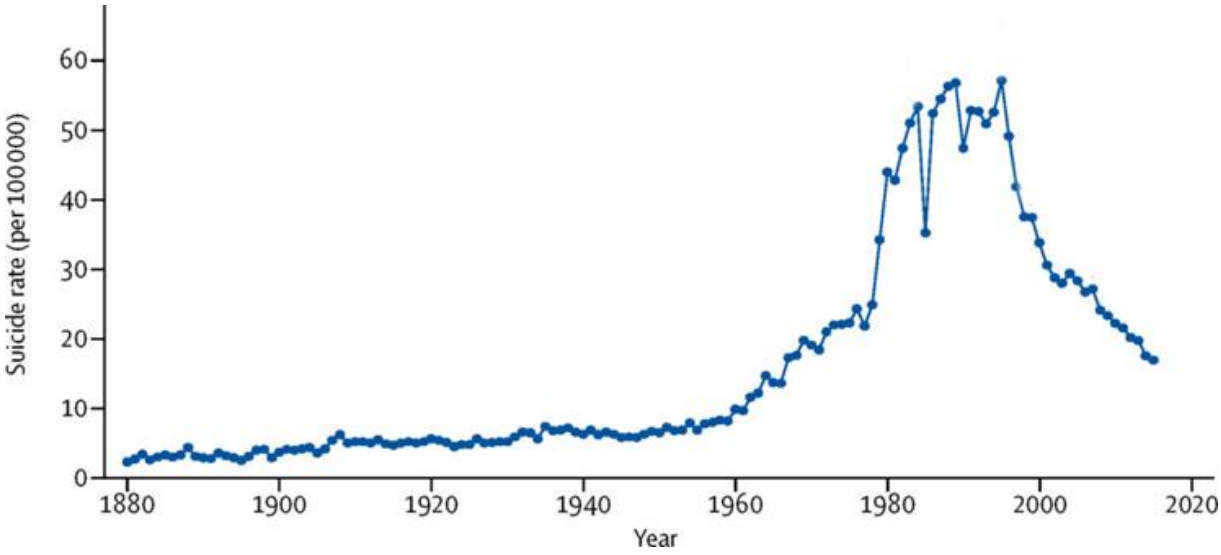



Figure 1. Incidences of Suicides in Sri Lanka from the Year 1880 to 2015

Suicide is a complex behavior and should not be oversimplified and need to be controlled through proper interventions and study since it is preventable. A person's risk of suicide is not only determined by personal characteristics but also influenced by their economical and sociological factors. Generally, there is no single factor but a cumulation of factors that lead an individual to suicide. Therefore, predicting the probable economical and sociological factors which trigger suicidal actions is important. If there is a proper mechanism to identify the vulnerable factors and association among those, which increase the likelihood of suicidal behavior, then the respective parties would be able to recognize the individuals who are at the risk and take necessary actions for supporting them since the best way to prevent suicide is to identify the early signs and respond appropriately.

## 1.3 Statement of the Problem

Suicide is preventable and everyone has a responsibility to prevent it at an individual, community, and national level. Suicide prevention is starting with identifying the risk factors and warning signs. There is massive interest in analyzing suicide data by using Machine Learning (ML) and Data Mining (DM) algorithms. A certain number of studies have been conducted using data analyzing techniques to identify the risk factors such as depression, anxiety, hopelessness, stress, or substance misuse. Suicidal behaviors have been predicted and have identified the features of the people who are having a risk for the second attempt.

But it has been noticed that not only mental and physical disorders but also a wide range of civil, educational, and professional factors are associated with an increased risk of suicide. But no research has considered the background of suicidal people and investigated the suicides by utilizing advanced analysis techniques.

## 1.4 Research Aims and Objectives

### 1.4.1 Aim

This research aims to predict the causative factors and the association among those by scanning the civil, educational, and professional backgrounds of the individuals in Sri Lanka who had intentionally end their own lives.

### 1.4.2 Objectives

In most cases, suicide does not cause by a single factor but a certain number of different factors. The combination of these factors could be the warning signs for suicide. One of the major objectives of the study is to increase the knowledge of the warning signs for suicide. Through that, it will be possible to educate the public on identifying the people who are at risk of suicide. If it is possible to spot those people, then the responsible parties would be able to send them support.

## 1.5 Scope

The study considered only the suicide incidents in Sri Lanka from the year 2014 to 2019 to predict the factors which affect the suicides in the country.

According to the effect of the action, suicidal behavior can be classified into two parts as completed and attempted. The completed suicide is an action that causes the death whereas if the person tries to commit suicide but still survives known as an attempted suicide. This research considered the completed suicides only.

This study is based on the discipline of DM and used the data that was collected annually by the Department of Police, Division of Statistics, Sri Lanka (“Crime Statistics,” 2020). The data set consists of the information of 18,906 suicide incidences. The attributes that have been considered are Age Group, Gender, Civil Status, Education Level, Nature of the Occupation, and Reason for Suicide.

## 1.6 Structure of the Report

The report consists of five chapters. Chapter 1 gives an overall introduction to the study. Chapter 2 critically reviews the work done on identifying the dimensions of suicide. Chapter 3 analyzes the problem and presents a detailed methodology of identifying the factors that influence suicides in Sri Lanka by analyzing the background of individuals. The chapter presents the contribution made to the field of computer science through the research work carried out. Chapter 4 describes the method used to evaluate the outcome of the research and present the results. Finally, Chapter 5 concludes the study and indicates the identified future works.

# CHAPTER 2: LITERATURE REVIEW

## 2.1 Introduction

Suicide is a deliberate attempt that results in the voluntary death of the person who does it. Suicide can be influenced by several risk factors and identifying those factors would be helpful to target the higher risk groups and send support to them. Several studies had investigated suicide around the world in different disciplines. A comprehensive review of the literature regarding suicides in the field of DM and ML is described and critically evaluated in the chapter.

## 2.2 Related Works

Joseph and Ramamurthy (Joseph and Ramamurthy, 2018) used DM techniques on the risk factors such as depression, anxiety, hopelessness, stress, or substance misuse to predict the suicidal behavior of individuals by creating a predictive model. The risk factors are calculated by using the suspected tweets which were used by the suicidal person to convey his feelings. They have compared five classification algorithms, Classification Via Regression, Logistic Regression, Decision Table, Random Forest and Sequential Minimal Optimization (SMO), and identified that the Classification Via Regression algorithm can predict with the highest accuracy. Abboute, Boudjeriou, Entringer, Aze, Bringay, and Poncelet (Abboute et al., 2014) also studied suicide prevention by mining Twitter messages. The methodology has consisted of four steps. As the first step, the research has identified nine topics that are usually talked about by suicidal people and then obtained the vocabulary. As the second step, a corpus of tweets containing the words of the defined vocabulary is collected. Next manually annotated the messages as risky and non-risky and finally used six classifiers, JRIP, IBK, IB1, J48, Naive Bayes, SMO to train the dataset. The study has identified that Naïve Bayes gives the best output. Braithwaite, Giraud-Carrier, West, Barnes, and Hanson (Braithwaite et al., 2016) have validated the use of ML algorithms to identify the people who are at risk of suicide by analyzing the data from Twitter. They have used the Decision Tree learning approach to differentiate the people who are at a suicidal risk from those who are not. The paper has concluded that those who are at high suicidal risk can be easily differentiated by ML algorithms. Cheng, Li, Kwok, Zhu, and Yip (Cheng et al., 2017) have researched to identify a Chinese individual's risk for committing suicide and his/her emotional distress by analyzing their Webio posts using Natural Language Processing and ML analysis. Although the accuracy that they have achieved is satisfactory, the performance of the classifiers needs to be improved in terms of depression and stress level. These researches have analyzed the social media posts of suicidal people and tried to predict the individuals who are at risk of suicide. To achieve that task, they have used Natural Language

Processing and ML techniques. None of these studies considered the background of the people as risk factors.

Bae, Lee, and Lee (Bae et al., 2015) have analyzed the middle and high school students in Korea using a decision tree. They have considered the dependent variable as suicide attempt and other social, demographic, interpersonal, and extra-personal factors as independent variables and identified that depression is the most suitable variable for predicting suicide attempts. Omprakash (Omprakash, 2013) has implemented a counseling system for predicting the suicidal behaviors of students by analyzing two kinds of data. The first one is known as Internal and the data was collected through student management systems. The second one is referred to as external data which denote the demographic details of the students, area, culture, the attitude of the people towards education, etc. The research has applied different algorithms on the data sets and identified that regression and tree M5P algorithms give the best prediction. These researches have considered the behavior, attitude, and social relationships to predict suiciding. Again, these studies also didn't pay attention to the background details of the people.

Boonkwang, Kasemvilas, Kaewhao, and Youdkang (Boonkwang et al., 2018) have collected a suicide and self-harm surveillance report and compared the ID3, C4.5, and naïve Bayes algorithms on the attributes of individuals and identified the characteristics of the people who are at the risk of a second attempt. The study concluded that C4.5 gives the best prediction. Choo, Diederich, Song, and Ho (Choo et al., 2014) have analyzed the medical records of patients who have attempted suicide, to identify the factors influencing suicide. As the first step, the study has applied text mining algorithms on the medical records, Further, they have used two-step cluster analysis on the preprocessed data to identify the groups which are having similar cases. These studies analyzed the medical records of the patient to identify the suiciding factors or the second attempt of suiciding. No consideration regarding the influence of civil, educational, and professional background of individuals towards suiciding.

Iliou, Konstantopoulou, Lymperopoulou, Anastasopoulos, Anastassopoulos, Margounakis, and Lymberopoulos (Iliou et al., 2019) have compared the Iliou preprocessing method and the Principal Component Analysis to predict suicide by analyzing the history of the family. The data set was collected through some clinical interviews by asking the situation of the family. The two preprocessing methods have been validated by using ten classification algorithms and identified that Iliou preprocessing method is more suitable to predict suicide.

## 2.3 Identified Research Gap

Many researchers have investigated the motivation or the reason for suicide and the risk factors associated with suicidal behavior and suicide. These researches did not consider the background details of suicidal people such as age group, gender, civil status, education level, nature of the occupation as risk factors. Rather, the studies are based on social media posts/comments/messages or medical records which reflect the psychological status of human beings. As a result, these researchers were able to recognize certain types of social media messages and mental disorders to detect suicides. No researches have been done in the field of DM or ML to investigate the influence of civil, educational, and professional background of individuals which could be a reason for someone to lead to such mental situations and commit suicide.

The other limitation is that there are no research studies that have explicitly investigated the factors affecting suicides in Sri Lanka in the field of DM. The factors which are leading someone for taking their own lives could be different according to the cultural, sociological, and economical background of the country. Therefore, Comparisons should be avoided and it is necessary to research the country itself.

So, comprehensive research is required to identify the factors influencing Sri Lankans to end their own lives by exploring the civil, educational, and professional backgrounds of the individuals who have committed suicide.

## 2.4 Summary

As a research area, suicide has evolved for so many years. Several types of research that addressed different research issues related to the dimensions of suicide have been found in the literature. These issues are suicidal behavior prediction, characteristics identification, second attempt prediction, etc. No formal study was found in investigating the risk factors of suicide by analyzing the background of the people (civil, educational, and professional) in Sri Lanka.

# CHAPTER 3: METHODOLOGY

## 3.1 Introduction

The comprehensive review of the literature showed that a certain number of studies have been conducted to identify the factors affecting suicide by using various ML and DM algorithms. Most of these researchers are based on the mental health and other clinical data or the behavioral changes of individuals. As a result, a research gap is identified and decided to explore the prospective efficiency of advanced analysis techniques for identifying the causative factors of suicide and the association among those by analyzing a data set that consists of the civil, educational, and professional background of the individuals who have committed suicide in Sri Lanka. Hence, as a solution, a new model is proposed and implemented in this research by using data analysis techniques.

## 3.2 Problem Analysis

Suicide in Sri Lanka is a critical and long-term social issue that is caused for the domestic, health, and economical costs of the country. According to the estimations, suicides among males are greater compared to the females in Sri Lanka. The rates of suicides are higher among older compared to the younger males and rates of young females are higher than older females. These rates are similar to the worldwide patterns (Kathriarachchi et al., 2019). The statistics show that the majority of victims were aged 15 to 44. District-wise, maximum suicide cases were reported in Jaffna, Vavuniya, Monaragala, and Polonnaruwa districts which are mostly considered as rural areas. Colombo and Galle are recorded as districts which are having minimum suicide records (“Suicide in Sri Lanka,” 2018).

Several researchers have identified many risk factors that caused someone to end their own lives, such as mental disorders, physical disorders, drug abuse, psychological states, cultural, family, and social situations, harassment or bullying, socioeconomic problems, discrimination, etc. Previously attempted suicide can be considered a risk factor as well (“Suicide,” 2021). Yet it has been noticed that suicide can be caused not only by the above-mentioned factors but also by the background of the people such as age, gender, civil status, educational level, and professional status as well. In other words, a person's background influences their risk for attempting suicide.

## 3.3 Complexity of the Data Set

The data set for the analysis was obtained from the Department of Police, Division of Statistics, Sri Lanka ("Crime Statistics," 2020) which is consisted of 18,906 suicide incidences for 6 years period from the year 2014 to 2019. The data set comprised of background information (age group, gender, civil status, educational level, nature of occupation) and the reason for suicide of each individual who committed suicide. The data set contains only categorical data. The values of each attribute are given in Table 1. Since the categories of each attribute are having lengthy values, codes are assigned to each distinct value as shown in Figure 2.

To understand the problem in-depth, five graphs have been created which present the suicide rates for the civil status, the standard of education, nature of the occupation, the reason for suicide, and gender categorized by the age group from the year 2014 to 2019.

Figure 3 demonstrates the suicide rates according to civil status. The graph illustrates that for both gender types, married people are having higher rates. In some age groups, a considerable amount of divorced and widows have committed suicide as well. Figure 4 shows the standard of education categorized by the age group. According to the graph, people who did not attend school and those who are having university degrees are having fewer suicide rates compared to other individuals. Figure 5 demonstrates the nature of occupation for each age group. It has been observed that for both males and females, unemployed people are having higher rates. Finally, Figure 6 illustrates the reason for suicide. According to the graph, both males and females committed suicide due to harassment by family members or due to economic problems, or some disease. For males, a considerable number of individuals have committed suicide due to drug addiction. Figure 7 shows the suicide rates according to gender. It has been noticed that males are having higher suicide rates rather than females in all age groups except for the age 8-16.

According to these graphs, it has been observed that the relationships among the attributes are complex and there can be several hidden associations. Thus, a study should be carried out to recognize those hidden patterns for identifying the risk factors of suicide by dealing with a large, complex data set. DM is considered a process that uncovers hidden patterns and relationships in large, complex data sets. Therefore, by considering the complexity associated with both the size and the content of the data set, it has been decided to use DM techniques to identify the hidden patterns.

Table 1. Details of Attributes of the Data Set

| ***Attribute*** | ***Values*** |
|---|---|
| Age | Age groups |
| Gender | Male, Female |
| Civil Status | Unmarried, Married, Illegal Married, Divorced, Widow, Legally Separate |
| Educational Level | School not attended, Grade 1 to 7, Passed grade 8, Passed G.C.E. O/L, Passed G.C.E. A/L, University degree or above, Other |
| Nature of Occupation | Professional Technical and related workers (Doctors/Engineers/Accountants/ Teachers/Authors/ Photographers), Administrative Executive Managerial and related workers, Clerical and related workers (Stenographers/ Typists/ etc.), Sales workers, Service workers (Cooks/Tailors/Barbers/ etc.), Agricultural Animal Husbandry Fisherman and related Forestry workers, Production process workers Craftsman and related workers transport equipment operators and laborers, Armed Services, Police, Security Personnel, Pensioners, Students, Politicians, Unemployed persons, Workers not classified by occupation |
| Reason for Suicide | Economic problems (Poverty indebtedness), Employment problems, Problems caused with the elders, Harassment by the husband and family disputes, Disappointment frustration caused through love affairs, Subjection to sexual harassment/Rape, Addiction to narcotic drugs, Aggrieved over the death of parents/relations, Loss of property, Failure at the examination, ill-treatment by the children, Sexual incapacity, Mental disorders, Chronic diseases, and Physical disabilities, Other reasons |

| Attribute | Value | Code |
|---|---|---|
| Age | 08 Yrs - 16 Yrs | 8-16y |
| | 17 Yrs - 20 Yrs | 17-20y |
| | 21 Yrs - 25 Yrs | 21-25y |
| | 26 Yrs - 30 Yrs | 26-30y |
| | 31 Yrs - 35 Yrs | 31-35y |
| | 36 Yrs - 40 Yrs | 36-40y |
| | 41 Yrs - 45 Yrs | 41-45y |
| | 46 Yrs - 50 Yrs | 46-50y |
| | 51 Yrs - 55 Yrs | 51-55y |
| | 56 Yrs - 60 Yrs | 56-60y |
| | 61 Yrs - 65 Yrs | 61-65y |
| | 66 Yrs - 70 Yrs | 66-70y |
| | Over 71 Yrs | 70y+ |
| Gender | Male | MALE |
| | Female | FEMALE |
| Civil Status | Unmarried | U |
| | Married | M |
| | Illegal Married | IM |
| | Divorced | D |
| | Widow | W |
| | Legally Separate | LS |
| Educational Level | School not attended | SN |
| | Grade 1 to 7 | G17 |
| | Passed grade 8 | P8 |
| | Passed G.C.E. O/L | POL |
| | Passed G.C.E. A/L | PAL |
| | University degree or above | UD |
| | Other | OTH |
| Nature of Occupation | Professional, Technical and related workers | PRO |
| | Administrative, Executive Managerial and related workers | ADM |
| | Clerical and related workers | CLE |
| | Sales workers | SALE |
| | Service workers | SER |
| | Agricultural, Animal, Husbandry, Fisherman and related Forestry workers | AGR |
| | Production process workers, Craftsman, Transport equipment operators and Laborers | PROD |
| | Armed Services | ARMY |
| | Police | POICE |
| | Security Personnel | SECURE |
| | Pensioners | PEN |
| | Students | STD |
| | Politicians | POLITI |
| | workers not classified by occupation | NOOCUP |
| | Unemployed persons | UNEMP |
| Reasons for Suicides | Economic problems | ECO |
| | Employment problems | EMP |
| | Problems caused with the elders | ELD |
| | Harassment by the husband and family disputes | HAR |
| | Disappointment, frustration caused through love affairs | DISLUV |
| | Subjection to sexual harassment/Rape | SEXHAR |
| | Addiction to narcotic drugs | DRUG |
| | Aggrieved over the death of parents/relations | AGRDEATH |
| | Loss of property | PROLOSS |
| | Failure at the examination | EXMFAIL |
| | Ill-treatment by the children | CHILD |
| | Sexual incapacity | SEXINC |
| | Mental disorders | MENDIS |
| | Chronic diseases and Physical disabilities | DISEASE |
| | Other | OTHREA |

Figure 2. Details of the Code Values Assigned

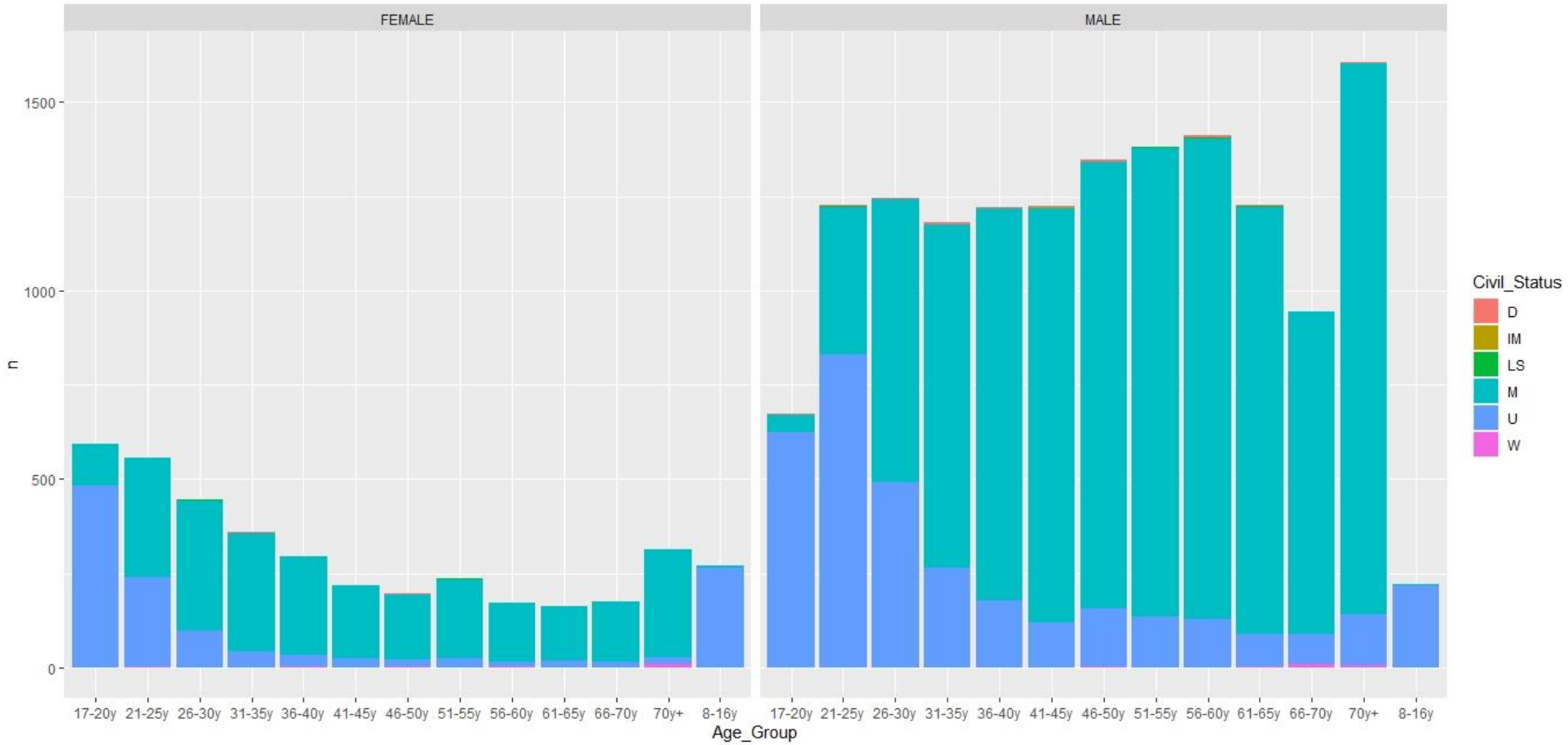



Figure 3. Civil Status Categorized by the Age Group

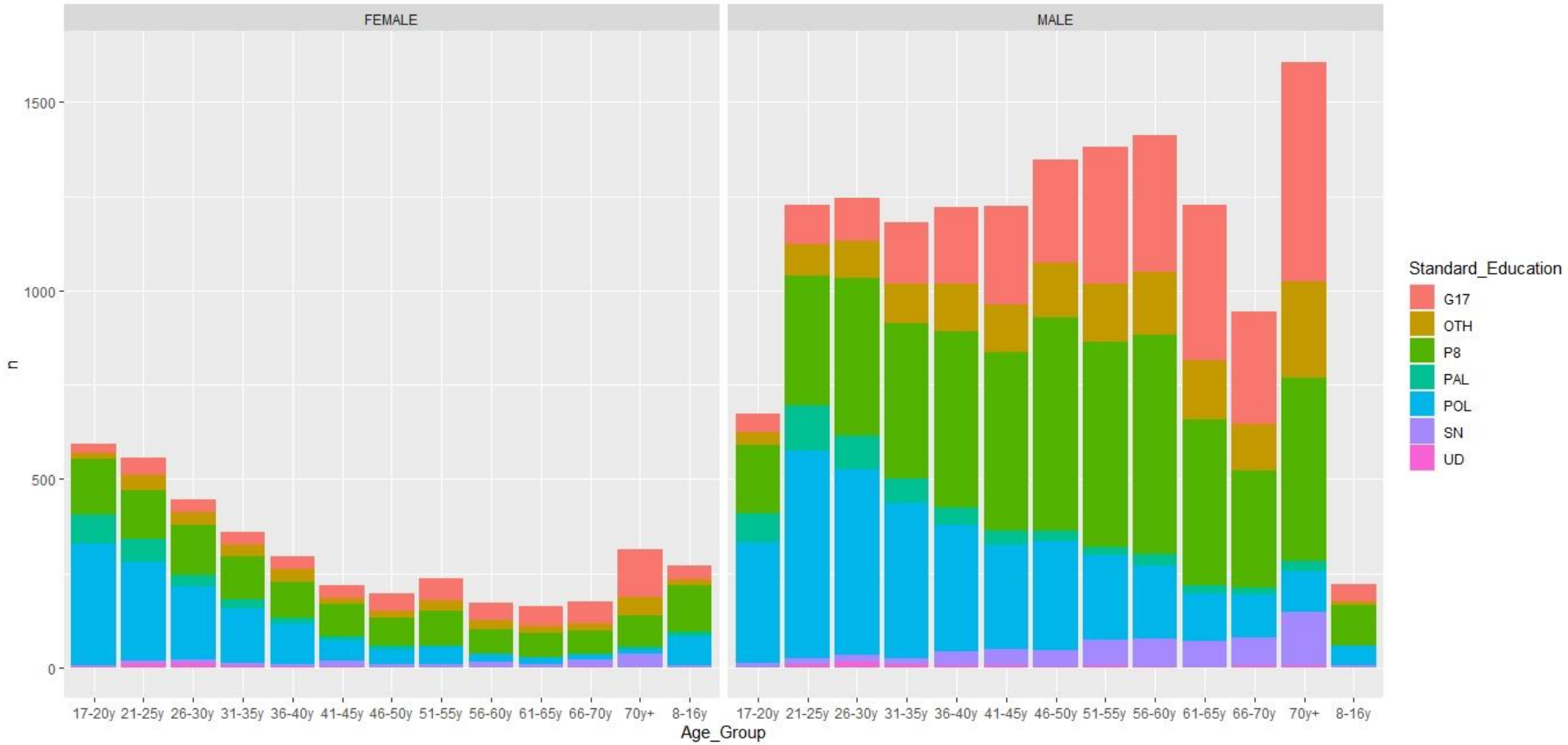



Figure 4. Standard of Education Categorized by the Age Group

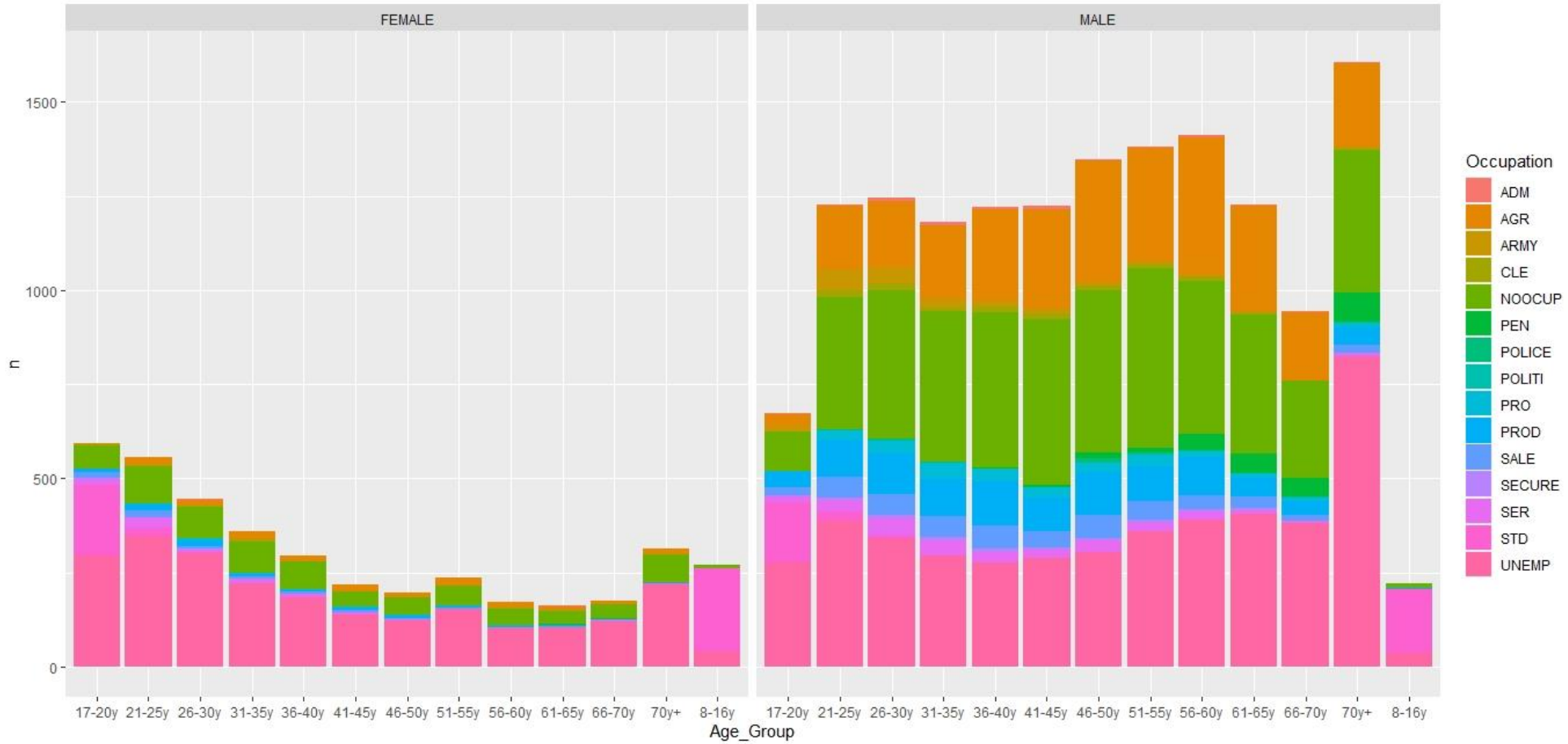



Figure 5. Nature of Occupation Categorized by the Age Group

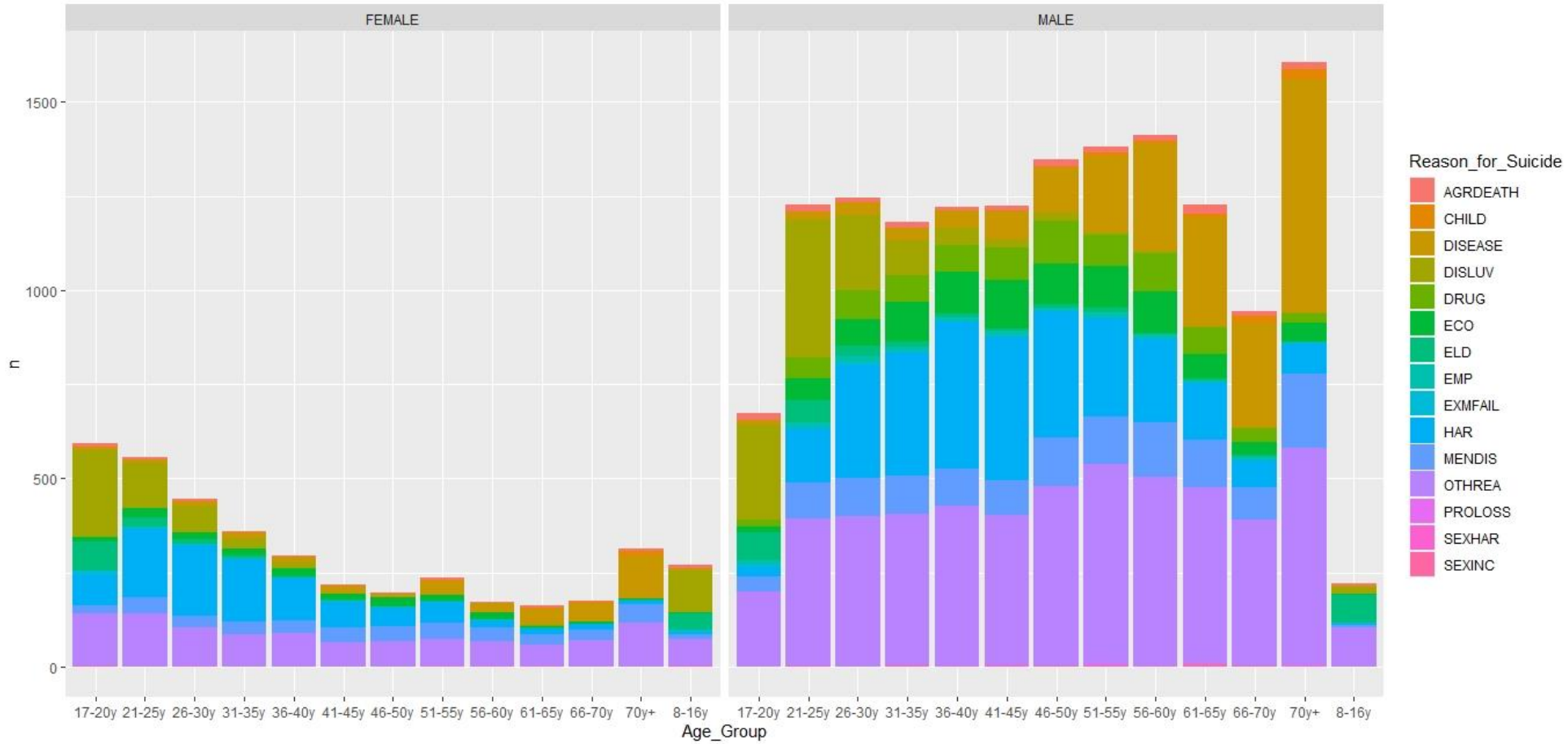



Figure 6. Reason for Suicide Categorized by the Age Group

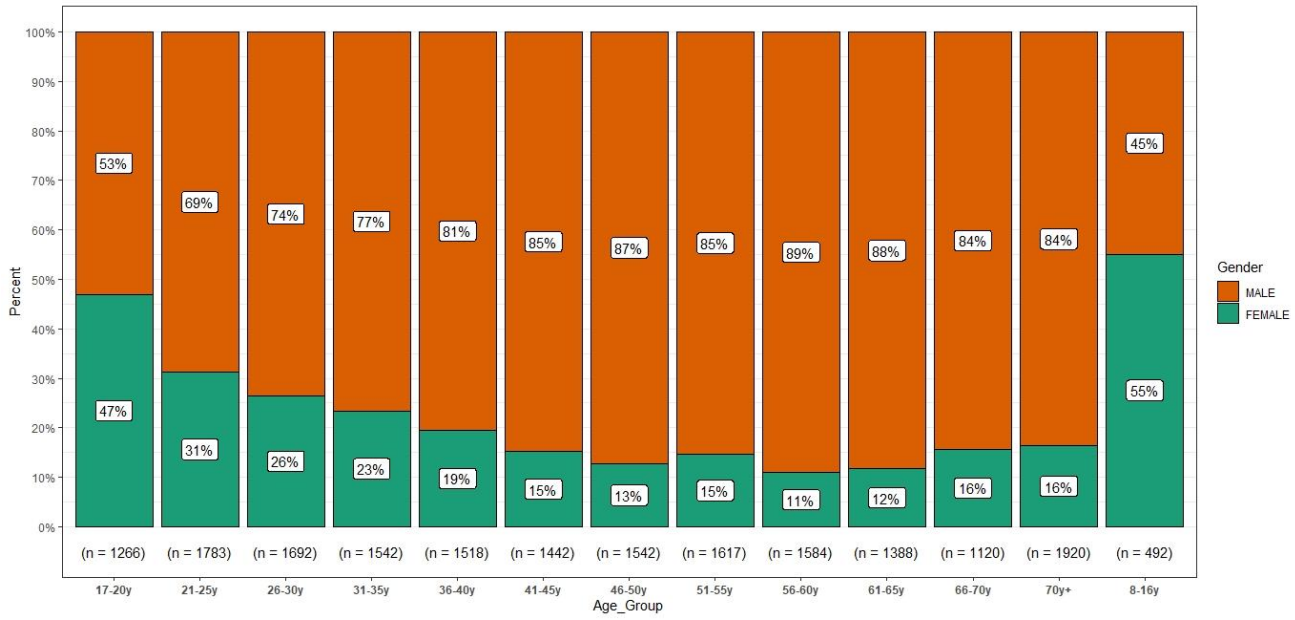



Figure 7. Suicide Rates According to Gender

## 3.4 Data Mining Overview

DM is a process used to extract implicit, unrevealed, and potentially useful patterns or knowledge from large amounts of raw data. DM has improved decision-making by uncovering hidden information and providing knowledge.

DM techniques can be divided into two methods as shown in Figure 8 (Weiss and Indurkhya, 1997). The two main techniques are:

1. Supervised Learning: supervised learning is used to train the machine using data that is labeled. That means, the supervised learning algorithm learns from labeled, trained data and makes predictions. This technique requires upfront human intervention to label the data. Supervised learning can be further subdivided into classification and regression.

2. Unsupervised Learning: Unsupervised learning allows the model to discover information on its own. It deals with unlabeled data. This technique can be used to find hidden patterns or to categorize data. Unsupervised learning can be further subdivided into association rule mining and clustering.

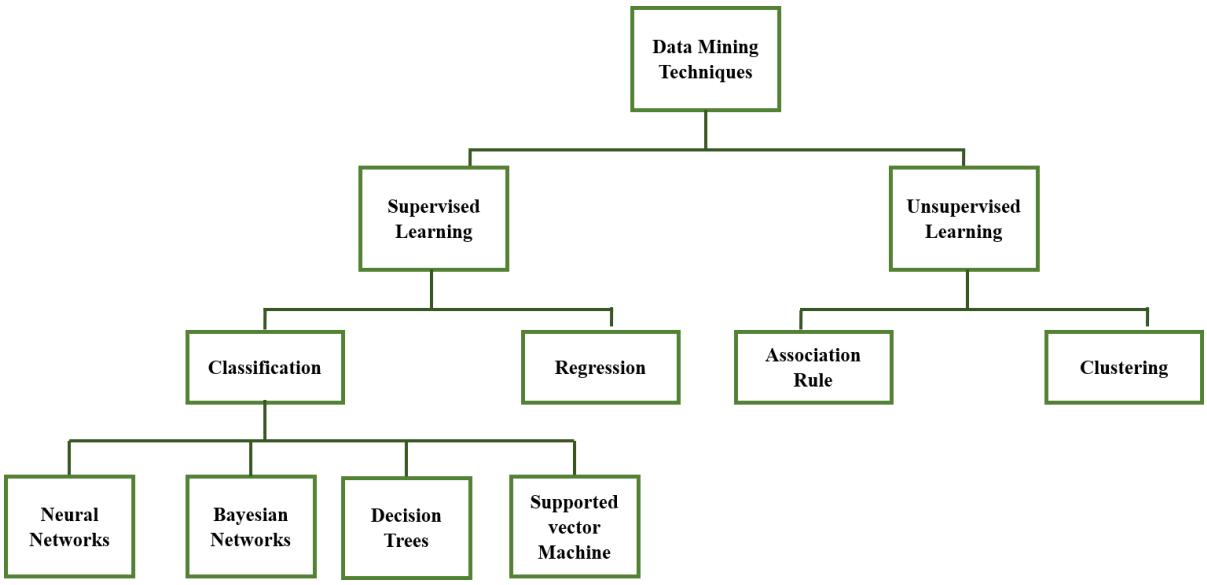



Figure 8. Data Mining Techniques

The DM process consists of several steps which are shown in Figure 9 (Han and Kamber, 2006). The steps are:

1. Data cleaning (The process of removing incomplete, incorrect, corrupted, or duplicate data)
2. Data integration (The process of combining data from multiple sources)
3. Data selection (The process of determining appropriate data from the database)
4. Data transformation (The process of converting data into format or structure appropriate for mining)
5. Data mining (The process of extracting data patterns)
6. Pattern evaluation (The process of recognizing the useful and interesting patterns)
7. Knowledge presentation (The process of presenting the mined knowledge to the user for visualization)

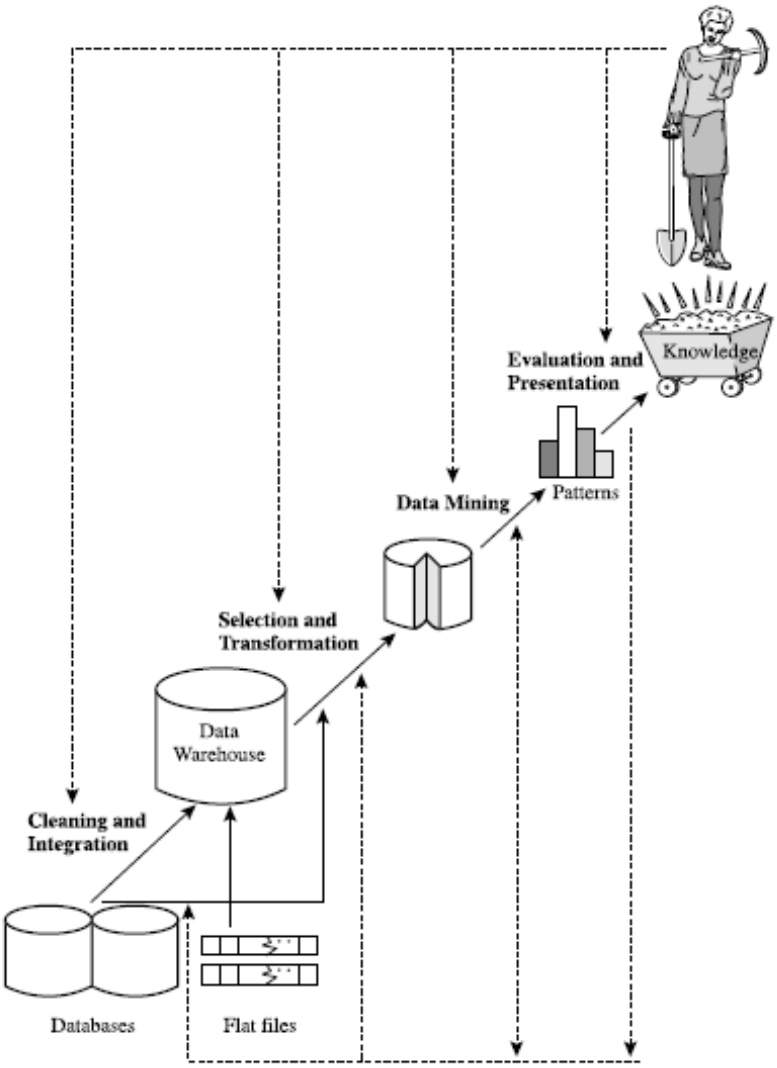



Figure 9. The process of Data Mining

## 3.5 Proposing Model

Figure 10 shows the proposing model of the study. Initially, the data set contain the details of individuals who have committed suicide from the year 2014 to 2019.

Since the data set is obtained through real-world data, which is incomplete and may contain missing data, noisy data, or inconsistent data, the data preprocessing techniques are applied before it sends through the model. As the next stage, it was planned to apply a clustering algorithm on the preprocessed data set to divide the entire data set into several homogeneous groups. To identify the associations between attributes of suicide data, an association rule mining algorithm is decided to use on each cluster and the entire data set

The reason behind this is that, if only the entire data set is analyzed through association rules, more important relations can be hidden. Hence, association rules will be generated for every cluster to discover the different suicide-prone circumstances (Kumar and Toshniwal, 2015).

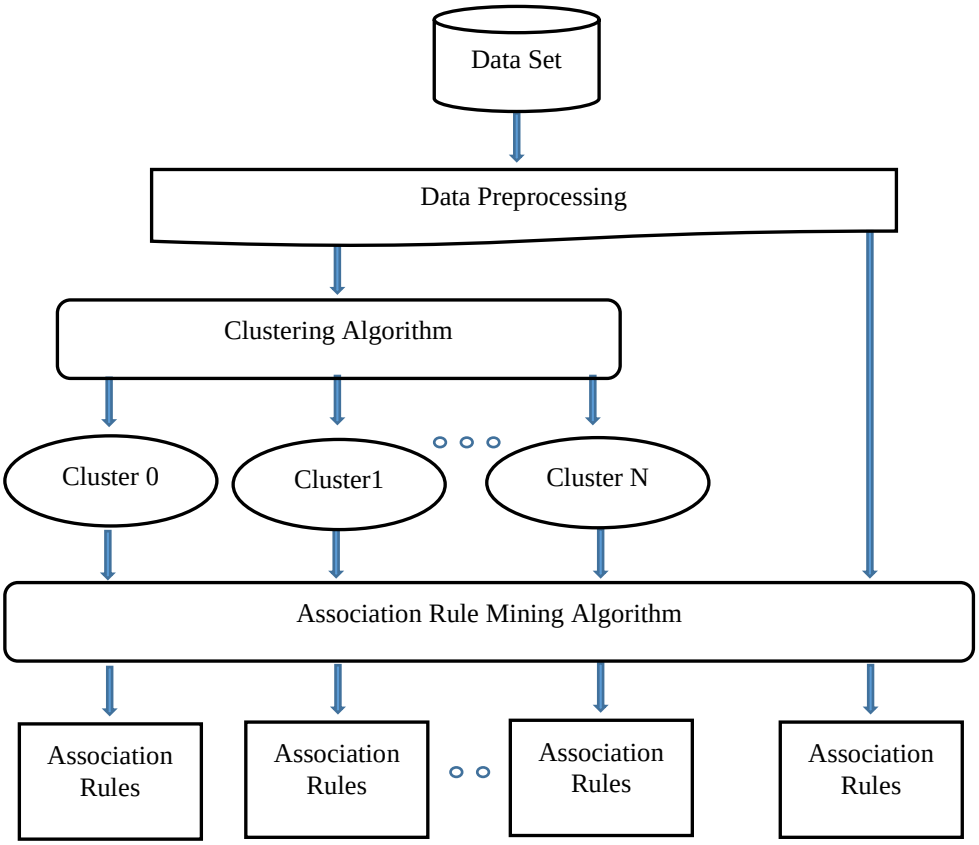



Figure 10. Proposing Model for Analysis

## 3.6 Data Preprocessing

Data Preprocessing involves transforming raw data into an efficient format. Real-world data usually consist of noise, missing values, and irrelevant attributes. Data preprocessing is a proven method for solving such issues and make the data set appropriate for the analysis and further processing.

As the first step data cleaning has been done. In some individual suicide incidents, missing values have been observed in certain attributes.

Missing values is a very common problem in a real-world data set. Five methods can be used to handle the missing values in a categorical data set. The methods are ("Missing Values | Treat Missing Values in Categorical Variables," 2021):

1. Delete the observations: If the dataset is large enough and the values which need to be predicted are well represented in the data set, then delete the entire row that contains the missing values.
2. Replace missing values with the most frequent value.
3. Develop a model to predict missing values: A training classifier can be used by considering the missing value as the dependent variable against the other attributes in the data set.
4. Delete the variable: If the number of missing values is exceptionally large, then delete the entire column which contains the missing values.
5. Apply unsupervised ML techniques: The idea behind this is that cluster the data set by skipping the columns which contain the missing values. Then it will be able to identify the category which contains the missing values.

Since the suicidal data set is large enough and few missing values are contained, it was decided to use the first method and delete the entire rows. Certain data entry errors were also found and the method used to handle the missing values is applied here to handle such errors as well.

In the original data set, the values for each attribute are having long names as shown in Table 1. To reduce both complexity and data storage, as the second step codes have been assigned for each value which is given in Figure 2.

Finally, it has been noticed that both the attributes "Educational level" and "Reason for suicide", are having values as "Other". Since the purpose of the study is to identify the relationship between the six attributes and the value "Other" would make meaningless associations, such rows have been removed by considering the “Other” values as irrelevant values. In the end, the data set is ready for analysis with 11,269 data.

## 3.7 Data Analyzing

### 3.7.1 R Programming

R is an integrated suite of software that is used in data science, statistics, and visualization. The popularity of R is increased in recent years among DM researchers due to the following benefits (“Pros and Cons of R Programming Language - Unveil the Essential Aspects!,” 2019):

- Open-source
- Platform independent

- Good for data analyzing
- Availability of various, large sets of libraries
- Powerful graphics
- Compatibility

Therefore, by considering the above-mentioned advantages, it has been decided to use R programming for the data analysis and visualization in the study.

### 3.7.2 Clustering

Clustering is an unsupervised learning technique that categorizes a set of objects into groups of similar objects such that the objects within a group are similar to one another and objects in different groups are dissimilar from each other. There are different types of clustering algorithms and usually, the most suitable algorithm for a particular data set depends on the type of the data set. For example, categorical data are different from numerical data since categorical data are discrete. Therefore, the clustering algorithms used for numerical data cannot be used for categorical data (Saini, 2021).

The suicide data set used in the study is a categorical data set. The standard k-means algorithm cannot apply to a categorical data set directly as it uses the Euclidean distance function and it is impossible to calculate the distance for categorical data points. Sample space for categorical data is discrete and does not have a natural origin. So, the Euclidean distance function on such space is meaningless.

Converting categorical data into numerical data can be considered as a solution to the above problem. In this case, the One-Hot encoding method can be used with the k-means algorithm to get the numerical presentation for the categorical data by assigning a binary value to each unique category. But this will increase the size of the data set extensively since the suicide data set have a large number of categories. And also the cluster means are meaningless as the 1 and 0 are not the real values (Huang, 1998).

Hence, it has decided to apply the k-modes algorithm on the data set by considering the following reasons (Huang, 1998):

a) k-modes algorithm is best for categorical data sets.
b) k-modes efficiently handle a large amount of data.

k-modes algorithm is an extension of the k-means algorithm which replaces the means of the clusters by modes. The k-modes algorithm also replaces the Euclidean distance function used in the k-means algorithm with the Simple matching distance by calculating the dissimilarities between the data points. If the dissimilarities are lesser, then data points are considered more similar. The dissimilarity is computed by counting the number of mismatches in all variables. (Saini, 2021), ("KModes Clustering Algorithm for Categorical data," 2021).

In this study, the simple matching distance is set as weighted. That means weighted distance will be used by considering the frequencies of the categories in the data. This will make it possible to generate clusters with stronger intra-similarities and will increase the performance of clustering (He et al., 2011).

The k-mode algorithm works as follows ("KModes Clustering Algorithm for Categorical data," 2021):

1. Specify the number of clusters (k).
2. Randomly pick k number of observations and use those as leaders/clusters.
3. Calculate the dissimilarities between each leader and each of the other observations in the data set and assign each observation into the closest cluster.
4. Considering one cluster at a time, for each feature, look for the mode and update the new leaders.
5. Repeat 3, 4 steps until no assignments are required.

The most fundamental issue that arises with the k-modes clustering, is that determining the optimal k to be formed by a clustering algorithm. Initially, it was tried to identify the k by using both the elbow method and the average silhouette method. But it was difficult to choose the optimal k with the above two methods due to the high-dimensional characteristics of the data set. Therefore, the Bayesian Information Criterion (BIC) is decided to use for identifying the optimum k. The k of the maximum BIC is the optimal number of clusters. Hence, the optimal k is identified as five as shown in Figure 11.

After identifying the k, the k-modes algorithm is applied through R programming to segment the data set. Cluster 0, Cluster 1, Cluster 2, Cluster 3 and Cluster 4 consists of 729, 1419, 2960, 1034 and 5127 data respectively.

It has been noticed that the five clusters are centered on five different reasons for suicide. According to that Cluster 0, Cluster 1, Cluster 2, Cluster 3, and Cluster 4 are labeled as "Mental Disorders", "Addiction to Narcotic Drugs", "Chronic Diseases", "Problems Caused with the Elders" and "Harassment by the Husband and Family Disputes" respectively.

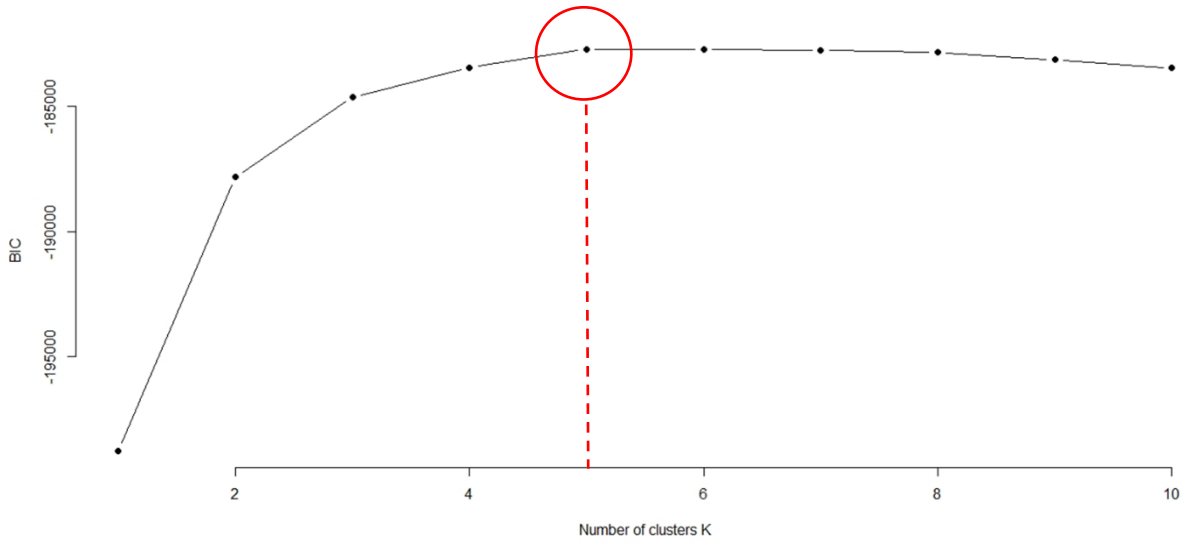



Figure 11. BIC for the Optimum Value of k

### 3.7.3 Association Rule Mining

Association rule mining is a rule-based technique used for finding associations or relationships among variables in large data sets. The technique identifies frequent if-then associations known as association rules that consist of an antecedent (if) and a consequent (then). For example, a rule X → Y indicates that if X occurs then Y will also occur.

There are three measures in association rule mining (Adekanmbi 'Yosola, 2018).

1. Support

   Support indicates how frequently an item set appears in the data set.

$$Support\ (\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Total\ number\ of\ transactions}$$

2. Confidence

   Confidence is a percentage value that shows the frequency of the if-then statement.

$$Confidence\ (\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

3. Lift

   Lift is used to compare the expected confidence with actual confidence.

$$Lift\ (\{X\} \rightarrow \{Y\} = \frac{Transactions\ containing\ both\ X\ and\ Y / Transactions\ containing\ X}{Fractions\ of\ transactions\ containing\ Y}$$

Apriori is an association rule mining algorithm that can be used to find the frequent itemsets by using the Breadth-First Search and Hash Tree. The algorithm is named Apriori since it uses the prior knowledge of frequent itemset properties.

To identify the associations between the background details of the individuals and the reason for suicide, the Apriori algorithm has been applied to each cluster and the entire data set. The Support for the analysis is considered as 0.01 and the Confidence for the analysis is set at 1. Usually, in the traditional market basket applications, the support is set to a relatively high level. The reason is, in business applications low attention is paid to the low frequency of goods and they are more concerned about the associations between the goods which are having a high frequency. But this scenario is not applicable with the analysis of suicide risk factors since, for the suicide data, the frequency of serious suicide incidences is very low, but the loss could be huge. Therefore, to identify the impact of serious suicide cases, it is needed to set relatively low support.

First, the Apriori algorithm has been applied to the entire data set. But it does not produce any single rule which shows the association between the five main attributes that could lead someone to suicide with a particular reason. Therefore, by considering the size of the suicide data set, it has been decided to use the Frequent Pattern (FP) Growth algorithm on the data set. FP Growth is a partition-based frequent itemset mining algorithm that is used to identify the hidden patterns of a large data set by partitioning the data set transactions into groups. But FP Growth also produced the same output as the Apriori algorithm (Pai, 2019). As the next step rules have been generated for each cluster as shown in Table 2.

Table 2. Cluster-wise Association Rules

| ***Cluster No.*** | ***Rule Body*** | ***Support*** | ***Lift*** |
|---|---|---|---|
| 0 | {21-25Yy, MALE, PROD}→{MENDIS} | 0.01369863 | 3.187773 |
| | {46-50y, M, MALE, NOOCUP, P8}→{MENDIS} | 0.02191781 | 3.187773 |
| | {46-50y, G17, M, MALE, NOOCUP}→{MENDIS} | 0.01369863 | 3.187773 |
| | {46-50y, NOOCUP, POL}→{MENDIS} | 0.01095890 | 3.187773 |
| | {46-50y, G17, M, UNEMP}→{MENDIS} | 0.01369863 | 3.187773 |
| | {46-50y, FEMALE, M, NOOCUP}→{MENDIS} | 0.01232877 | 3.187773 |
| 1 | {41-45y, M, MALE, NOOCUP}→{DRUG} | 0.02042254 | 2.204969 |
| | {41-45y, M, MALE, P8}→{DRUG} | 0.01830986 | 2.204969 |
| | {56-60y, M, MALE, P8}→{DRUG} | 0.03098592 | 2.204969 |
| | {66-70y, M, MALE}→{DRUG} | 0.01971831 | 2.204969 |
| | {G17, M, MALE, UNEMP}→{DRUG} | 0.02253521 | 2.204969 |
| 2 | {51-55y, M, MALE, P8}→{DISEASE} | 0.01452212 | 1.790206 |
| | {51-55y, P8, UNEMP}→{DISEASE} | 0.01013171 | 1.790206 |
| | {56-60y, AGR, MALE, P8}→{DISEASE} | 0.01046944 | 1.790206 |
| | {56-60y, M, MALE, P8, UNEMP}→{DISEASE} | 0.01013171 | 1.790206 |
| | {61-65y, M, MALE, P8, UNEMP}→{DISEASE} | 0.01215805 | 1.790206 |
| | {61-65y, AGR, M, P8}→{DISEASE} | 0.01046944 | 1.790206 |
| | {61-65y, G17, NOOCUP}→{DISEASE} | 0.01013171 | 1.790206 |
| | {61-65y, AGR, MALE, P8}→{DISEASE} | 0.01013171 | 1.790206 |
| | {66-70y, M, MALE, P8, UNEMP}→{DISEASE} | 0.01215805 | 1.790206 |
| 3 | {8-16y, MALE, P8, STD, U}→{ELD} | 0.02415459 | 2.320628 |
| | {8-16y, MALE, U, UNEMP}→{ELD} | 0.01642512 | 2.320628 |
| | {26-30y, MALE, U}→{ELD} | 0.01545894 | 2.320628 |
| | {G17, M, MALE}→{ELD} | 0.01739130 | 2.320628 |
| | {M, P8, UNEMP}→{ELD} | 0.01545894 | 2.320628 |
| | {M, MALE, P8}→{ELD} | 0.02608696 | 2.320628 |

*Continued*

Table 2. Concluded

| ***Cluster No.*** | ***Rule Body*** | ***Support*** | ***Lift*** |
|---|---|---|---|
| 4 | {26-30y, M, POL, UNEMP}→{HAR} | 0.01306552 | 2.131338 |
| | {26-30y, FEMALE, POL, UNEMP}→{HAR} | 0.01053042 | 2.131338 |
| | {46-50y, AGR, M, MALE}→{HAR} | 0.01306552 | 2.131338 |
| | {46-50y, M, UNEMP}→{HAR} | 0.01189548 | 2.131338 |
| | {46-50y, MALE, UNEMP}→{HAR} | 0.01014041 | 2.131338 |

## 3.8 Summary

This chapter has analyzed the problem and the data set and described the basic model that has been followed to achieve the main aim and the objectives of the research. The methodology can be divided into three parts as preprocessing, clustering, and association rule mining. Mainly two data analysis algorithms have been used, k-mode and Apriori. Through the k-mode clustering, the data set have been divided into five clusters and for both the entire data set and each cluster, the Apriori algorithm has been applied to identify the association rules among the attributes. Association rules gained by applying the Apriori algorithm to the entire data set do not contain any significant rule. The other five clusters generated 6, 5, 9, 6, and 5 number of rules Orderly.

# CHAPTER 4: EVALUATION AND RESULTS

## 4.1 Introduction

The study was able to identify associations between the attributes through the proposed model. The associations have gained five reasons for suicide by applying the Apriori algorithm for each cluster. This chapter discussed and critically evaluates the research work by analyzing those obtained associations rules. The chapter will include aspects such as the discussion of the results gained and the evaluation.

## 4.2 Discussion

Association rules for Cluster 0 show the background details of the people who could commit suicide due to mental disorders. The rules show that both married males and females who are within the age group 46-50 (middle age) and who do not have a proper occupation tend to suicide due to mental disorders. The reason may be that as a married person they may not able to support the family without any proper occupation. This could lead to mental disorders and finally to suicide. When it is considered the education level of these people, the rules show that they are having poor educational backgrounds. This low level of education may be the reason for their unemployment. Further, education is a powerful method that helps someone to control their emotions and heal themselves. These people might not be able to control themselves due to the lack of education. The rules also indicate that male production process workers/Craftsman and related workers/transport equipment operators/laborers who are within the 21-25 age group may commit suicide because of mental disorders. Usually, these kinds of employees are working continuously without having proper rest as well as proper salary. This could be lead them for suiciding with mental disorders.

Association rules for the Cluster 1 presents the rules regarding the reason “Addiction to narcotic drugs". According to the rules, unemployed married males who are having a poor educational background have a possibility of suicide because of their addiction to narcotic drugs. Although males in any age group can be addicted to drugs usually, the rules show that middle age and old males having a higher chance to commit suicide due to this addiction. The reason might be the physical and emotional changes that occurred as a result of the drug addiction over the years.

Association rules for the Cluster 2 shows the background details of the people who could commit suicide due to chronic diseases. Although the rules are scattered it is shown that either old or middle age people could be affected by chronic diseases and commit suicide.

Association rules for Cluster 3 targeted the reason "Problems caused with the elders". These rules show that young males, despite their civil status, could be suicided due to the problems caused with elders. And also, according to the rules these youngsters are unemployed or/and uneducated.

Association rules for Cluster 4 targeted the reason "Harassment by the husband and family disputes". These rules show that within the age group 26-30, unemployed married people or females can commit suicide due to the harassment of the husband and the family. And also, within the age group 46-50, unemployed married people or males may commit suicide as a result of the harassment. The rules clearly show that unemployed married people, despite their gender, are harassed by the husband or family members and commit suicide.

According to the rules obtained for all five clusters, it was observed that rather than females, males are tended to commit suicide more. The reason could be that women are more emotionally literate than men. Women used to discuss their feelings with others while men tried to hide those or use alcohol or drugs to cope with the stress.

It was observed that the rules are created only for five reasons which were identified as labels of the clusters. Therefore, the study was able to identify the factors influencing suicide because of mental disorders, addiction to narcotic drugs, chronic diseases, problems caused with the elders, and harassment by the husband and family disputes.

## 4.3 Evaluation

The Chi-squared statistical analysis approach is used to evaluate the rules obtained by applying the Apriori algorithm. The Chi-squared test is used to show the relationship between two categorical attributes. The method produces a value that shows the difference between the observed count and the expected count. If the significant value counted by the Chi-squared is less than 0.05, then the association is considered as a positive correlation (Mohd Shaharanee et al., 2009). Table 3, Table 4, Table 5, Table 6, and Table 7 show the Chi-squared test results for Cluster 0, Cluster 1, Cluster 2, Cluster 3, and Cluster 4. The negative correlations are highlighted in the tables.

Table 3. Chi-squared Test Results for the Cluster 0

| | ***Age Group*** | ***Gender*** | ***Civil Status*** | ***Education Level*** | ***Nature of Occupation*** |
|---|---|---|---|---|---|
| ***Reason for Suicide*** | p-value < 2.2e-16 | p-value = 0.02526 | p-value = 0.0008033 | p-value = 0.03302 | p-value < 2.2e-16 |

Table 4. Chi-squared Test Results for the Cluster 1

| | ***Age Group*** | ***Gender*** | ***Civil Status*** | ***Education Level*** | ***Nature of Occupation*** |
|---|---|---|---|---|---|
| ***Reason for Suicide*** | p-value < 2.2e-16 | p-value < 2.2e-16 | p-value = 0.4077 | p-value = 6.97e-07 | p-value = 0.01915 |

Table 5. Chi-squared Test Results for the Cluster 2

| | ***Age Group*** | ***Gender*** | ***Civil Status*** | ***Education Level*** | ***Nature of Occupation*** |
|---|---|---|---|---|---|
| ***Reason for Suicide*** | p-value < 2.2e-16 | p-value = 1.871e-15 | p-value = 8.326e-10 | p-value < 2.2e-16 | p-value < 2.2e-16 |

Table 6. Chi-squared Test Results for the Cluster 3

| | ***Age Group*** | ***Gender*** | ***Civil Status*** | ***Education Level*** | ***Nature of Occupation*** |
|---|---|---|---|---|---|
| ***Reason for Suicide*** | p-value < 2.2e-16 | p-value = 1.043e-07 | p-value = 3.816e-16 | p-value = 0.0005053 | p-value = 0.8056 |

Table 7. Chi-squared Test Results for the Cluster 4

| | ***Age Group*** | ***Gender*** | ***Civil Status*** | ***Education Level*** | ***Nature of Occupation*** |
|---|---|---|---|---|---|
| ***Reason for Suicide*** | p-value < 2.2e-16 | p-value < 2.2e-16 | p-value < 2.2e-16 | p-value < 2.2e-16 | p-value < 2.2e-16 |

According to Table 3, Table 5, and Table 7 all five main attributes, Age Group, Gender, Civil Status, Education Level and Nature of the Occupation, in the Cluster 0, Cluster 2, and Cluster4 have positive correlations with the attribute Reason for Suicide which suggests that the rules obtained for above mentioned three clusters can be considered as significant.

Table 4 shows that the attribute Civil Status has a negative correlation with Reason for Suicide in Cluster 1. For Cluster 1, it was able to acquire 5 rules none of them are significant according to the Chi-squared test results.

As shown in Table 6, the attribute Nature of Occupation also has a negative correlation with Reason for Suicide in Cluster 3. It was able to acquire 6 rules and among those 3 rules contained the nature of the occupation. Hence those 3 rules can be considered insignificant.

The study was able to produce 31 rules and according to the Chi-squared analysis 8 are insignificant. Therefore, it can be concluded that 74% of rules produced by the study are accurate.

## 4.4 Summary

The Chi-squared statistical analysis approach is applied to the data of each cluster to identify the correlation between the main five attributes and the Reason for Suicide. According to the values obtained in Cluster 1, Civil Status, and in Cluster 3, Nature of Occupation have no relation with the Reason for Suicide. Hence, among 31 total rules, 8 rules can be considered insignificant rules. An average, 74% of rules are significant and accurate.

# CHAPTER 5: CONCLUSION AND FUTURE WORKS

This research work aims to identify the risk factors and the association among those by examining the civil, educational, and professional backgrounds of the Sri Lankan people who had committed suicide. A new model has been proposed in the study by using the k-mode and Apriori algorithms. The study was able to cluster the data set into five groups based on five different reasons such as mental disorders, addiction to narcotic drugs, chronic diseases, problems caused with the elders, and harassment by the husband and family disputes. By applying the Apriori algorithm for each cluster, altogether 31 association rules were produced. It has been noticed that applying the association rules on each cluster is revealed important associations which may remain hidden if only the entire data set is analyzed. The rules show that generally unemployment, low education, gender, civil status, and certain age groups are having a close relationship with suicide. Finally, the study concludes that 74% of rules are significant.

## 5.1 Future Works

The research was able to produce the rules for five reasons of suicide only. But, according to the data set, there are fourteen different reasons of suicide. Therefore, it is necessary to identify the factors influencing suicide due to other reasons as well.

The study considered the completed suicide incidences only. By collecting the details regarding the attempted suicide incidences, it will be able to increase the size of the data set. This will produce more accurate results.

This research work considered the background details of the individuals such as age, gender, civil status, educational level, professional background only. But there are some other influencing environmental factors such as suicide attempts of family members, household socioeconomic position, household pesticide access, alcohol consumption in the household, etc. It is necessary to consider these factors as well.

# REFERENCES


Abboute, A., Boudjeriou, Y., Entringer, G., Azé, J., Bringay, S., Poncelet, P., 2014. Mining Twitter for Suicide Prevention, in: Natural Language Processing and Information Systems, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 250–253. https://doi.org/10.1007/978-3-319-07983-7_36

Adekanmbi 'Yosola, 2018. Association Rule Mining - Apriori Algorithm [WWW Document]. Medium. URL https://blog.usejournal.com/association-rule-mining-apriori-algorithm-c517f8d7c54c (accessed 5.16.21).

Bae, S.M., Lee, S.A., Lee, S.-H., 2015. Prediction by data mining, of suicide attempts in Korean adolescents: a national study. Neuropsychiatr. Dis. Treat. 11, 2367–2375. https://doi.org/10.2147/NDT.S91111

Boonkwang, K., Kasemvilas, S., Kaewhao, S., Youdkang, O., 2018. A Comparison of Data Mining Techniques for Suicide Attempt Characteristics Mapping and Prediction, in: 2018 International Seminar on Application for Technology of Information and Communication. Presented at the 2018 International Seminar on Application for Technology of Information and Communication, pp. 488–493. https://doi.org/10.1109/ISEMANTIC.2018.8549835

Braithwaite, S.R., Giraud-Carrier, C., West, J., Barnes, M.D., Hanson, C.L., 2016. Validating Machine Learning Algorithms for Twitter Data Against Established Measures of Suicidality. JMIR Ment. Health 3, e21. https://doi.org/10.2196/mental.4822

Cheng, Q., Li, T.M., Kwok, C.-L., Zhu, T., Yip, P.S., 2017. Assessing Suicide Risk and Emotional Distress in Chinese Social Media: A Text Mining and Machine Learning Study. J. Med. Internet Res. 19, e243. https://doi.org/10.2196/jmir.7276

Choo, C., Diederich, J., Song, I., Ho, R., 2014. Cluster analysis reveals risk factors for repeated suicide attempts in a multi-ethnic Asian population. Asian J. Psychiatry 8, 38–42. https://doi.org/10.1016/j.ajp.2013.10.001

Crime Statistics [WWW Document], 2020. URL https://www.police.lk/index.php/item/138 (accessed 12.3.20).

Han, J., Kamber, M., 2006. Data Mining: Concepts and Techniques, 2nd ed. Diane Cerra.

He, Z., Xu, X., Deng, S., 2011. Attribute value weighting in k-modes clustering. Expert Syst. Appl. 38, 15365–15369. https://doi.org/10.1016/j.eswa.2011.06.027

Huang, Z., 1998. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Min. Knowl. Discov. 2, 283–304. https://doi.org/10.1023/A:1009769707641

Iliou, T., Konstantopoulou, G., Lymperopoulou, C., Anastasopoulos, K., Anastassopoulos, G., Margounakis, D., Lymberopoulos, D., 2019. Iliou Machine Learning Data Preprocessing Method for Suicide Prediction from Family History. Presented at the 15th IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI), Springer International Publishing, p. 512. https://doi.org/10.1007/978-3-030-19823-7_43

Joseph, A., Ramamurthy, B., 2018. Suicidal behavior prediction using data mining techniques. IAEME 9, 293–301.

Kathriarachchi, S., Rajapakse, T., Seneviratne, L., Ferdinando, R., Ranaweera, S., Wijesundere, A., Samaraweera, S., Gunasekera, M., Jayaratne, K., Suveendran, T., Hettiarachchi, K., 2019. Suicide Prevention in Sri Lanka: Recommendations for Action. Sri Lanka Medical Association (SLMA) Expert Committee on Suicide Prevention.

KModes Clustering Algorithm for Categorical data, 2021. . Anal. Vidhya. URL https://www.analyticsvidhya.com/blog/2021/06/kmodes-clustering-algorithm-for-categorical-data/ (accessed 7.29.21).

Kumar, S., Toshniwal, D., 2015. A data mining framework to analyze road accident data. J. Big Data 2, 26. https://doi.org/10.1186/s40537-015-0035-y

Missing Values | Treat Missing Values in Categorical Variables, 2021. . Anal. Vidhya. URL https://www.analyticsvidhya.com/blog/2021/04/how-to-handle-missing-values-of-categorical-variables/ (accessed 7.29.21).

Mohd Shaharanee, I.N., Hadzic, F., Dillon, T., 2009. Interestingness of Association Rules Using Symmetrical Tau and Logistic Regression. pp. 422–431. https://doi.org/10.1007/978-3-642-10439-8_43

Omprakash, L.M., 2013. Data Mining Tool for Prediction of Suicides among Students, in: National Conference on New Horizons in IT. pp. 178–181.

Pai, K.S.K. nbsp and K.K., 2019. Determining Frequent Item Sets using Partitioning Technique for Large Transaction Database. Indian J. Sci. Technol. 12, 1. https://doi.org/10.17485/ijst/2019/v12i3/140766

Pros and Cons of R Programming Language - Unveil the Essential Aspects! [WWW Document], 2019. . DataFlair. URL https://data-flair.training/blogs/pros-and-cons-of-r-programming-language/ (accessed 5.16.21).

Saini, B., 2021. K-Means and K-Modes Clustering Algorithm [WWW Document]. Medium. URL https://ai.plainenglish.io/k-means-and-k-modes-clustering-algorithm-4ff51395fa8d (accessed 5.16.21).

Suicide, 2021. . Wikipedia.

Suicide in Sri Lanka, 2018. . Wikipedia.

Suicide Rate by Country 2020 [WWW Document], 2020. URL https://worldpopulationreview.com/country-rankings/suicide-rate-by-country (accessed 12.3.20).

Thalagala, N., 2011. Suicide Trends in Sri Lanka 1880-2006; Social, Demographic and Geographical Variations. J. Coll. Community Physicians Sri Lanka 14, 24–32. https://doi.org/10.4038/jccpsl.v14i1.2945

The Reality of Suicide in Sri Lanka: Need for Data-driven analysis, 2017. . Groundviews. URL https://groundviews.org/2017/10/06/hype-and-reality-of-suicides-in-sri-lanka-need-for-data-driven-analysis/ (accessed 12.4.20).

Weiss, S.M., Indurkhya, N., 1997. Predictive Data Mining: A Practical Guide, 1st edition. ed. Morgan Kaufmann, San Francisco.

WHO, 2019. Suicide [WWW Document]. URL https://www.who.int/news-room/fact-sheets/detail/suicide (accessed 12.3.20).

World Health Organization, 2019. Suicide in the world: global health estimates. World Health Organization, Geneva.

# APPENDICES

## I. Finding the Optimal Number of Clusters Using BIC

```
#Import the package
library(LCAvarsel)

#Set the path of the working directory
setwd("E:/MSc in MCS/MCS3204-2nd Attempt/Implementation/kmode_R")

#Read the data set
suicideDf <- read.csv("all_without_oth.csv")

#Set the maximum number of clusters
k.max <- 10

#Calculate the BIC value for each cluster
fit <- sapply(1:k.max, function(k) {
                set.seed(100000)
                fitLCA(
                     suicideDf[,2:7],
                     G = k,
                     X = NULL,
                     ctrlLCA = controlLCA()
                )$bic
          })

#Plot the BIC values in a graph with respect to the number of clusters
plot(1:k.max, fit,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="BIC")
```

## II. K-Mode Clustering

```
#Import the package
library(klaR)

#Set the path of the working directory
setwd("E:/MSc in MCS/MCS3204-2nd Attempt/Implementation/kmode_R")

#Read the data set
suicideDf <- read.csv("all_without_oth.csv")

#Apply the K-mode clustering algorithm
cluster.results <-kmodes(
    suicideDf[,2:7], #Data set
    5, #Number of clusters
    iter.max = 10, #Number of iterations
    weighted = TRUE #Weighted version of the distance is used
)

#Assign cluster number to each record in a new column
cluster.output <- cbind(
    suicideDf,
    cluster.results$cluster
)

#Save as a .csv file
write.csv(
    cluster.output,
    file = "kmode_clusters.csv",
    row.names = TRUE
)
```

## III. Association Rule Mining Using Apriori Algorithm

```
#Import packages
library(arules)
library(arulesViz)

#Set the path of the working directory
setwd("E:/MSc in MCS/MCS32042nd Attempt/Implementation/association/cluster
1")

#Convert each record to a transaction
dataset = read.transactions(
    'kmode_cluster_1 - Copy.csv',
    sep = ',',
    rm.duplicates = TRUE
)

set.seed = 220

#Apply the Apriori algorithm with the minimum support=0.01 and confidence=
1
associa_rules = apriori(
    data = dataset,
    parameter = list(support = 0.01, confidence = 1)
)

#Sort the rules according to the lift
inspect(sort(associa_rules, by = 'lift'))

#Filter the rules which contain the “Reason for suicide” as consequent
reason_rules <- subset(
                associa_rules,
                subset=rhs %in% c('ECO','EMP','ELD','HAR','DISLUV','SEXHAR
','DRUG','AGRDEATH','PROLOSS','EXMFAIL','CHILD','SEXINC','MENDIS','DISEASE
')
            )

#Sort the rules according to the lift value
inspect(sort(reason_rules, by = 'lift'))
```