FINDING CAUSES OF SMOKING USING ASSOCIATION ANALYSIS

BY

MANASH KUMAR MONDAL
ID: 161-15-7245

AND

TAUKIR AHMMED
ID: 161-15-7165

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Md. Tarek Habib
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

OCTOBER, 2020

APPROVAL

This Project titled "**Finding Causes of Smoking using Association Analysis**", submitted by Manash Kumar Mondal (ID- 161-15-7245) and Taukir Ahmmed (ID- 161-15-7165) to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 7th to 8th October, 2020.

BOARD OF EXAMINERS

Dr. Syed Akhter Hossain **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Azmoon

Nazmun Nessa Moon **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Zahir

Gazi Zahirul Islam **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Saddam

Dr. Md. Saddam Hossain **External Examiner**
Assistant Professor
Department of Computer Science and Engineering
United International University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Md. Tarek Habib, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

THabib

Md. Tarek Habib
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

ManashKumarMondal

Manash Kumar Mondal
ID: 161-15-7245
Department of CSE
Daffodil International University

Ahmed

Taukir Ahmmed
ID: 161-15-7165
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Md. Tarek Habib, Assistant Professor,** Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine learning*" to carry out this project. His endless patience ,scholarly guidance ,continual encouragement, constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Dr. Syed Akhter Hossain, Professor, and Head of the Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Smoking is one of the most prominent causes of illness and death around the world. Throughout the 21st century, death attributes to smoking are projected to rise substantially and much of the increase will occur in developing countries such as Bangladesh. Based on a study over 1 lakh deaths in Bangladesh every year and second-hand smoke causes 24,757 deaths last year. It also causes an economic loss for our country due to health hazards. The young generation is the biggest share of tobacco user in Bangladesh. People are enough to concern about the bad effect of smoking, however, the number of smokers also growing in the same manner. In Bangladesh most the tobacco user is men and they have started smoking at an early age. So we need to find the reason for starting smoking. In our study, we surveyed the student of Daffodil International University to getting information about the reason for smoking. After that, we used the Data Mining technique to get a summary. We did association rule mining using the Apriori algorithm for finding the relationship between factors.

TABLE OF CONTENTS

CHAPTER

CHAPTER 1: INTRODUCTION

CHAPTER 2: BACKGROUND

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Introduction

Tobacco smoking is the practice of burning tobacco and ingesting the smoke that is produced. The smoke may be breathed as is done with cigarettes, or simply released from the mouth, as is commonly done with pipes and cigars [1]. On average, the life expectancy of a smoker is 10 years less than a nonsmoker [2]. Cigarette smoking is the main source of SHS exposure because it is the most prevalent form of tobacco smoking although specific differences between countries. Tobacco smoke contains thousands of chemicals that are released through burning as gases, vapors, and particles. Both smoking and passive smoking are the biggest causes of death of Bangladesh [3]. Over 70 lakh people suffering from various tobacco related diseases [4]. More than half of Bangladeshi men over the age of 25 years smoke cigarettes or biris [5]. The government takes many steps to control tobacco use but the overall number of users didn’t reduce that much. Most people started smoking at a young age. There are many influencing factors related to started to smoking which differs from various people

1.2 Motivation

According to this study, male smoking prevalence was 60.01% and male smokeless tobacco use prevalence was 21.35%. that GATS found that 23.0% of adults aged 15 years or above were smokers of tobacco in Bangladesh [3]. Most smokers started smoking while they were teen [6]. There is so many influence factor related to staring smoking. These factors are highly related to a person’s characteristics. And that is the motivation for us to conduct the research and use of Machine learning to specifically association analysis find the causes of smoking as this percent of people are involved in smoking.

1.3 Rationale of the Study

As number of smokers is growing day by day, so need to find the root of that problem. Anyone who starts using tobacco can become addicted to nicotine [6]. Studies show that smoking is most likely to become a habit during the teen years [6]. The younger you are when you begin to smoke, the more likely you are to become addicted to nicotine. So we think we have to address that problem.

This research is focused on the objectives of obtaining “why” people are tempted to smoking and continuing smoking employing the association rule mining method that shows us the interconnection among factors. We want to make this study for scholarly purpose

1.4 Research Dataset

For this study, we have used our own dataset. Since we didn't find any available dataset as per our need. We have decided to make our own and collected data using a survey physically. All the data were taken from the DIU Main Campus students. Our raw dataset contains 1011 responses and 26 columns.

1.5 Research Questions

We haven’t found relevant research on this topic. Our main objective is to find the relationship between factors based on responses from both smokers and non-smokers and also test to see how the model works on our dataset.

1.6 Expected Outcome

The task of mining association rules consists of finding frequent item-sets and generating high confidence among them [7]. The objective of our study is to find reasons/ influences which are related to making interest to start smoking. It will help to find patterns among factors. As every person is different, so their influencing factors will differ. Our study will show rules that contain high confidence value.

Here we present the experiment results of the Apriori algorithm on our survey datasets which will encourage others to apply rule mining by making a dataset for solving a particular problem. Besides we believe our study will help people working on preventing smoking.

1.7 Layout of Report

We have divided our report into five chapters so that it will be well organized to examiners. Each chapter has a particular purpose to explain our research works. The chapter-wise reporting summary are as follows

- Chapter 1: It contains an introductory part of our study. The motivation, rationale of the study are discussed in this segment.
- Chapter 2: This chapter explained the background knowledge of association rule mining. Why we adopted the Apriori algorithm, relevant works, and difficulties of the study
- Chapter 3: Here we discussed how we collected our data and did it digitized. Then how we have cleaned, checked missing and null values, and pre-processed our data. On the other hand why we have decided to adopt the Apriori algorithm
- Chapter 4: In this chapter, we have presented study results and did a concise analysis of the results.
- Chapter 5: Here we have addressed the impact of our study on society and environments
- Chapter 6: We have done the conclusion, future works, and project scope to enhance our works

CHAPTER 2
BACKGROUND STUDY

2.1 Introduction

Those who have friends or parents who smoke are more prone to start smoking than those who don’t [6]. Some teens state that they “just desired to try it,” or they believed it was “cool” to smoke [6]. Social media, TV ads, and other promotional videos are influencing people to start smoking. Movies and cinema are also playing a high role in addicted to smoking.

Besides some personal and psychological aspects also influence an individual to start smoking. Young people are the most sufferers of those influences. So perceiving association with distinctive item-sets, we can find some comprehensive insight into the causes of smoking.

2.2 Related Works

This section of this paper is focused on other near past works is done by other researchers on the several problems for rule mining. Their works guided us to understand the process, methods, and way to perform our study.

Association Rule Mining is a common and well-researched method for exploring intriguing relations among variables in extensive databases and to pick exciting rules from the set of all potential rules, constraints on various degrees of significance and concern can be used [8].

Bioinformatics is one of the prominent domains of data mining. By using associating rule mining can find patterns that explain the correlation between the binary attributes [9][10][11]. One of the most prominent biological data is the functions and other properties of proteins at a genomic order are protein interaction networks [12]. Protein function relationship can be found by using an association rule base. The most prominent application is on market basket analysis. According to transaction data, one can predict the frequent item-set people buy for a better recommendation system [13] and also can analysis a consumers behavior [14]. Data analyzed in this case consist of all buying transaction of product on a certain unit of time and analyzed them to find structure sales of different product available [14].Main aim is find recurrent rules within transactions[15].

Analyze and mine knowledge on significant factors causing infertility in women through Frequent Item-set Mining have been used. Even, there are a number of factors causing infertility in women, only three significant factors namely Age, Body Mass Index and Thyroid Stimulating Hormone Levels during prenatal periods have been taken for analysis [16].

2.3 Research Summary

In rule mining most of the research performed on recommendation system and biological data such as market basket analysis and Patterns of Numerously Occurring Heart Diseases[17].

In our research paper, we have done our investigation on a dataset having numerous factors that are associated with smoking. We use the Apriori algorithm to generate frequent itemsets, filter out using our desire confidence, and sort them according to support and confidence.

2.4 Scope of the Problem

Our study is based on rule mining. Here we generate only interesting correlated factors but not a prediction. So there is a scope to foretell smoker and non-smoker using these factors, frequent itemsets, and existence dataset.

2.5 Challenges

We have faced several challenges for completing our research study. These are follows

- There isn’t any existence dataset
- Selecting factors for doing survey

- Collecting responses from people physically
- Inputting these paper data into excel file
- Selecting the algorithm

CHAPTER 3

Research Methodology

3.1 Data Collection

Since there is not any available dataset, we have make our own. Firstly we finalized the factors for causes of smoking. We print the multiple choice based paper form and it contains 26 individual questions. The number of choice and choose option varies according to question.

Table 3.1: All question for survey with options

No	Question	Options
1	gender?	Male Female
2	smoker?	Yes No
3	smoke first time age?	13-16 16-19 19-25 25 plus
4	academic performance started smoking?	Excellent Very Good Satisfactory Fair Poor
5	attachment to school started smoking?	Excellent Very Good Satisfactory Fair Poor

No	Question	Options
6	self-regulation skills started smoking?	Excellent Very Good Satisfactory Fair Poor
7	allow watch age-restricted movies?	Yes No
8	influnce tobacco advertisements?	Yes No
9	favorite film star smokes on screen?	Yes No
10	smoking scenes in film?	Yes No
11	watching people smoking?	Yes No
12	watching family member is smoking?	Yes No
13	Influencing factor starting cigarette smoking?	Friends influence Father's influence Brother influence Uncle's influence Grandfather influence Female family member influence
14	personality characteristics?	Impulsivity Rebelliousness Risk taking property Self esteem Sensation seeking Problematic interpersonal relationship

No	Question	Options
15	Continuing cigarette smoking?	Mental depression Bad family relation Education problem Friend’s circle Difficulties with girlfriend Difficulties with boyfriend
16	Smoking helps?	Sadness Loneliness Boringness Depression Working pressure To feel cool
17	do your family know smoking status?	Yes No
18	do your family monitor about your smoking habit?	Yes No
19	Should student smoke?	Yes No
20	any trouble in school?	Yes No
21	curiosity about smoking?	Yes No
22	intention to smoke in future?	Yes No
23	influence other for smoking?	Yes No

No	Question	Options
24	guardians educational level?	College/university/tertiary No formal education Secondary / high school Primary / vocational training Diploma
25	educational status when you started?	Secondary Primary University Illiterate
26	reasons or influences of start smoking?	Watching tv/cinema Family influence To feel mature Performance in class To attract girls To attract boys To follow seniority Poor

We have collected our data from the students of DIU Main Campus. We collected data physically. After that make these paper data into excel formal by imputing data manually.

3.2 Data Preprocessing

3.2.1 Data cleaning

Some of data contains missing values. So we have found those values by using pandas library and we remove these using excel filter method. We also remove extra space in value. Tools we used for data cleaning are-

- Numpy Library
- Pandas Library
- Microsoft Excel

3.2.2 Data Preprocessing

As we handle null values and other validation. Then we did preparing our data frame for applying algorithm. We filtered the data into smoker and non-smoker data frame. Then again we filtered out male and female data of smokers. After that we do some filtering for various types rule mining. For this processing purpose we used python3 pandas library.

3.3 Data Overview

We have collected data from both male and female. About 75.76% male and 24.03% female are responded to our survey.

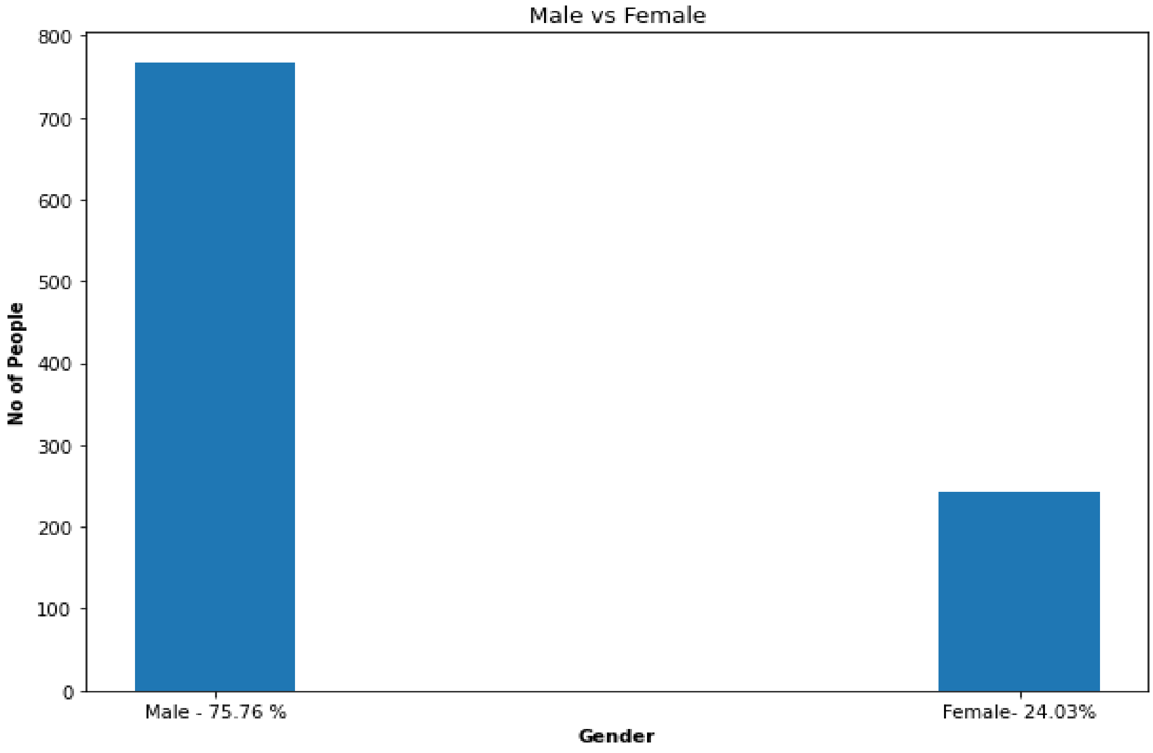


Figure **3.1**: Male vs Female responses

In our dataset we have found the smoker and non-smoker records are respectively 482 and 527. So about 47.6% are smokers' records and rest of the 52.4% are non-smokers records.

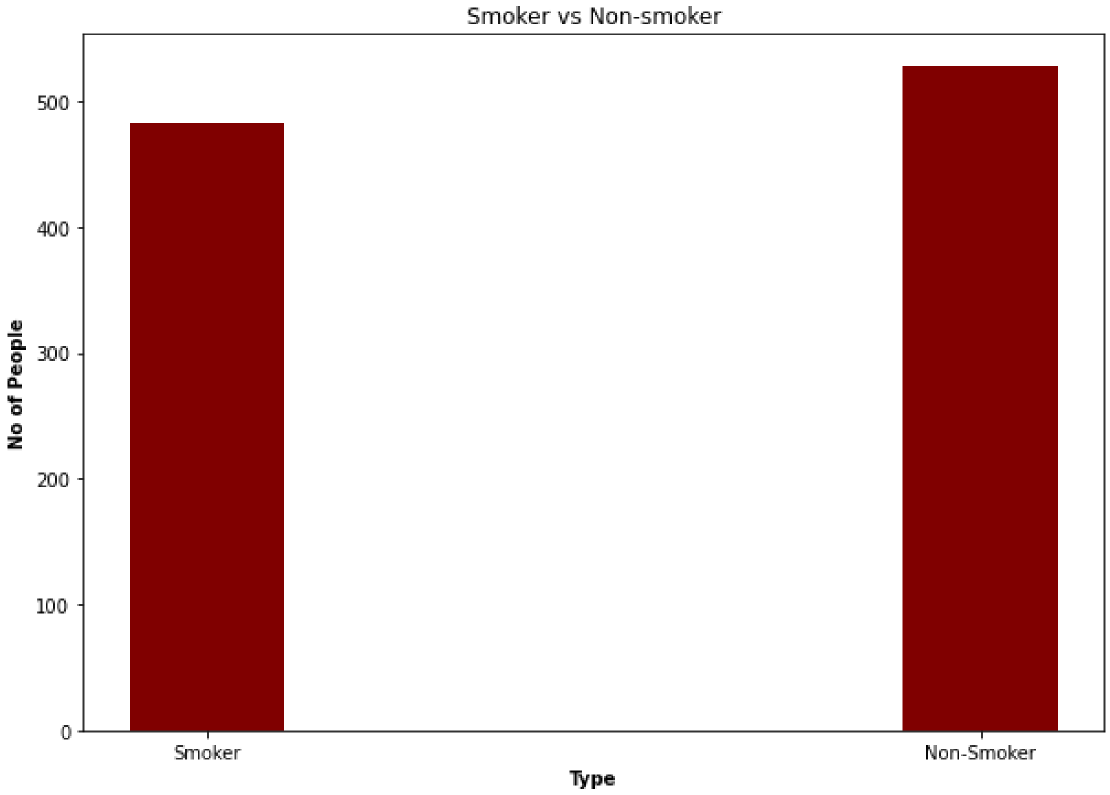


Figure 3.2 Number of Smoker and non-smoker

The following graphs shows what both smoker and non-smoker thinks about student should smoker or not. We have find that majority people think student shouldn’t smoke.

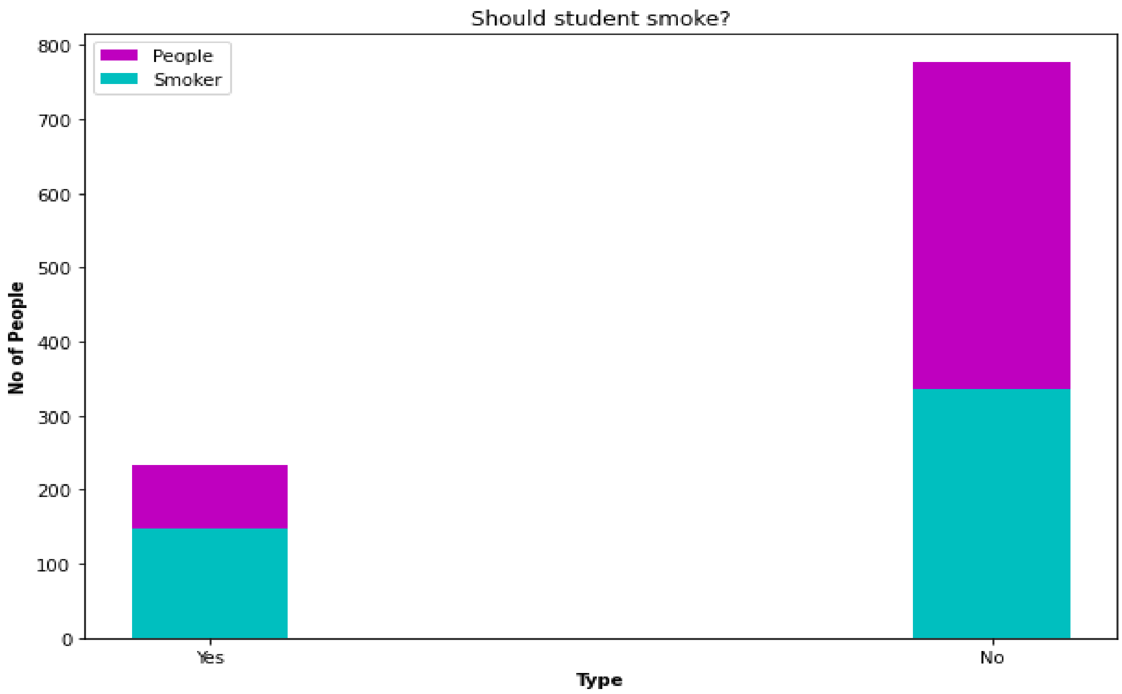


Figure 3.3 People thinks student should smoke or not

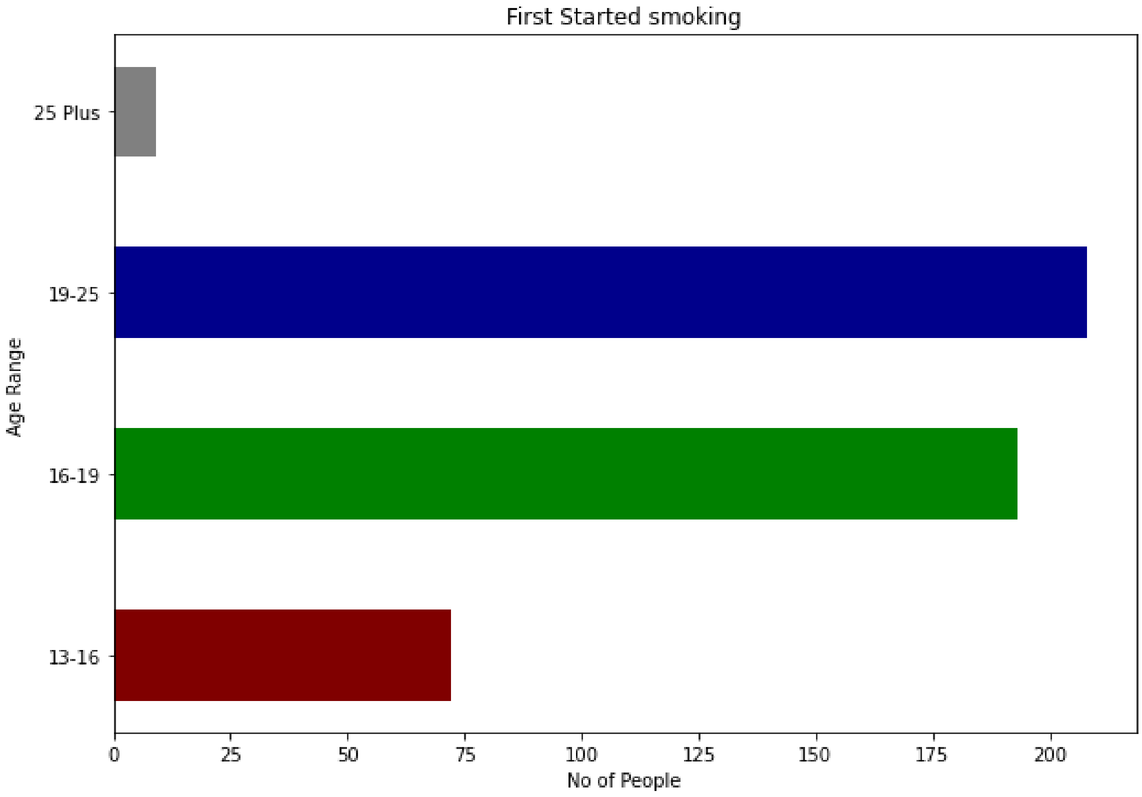


Figure 3.4 Age distribution of first time smoke

The Previous graphs shows the age distribution of all of our records when they started smoking. We have find that most the people started smoking at the age between 19-25 ages.

The following graph shows the percent of people having depression. From that analysis about 24% people have depression and almost 50% smoker have depression.

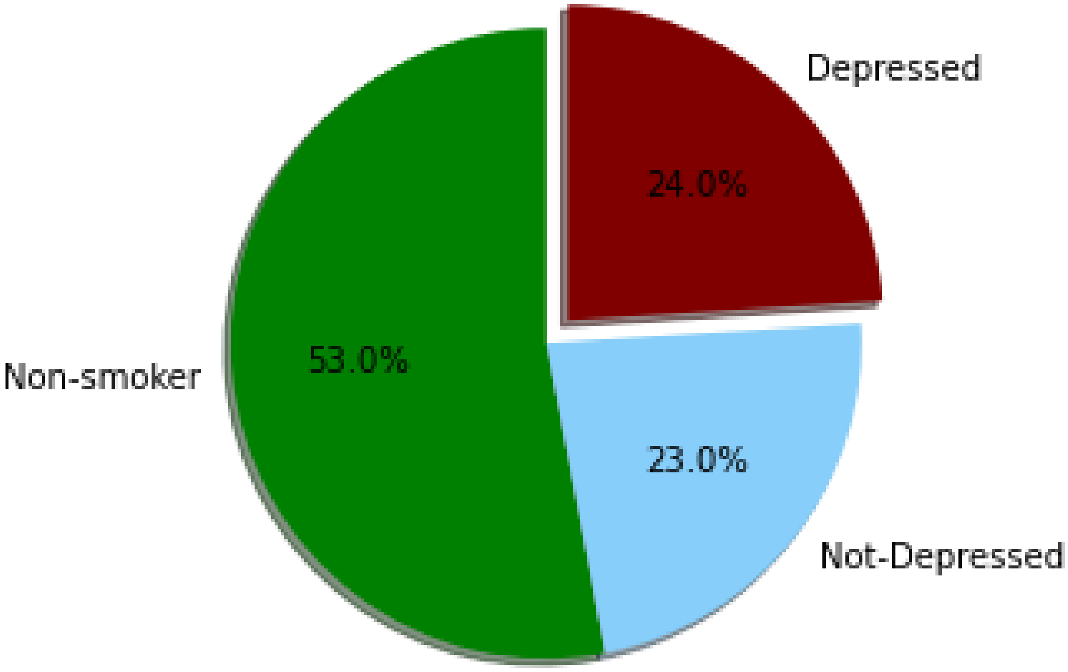


Figure 3.5 Percent of smokers are in depression

This time we have done an analysis what was their educational status when started smoking. Most the people around 57.3% people started smoking when they are on secondary level. Also we have found and interesting data that indicated 34% people do their first attempt when they are on their primary education level.

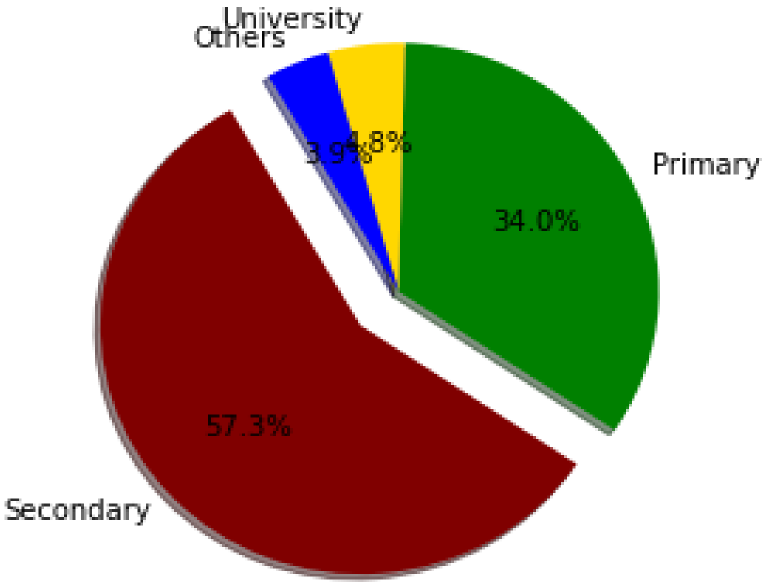
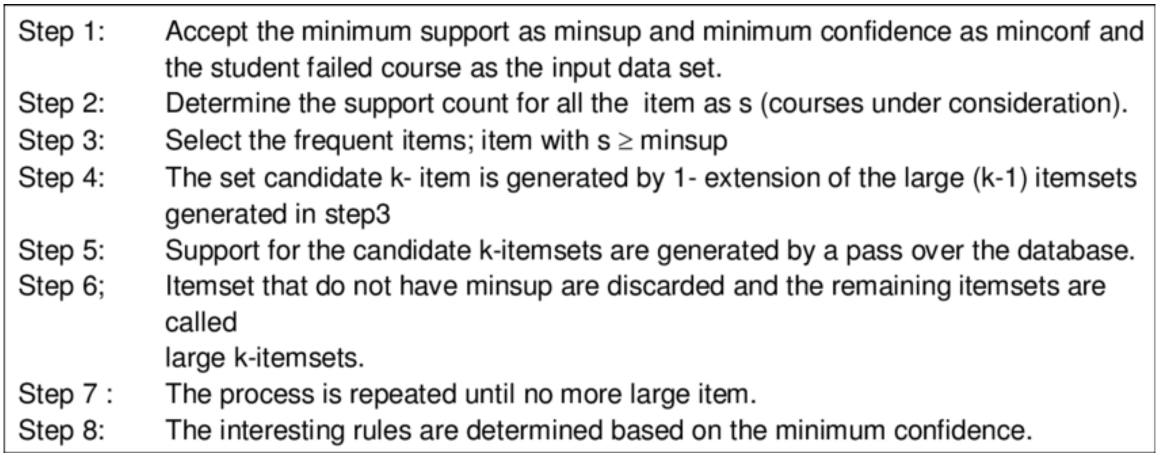


Figure 3.6 Educational status when started smoking

3.3 Proposed Methodology

This research make extensive us of association rule mining [18]. It is a very popular data mining technique. It identifies frequent patterns and associations (relations) among a set of items. Ex: If you go to buy a keyboard, you might also get a mouse. So place them aside in your market to get more profit. For that purpose you have to assumption how likely a customer buy keyboard and then mouse also.

An association rule consists of two parts, an antecedent (if) and a consequent (then) as it is similar to if/then statement [19]. There are two different algorithm for mining association rule, Apriori and FP-Growth. General Pseudo code is given below



```
Step 1:   Accept the minimum support as minsup and minimum confidence as minconf and
          the student failed course as the input data set.
Step 2:   Determine the support count for all the  item as s (courses under consideration).
Step 3:   Select the frequent items; item with s ≥ minsup
Step 4:   The set candidate k- item is generated by 1- extension of the large (k-1) itemsets
          generated in step3
Step 5:   Support for the candidate k-itemsets are generated by a pass over the database.
Step 6;   Itemset that do not have minsup are discarded and the remaining itemsets are
          called
          large k-itemsets.
Step 7 :  The process is repeated until no more large item.
Step 8:   The interesting rules are determined based on the minimum confidence.
```

Figure 3.7 General Pseudo code of association rule mining [20]

In this study we will use Apriori algorithm because it’s much suitable for our dataset. It is seminal algorithm for mining frequent itemsets for Boolean association rules [10]. It uses prior knowledge of frequent properties that why it was named Apriori.

CHAPTER 4

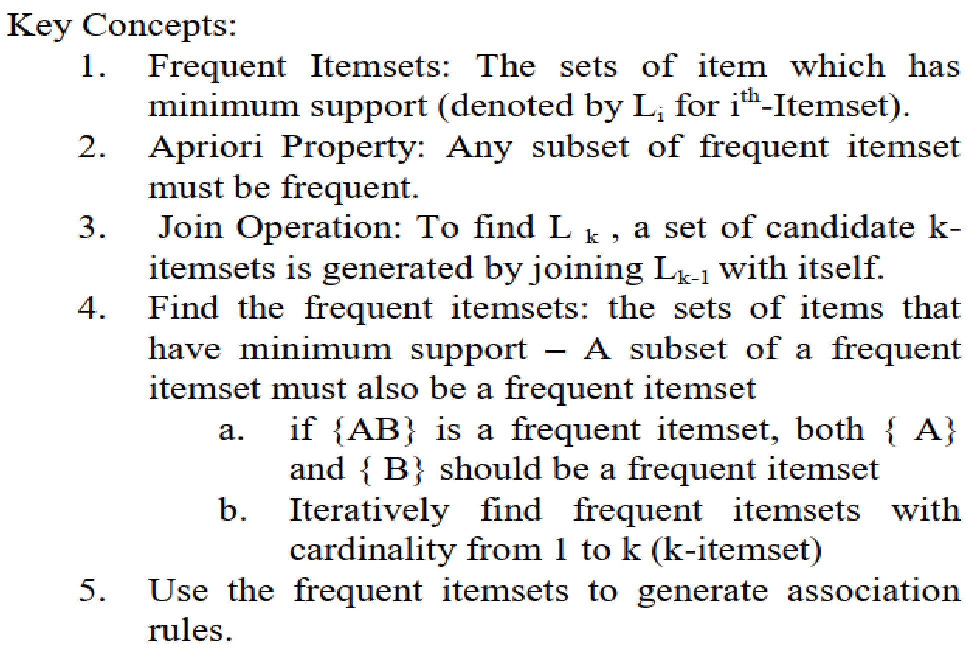
Experimental Evaluation

4.1 Introduction:

Apriori algorithm is our key algorithm in this study. It has basically 3 steps. First find frequent itemsets which has minimum support, then from these frequents sets generate association rules and find the correlation between these rules to filter out uninterested rules [21].

4.2 Algorithm Implementation:

Apriori is a Machine Learning algorithm that is used to gain insight into the structured relationships between different items involved. It's a data mining technique that is used for mining frequent itemsets and relevant association rules. For example, a customer tends to buy a keyboard and a mouse at the same time can be found using Apriori rule mining. The key concepts of the algorithm is given below



Key Concepts:

1. Frequent Itemsets: The sets of item which has minimum support (denoted by L_i for i^{th}-Itemset).
2. Apriori Property: Any subset of frequent itemset must be frequent.
3. Join Operation: To find L_k , a set of candidate k-itemsets is generated by joining L_{k-1} with itself.
4. Find the frequent itemsets: the sets of items that have minimum support – A subset of a frequent itemset must also be a frequent itemset
   a. if {AB} is a frequent itemset, both { A} and { B} should be a frequent itemset
   b. Iteratively find frequent itemsets with cardinality from 1 to k (k-itemset)
5. Use the frequent itemsets to generate association rules.

Figure 4.1 Key Concepts of Apriori Algorithm

Support:

Support refers to the default popularity of an item and can be calculated by finding the number of transactions containing a particular item divided by total number of transactions

$$\text{Support (Keyboard)} = \frac{\text{Transactions containing (Keyboard)}}{\text{Total Transactions}}$$

Confidence:

Confidence refers to the likelihood that an item B (mouse) is also bought if item A (keyboard) is bought. Like our keyboard and mouse example.

$$\text{Confidence}(\text{Keyboard} \rightarrow \text{Mouse}) = \frac{\text{Transactions containing Keyboard and Mouse}}{\text{Transactions containing Keyboard}}$$

Correlation Analysis:

As support and confidence are not enough for filtering interesting rule, so overcome this weakness, a correlation measure can be used [21]. This leads to correlation rules of the form

$$A \Rightarrow B\ [support, confidence, correlation]$$

Lift can be used to measure correlation. As example Lift (Keyboard -> Mouse) refers to the increase in the ratio of sale of Mouse when the Keyboard is sold. Lift (Keyboard ⇒ Mouse) can be calculated by dividing Confidence (Keyboard → Mouse) divided by Support (Mouse).

$$\text{Lift}(\text{Keyboard} \rightarrow \text{Mouse}) = \frac{\text{Confidence (Keyboard} \rightarrow \text{Mouse)}}{\text{Support (Mouse)}}$$

If the value of lift is less than 1, they are negatively correlated, if greater than 1 then they are positively correlated or otherwise there is no correlation between them [21].

CHAPTER 5

Results and Discussion

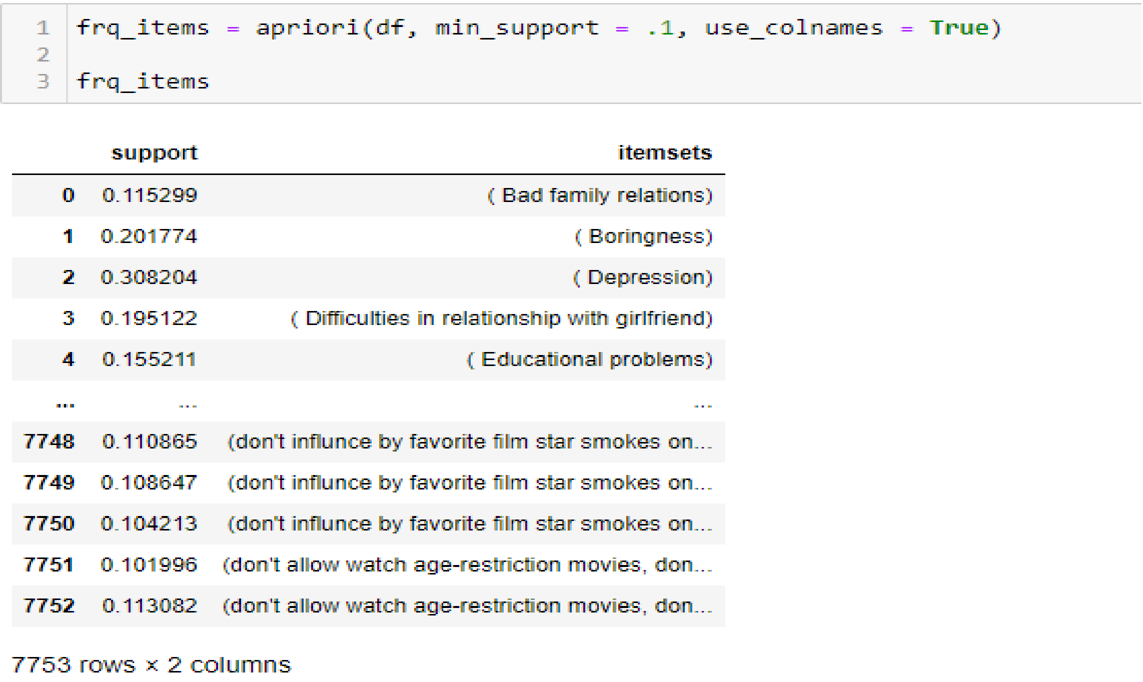
5.1 Introduction

This section discussed about the practical findings on dataset and the overall outcome the project. All different specific rule mining will be disclosed here and finding the cause of smoking for different scenario type data.

5.2 Experimental Result

In our study we used Python “mlxtend” library to apply apriori

Rule 1: Mining only for male smoker using all possible factors except “should student smoke”, “influence other for smoking”, “guardian education level” and your education level’ to generate rules. The frequents itemsets and generated rules are given below

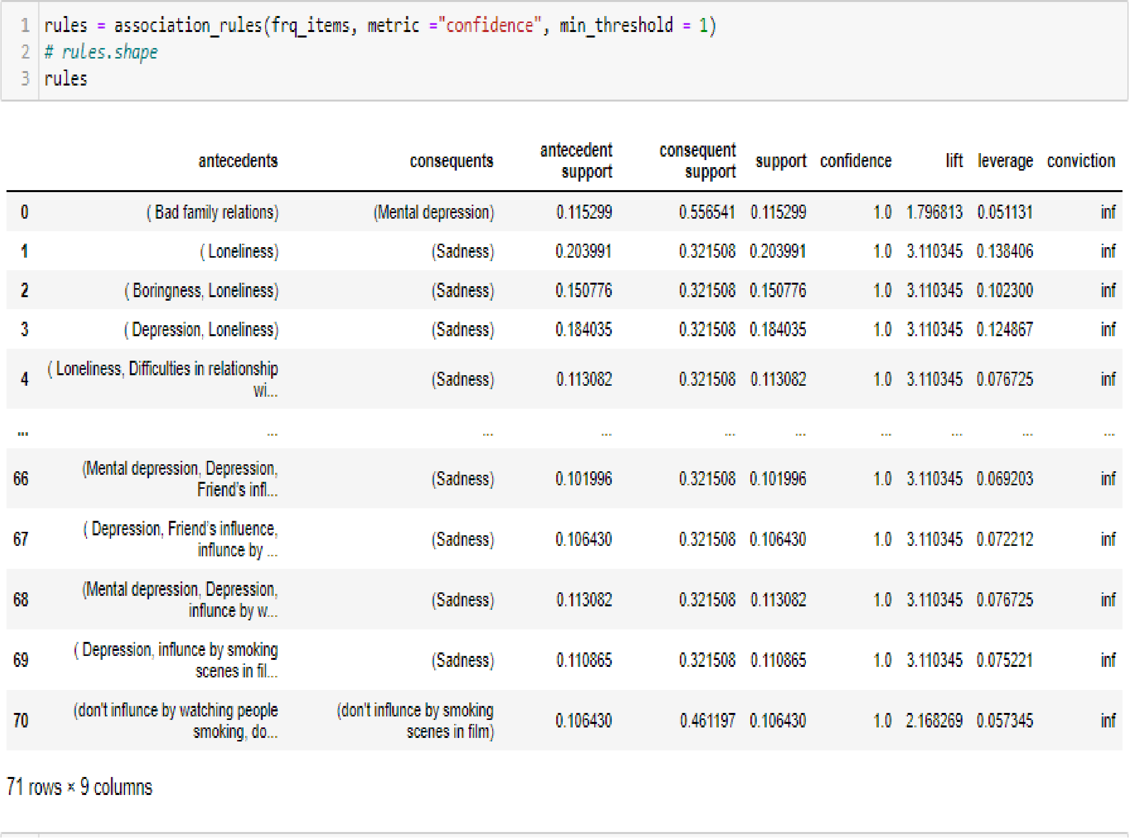


```
frq_items = apriori(df, min_support = .1, use_colnames = True)

frq_items
```

	support	itemsets
0	0.115299	(Bad family relations)
1	0.201774	(Boringness)
2	0.308204	(Depression)
3	0.195122	(Difficulties in relationship with girlfriend)
4	0.155211	(Educational problems)
...	...	...
7748	0.110865	(don't influnce by favorite film star smokes on...
7749	0.108647	(don't influnce by favorite film star smokes on...
7750	0.104213	(don't influnce by favorite film star smokes on...
7751	0.101996	(don't allow watch age-restriction movies, don...
7752	0.113082	(don't allow watch age-restriction movies, don...

7753 rows × 2 columns

Figure 5.1: Male smoker with all possible factors frequent itemsets



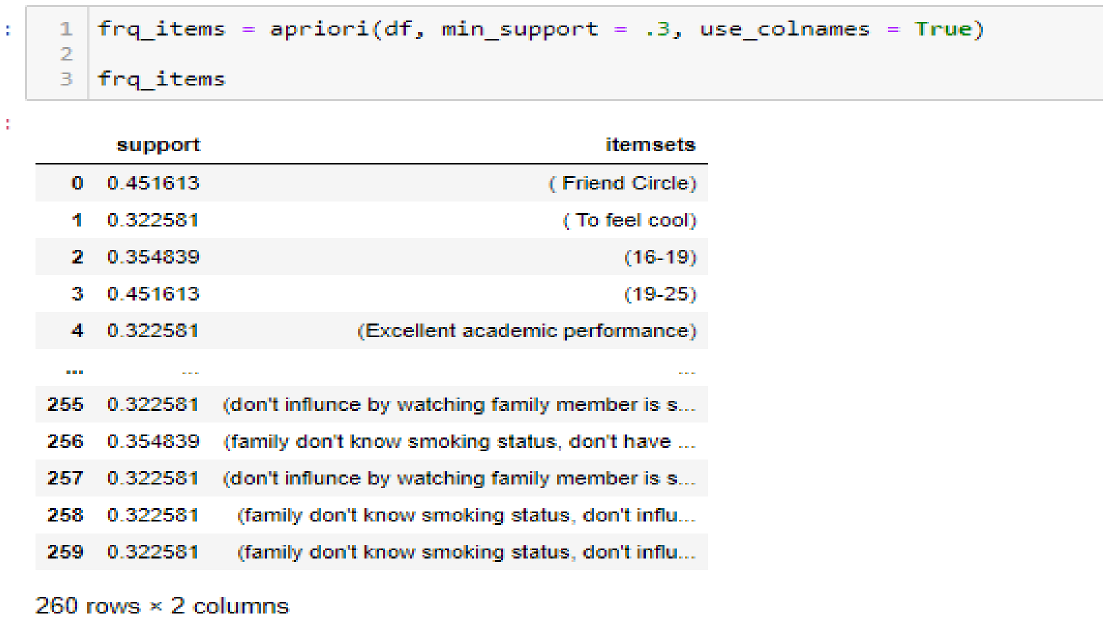
```python
rules = association_rules(frq_items, metric ="confidence", min_threshold = 1)
# rules.shape
rules
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(Bad family relations)	(Mental depression)	0.115299	0.556541	0.115299	1.0	1.796813	0.051131	inf
1	(Loneliness)	(Sadness)	0.203991	0.321508	0.203991	1.0	3.110345	0.138406	inf
2	(Boringness, Loneliness)	(Sadness)	0.150776	0.321508	0.150776	1.0	3.110345	0.102300	inf
3	(Depression, Loneliness)	(Sadness)	0.184035	0.321508	0.184035	1.0	3.110345	0.124867	inf
4	(Loneliness, Difficulties in relationship wi...	(Sadness)	0.113082	0.321508	0.113082	1.0	3.110345	0.076725	inf
...	...	...	...	...	...	...	...	...	...
66	(Mental depression, Depression, Friend's infl...	(Sadness)	0.101996	0.321508	0.101996	1.0	3.110345	0.069203	inf
67	(Depression, Friend's influence, influnce by ...	(Sadness)	0.106430	0.321508	0.106430	1.0	3.110345	0.072212	inf
68	(Mental depression, Depression, influnce by w...	(Sadness)	0.113082	0.321508	0.113082	1.0	3.110345	0.076725	inf
69	(Depression, influnce by smoking scenes in fil...	(Sadness)	0.110865	0.321508	0.110865	1.0	3.110345	0.075221	inf
70	(don't influnce by watching people smoking, do...	(don't influnce by smoking scenes in film)	0.106430	0.461197	0.106430	1.0	2.168269	0.057345	inf

71 rows × 9 columns

Figure 5.2: Male smoker with all possible factors generated rules

Rule 2: Mining only for female smoker using all possible factors except "should student smoke", "influence other for smoking", "guardian education level" and your education level'

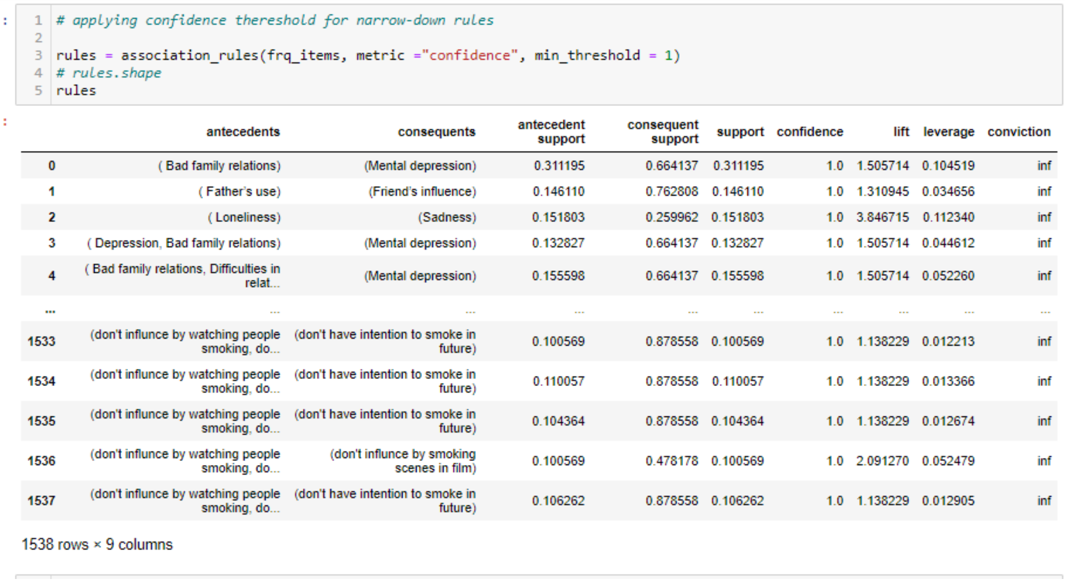


```
frq_items = apriori(df, min_support = .3, use_colnames = True)

frq_items
```

	support	itemsets
0	0.451613	(Friend Circle)
1	0.322581	(To feel cool)
2	0.354839	(16-19)
3	0.451613	(19-25)
4	0.322581	(Excellent academic performance)
...	...	...
255	0.322581	(don't influnce by watching family member is s...
256	0.354839	(family don't know smoking status, don't have ...
257	0.322581	(don't influnce by watching family member is s...
258	0.322581	(family don't know smoking status, don't influ...
259	0.322581	(family don't know smoking status, don't influ...

260 rows × 2 columns

Figure 5.3: Female smoker with all possible factors frequent itemsets



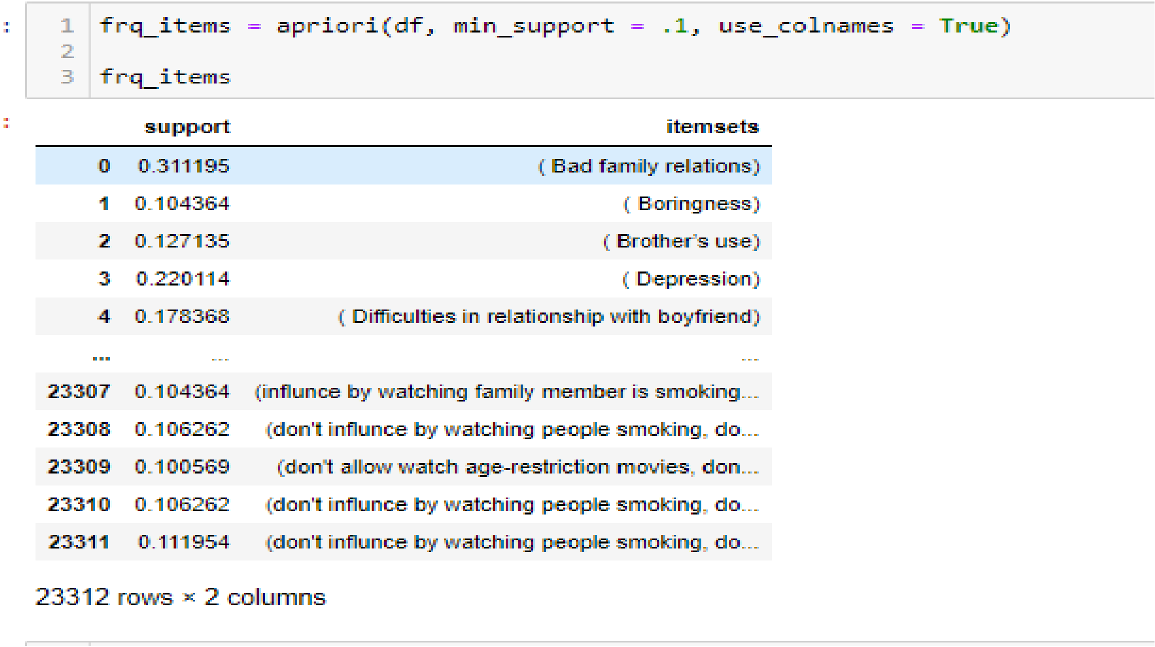
```
# applying confidence thereshold for narrow-down rules

rules = association_rules(frq_items, metric ="confidence", min_threshold = 1)
# rules.shape
rules
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(Bad family relations)	(Mental depression)	0.311195	0.664137	0.311195	1.0	1.505714	0.104519	inf
1	(Father's use)	(Friend's influence)	0.146110	0.762808	0.146110	1.0	1.310945	0.034656	inf
2	(Loneliness)	(Sadness)	0.151803	0.259962	0.151803	1.0	3.846715	0.112340	inf
3	(Depression, Bad family relations)	(Mental depression)	0.132827	0.664137	0.132827	1.0	1.505714	0.044612	inf
4	(Bad family relations, Difficulties in relat...	(Mental depression)	0.155598	0.664137	0.155598	1.0	1.505714	0.052260	inf
...	...	...	...	...	...	...	...	...	...
1533	(don't influnce by watching people smoking, do...	(don't have intention to smoke in future)	0.100569	0.878558	0.100569	1.0	1.138229	0.012213	inf
1534	(don't influnce by watching people smoking, do...	(don't have intention to smoke in future)	0.110057	0.878558	0.110057	1.0	1.138229	0.013366	inf
1535	(don't influnce by watching people smoking, do...	(don't have intention to smoke in future)	0.104364	0.878558	0.104364	1.0	1.138229	0.012674	inf
1536	(don't influnce by watching people smoking, do...	(don't influnce by smoking scenes in film)	0.100569	0.478178	0.100569	1.0	2.091270	0.052479	inf
1537	(don't influnce by watching people smoking, do...	(don't have intention to smoke in future)	0.106262	0.878558	0.106262	1.0	1.138229	0.012905	inf

1538 rows × 9 columns

Figure 5.4: Female smoker with all possible factors generated rules

Rule 3: What's is thinking of non-smoker (both male and female) people the causes of smoking using factors all except "should student smoke", "influence other for smoking", "guardian education level" and your education level'

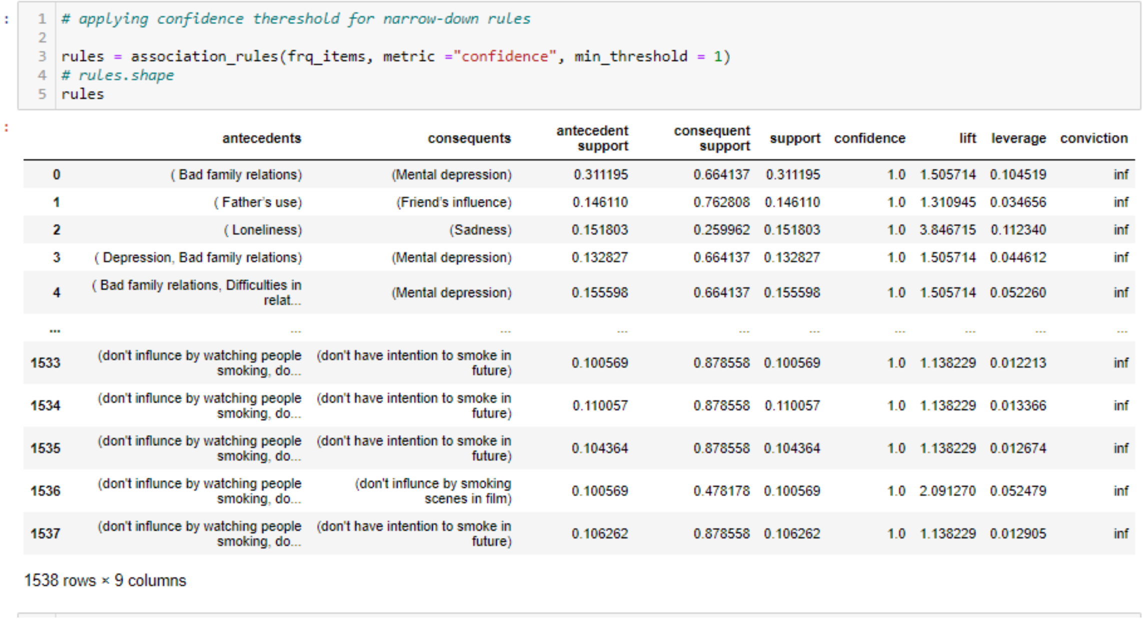


```
frq_items = apriori(df, min_support = .1, use_colnames = True)

frq_items
```

	support	itemsets
0	0.311195	(Bad family relations)
1	0.104364	(Boringness)
2	0.127135	(Brother's use)
3	0.220114	(Depression)
4	0.178368	(Difficulties in relationship with boyfriend)
...	...	...
23307	0.104364	(influnce by watching family member is smoking...
23308	0.106262	(don't influnce by watching people smoking, do...
23309	0.100569	(don't allow watch age-restriction movies, don...
23310	0.106262	(don't influnce by watching people smoking, do...
23311	0.111954	(don't influnce by watching people smoking, do...

23312 rows × 2 columns

Figure 5.5: Both male and female smoker with all possible factors frequent itemsets



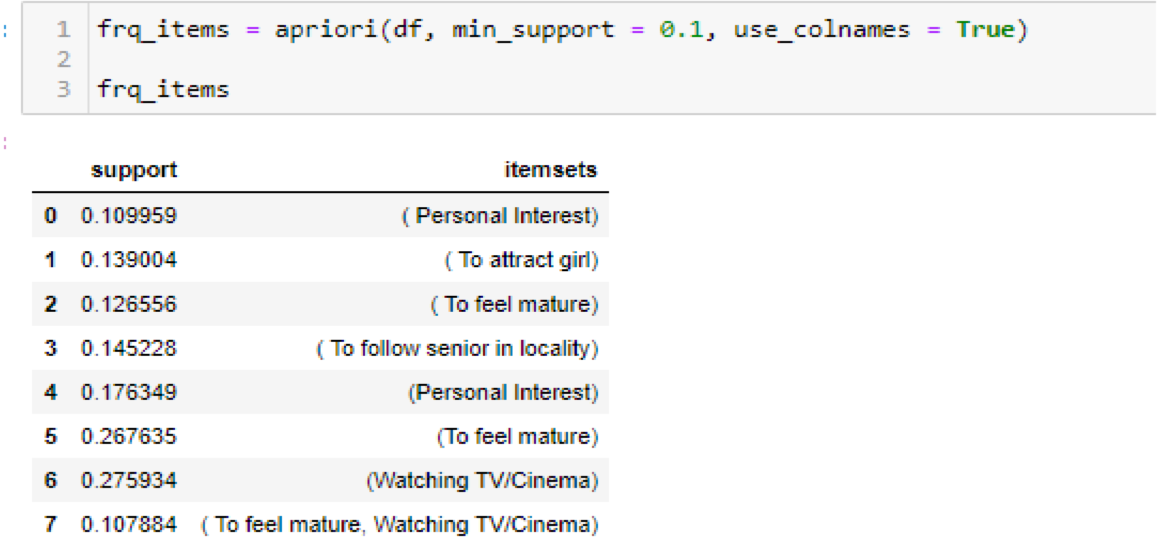
```
# applying confidence thereshold for narrow-down rules

rules = association_rules(frq_items, metric ="confidence", min_threshold = 1)
# rules.shape
rules
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(Bad family relations)	(Mental depression)	0.311195	0.664137	0.311195	1.0	1.505714	0.104519	inf
1	(Father's use)	(Friend's influence)	0.146110	0.762808	0.146110	1.0	1.310945	0.034656	inf
2	(Loneliness)	(Sadness)	0.151803	0.259962	0.151803	1.0	3.846715	0.112340	inf
3	(Depression, Bad family relations)	(Mental depression)	0.132827	0.664137	0.132827	1.0	1.505714	0.044612	inf
4	(Bad family relations, Difficulties in relat...	(Mental depression)	0.155598	0.664137	0.155598	1.0	1.505714	0.052260	inf
...	...	...	...	...	...	...	...	...	...
1533	(don't influnce by watching people smoking, do...	(don't have intention to smoke in future)	0.100569	0.878558	0.100569	1.0	1.138229	0.012213	inf
1534	(don't influnce by watching people smoking, do...	(don't have intention to smoke in future)	0.110057	0.878558	0.110057	1.0	1.138229	0.013366	inf
1535	(don't influnce by watching people smoking, do...	(don't have intention to smoke in future)	0.104364	0.878558	0.104364	1.0	1.138229	0.012674	inf
1536	(don't influnce by watching people smoking, do...	(don't influnce by smoking scenes in film)	0.100569	0.478178	0.100569	1.0	2.091270	0.052479	inf
1537	(don't influnce by watching people smoking, do...	(don't have intention to smoke in future)	0.106262	0.878558	0.106262	1.0	1.138229	0.012905	inf

1538 rows × 9 columns

Figure 5.6: Both male and female smoker with all possible factors generated rules

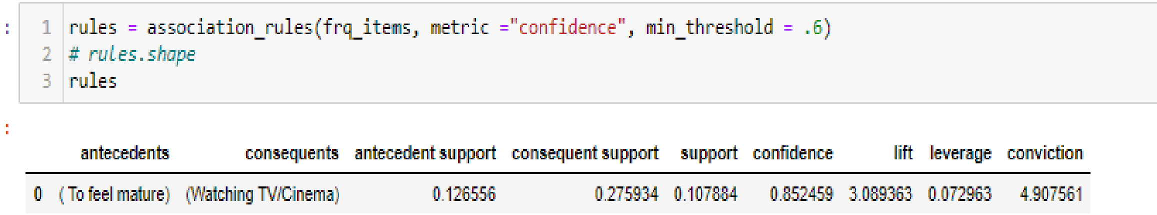
Rule 4: Only “Reasons or influences of starting smoking” for smoker both male and female using all possible factors except “should student smoke”, “influence other for smoking”, “guardian education level” and your education level’ to generate rules



```
frq_items = apriori(df, min_support = 0.1, use_colnames = True)

frq_items
```

	support	itemsets
0	0.109959	(Personal Interest)
1	0.139004	(To attract girl)
2	0.126556	(To feel mature)
3	0.145228	(To follow senior in locality)
4	0.176349	(Personal Interest)
5	0.267635	(To feel mature)
6	0.275934	(Watching TV/Cinema)
7	0.107884	(To feel mature, Watching TV/Cinema)

Figure 5.7 Both male and female Smoker with only Reasons or influences of starting smoking

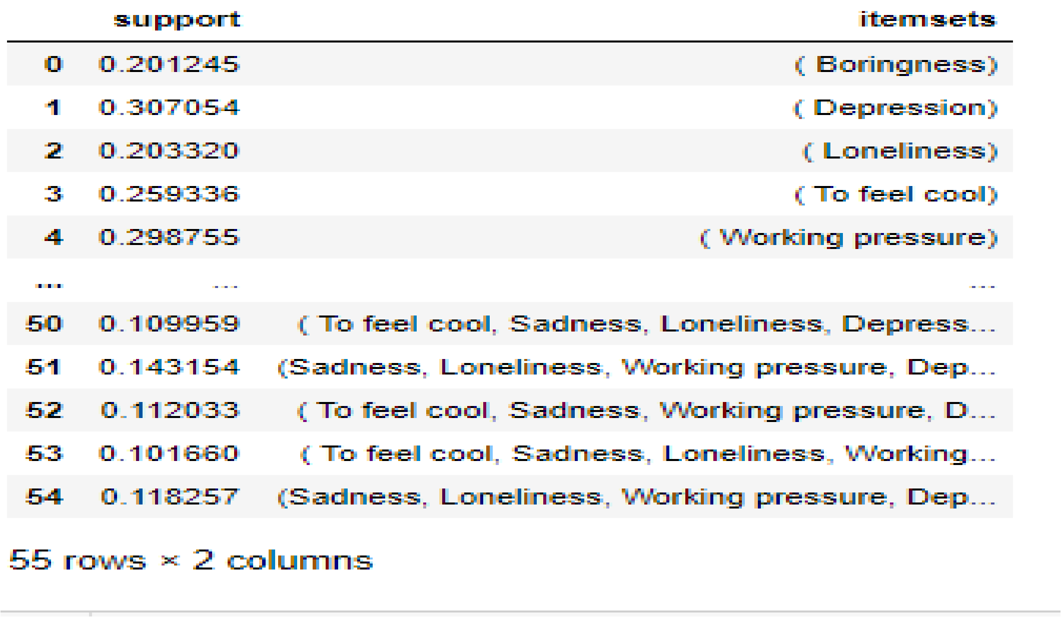


```
rules = association_rules(frq_items, metric ="confidence", min_threshold = .6)
# rules.shape
rules
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(To feel mature)	(Watching TV/Cinema)	0.126556	0.275934	0.107884	0.852459	3.089363	0.072963	4.907561

Figure 5.8 male and female Smoker with only Reasons or influences of starting smoking generated rules

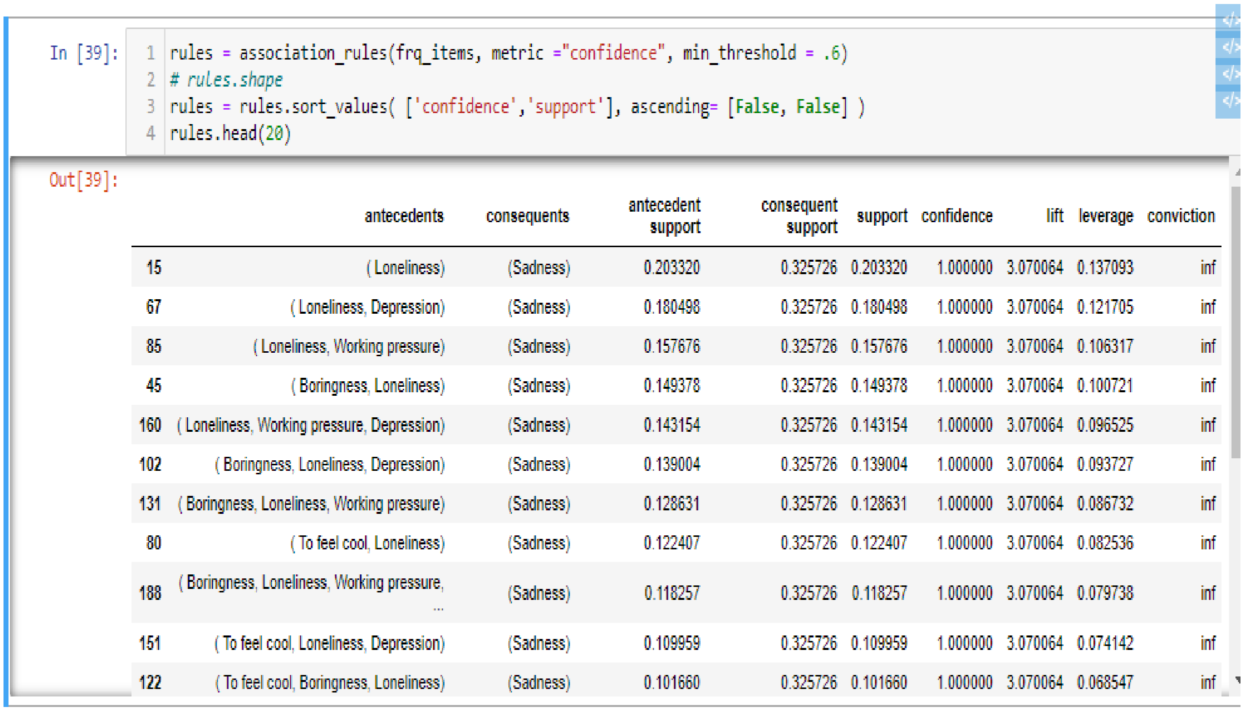
Rule 5: Now try to find some rule mining on factors that helps smoker to overcome his personal issues.



	support	itemsets
0	0.201245	(Boringness)
1	0.307054	(Depression)
2	0.203320	(Loneliness)
3	0.259336	(To feel cool)
4	0.298755	(Working pressure)
...	...	...
50	0.109959	(To feel cool, Sadness, Loneliness, Depress...
51	0.143154	(Sadness, Loneliness, Working pressure, Dep...
52	0.112033	(To feel cool, Sadness, Working pressure, D...
53	0.101660	(To feel cool, Sadness, Loneliness, Working...
54	0.118257	(Sadness, Loneliness, Working pressure, Dep...

55 rows × 2 columns

Figure 5.9: Smoker on factors that helps smoker to overcome itemsets



In [39]:

```
rules = association_rules(frq_items, metric ="confidence", min_threshold = .6)
# rules.shape
rules = rules.sort_values( ['confidence','support'], ascending= [False, False] )
rules.head(20)
```

Out[39]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
15	(Loneliness)	(Sadness)	0.203320	0.325726	0.203320	1.000000	3.070064	0.137093	inf
67	(Loneliness, Depression)	(Sadness)	0.180498	0.325726	0.180498	1.000000	3.070064	0.121705	inf
85	(Loneliness, Working pressure)	(Sadness)	0.157676	0.325726	0.157676	1.000000	3.070064	0.106317	inf
45	(Boringness, Loneliness)	(Sadness)	0.149378	0.325726	0.149378	1.000000	3.070064	0.100721	inf
160	(Loneliness, Working pressure, Depression)	(Sadness)	0.143154	0.325726	0.143154	1.000000	3.070064	0.096525	inf
102	(Boringness, Loneliness, Depression)	(Sadness)	0.139004	0.325726	0.139004	1.000000	3.070064	0.093727	inf
131	(Boringness, Loneliness, Working pressure)	(Sadness)	0.128631	0.325726	0.128631	1.000000	3.070064	0.086732	inf
80	(To feel cool, Loneliness)	(Sadness)	0.122407	0.325726	0.122407	1.000000	3.070064	0.082536	inf
188	(Boringness, Loneliness, Working pressure, ...	(Sadness)	0.118257	0.325726	0.118257	1.000000	3.070064	0.079738	inf
151	(To feel cool, Loneliness, Depression)	(Sadness)	0.109959	0.325726	0.109959	1.000000	3.070064	0.074142	inf
122	(To feel cool, Boringness, Loneliness)	(Sadness)	0.101660	0.325726	0.101660	1.000000	3.070064	0.068547	inf

Figure 5.10: Smoker on factors that helps smoker to overcome generated rules

5.3 Result Evaluation

For evaluating the result we need to focus on the lift of our generated rules. We studied on five different criteria for mining and we got the lift value greater than 1. So we can say that all of our generated rules are positively correlated.

CHAPTER 6

Conclusion and Future Research

6.1 Summary

Controlling the growing number of smoker is a great challenge. More people are starting smoking now-a-days than before. Find the causes of why people starting is our main objectives. There are various rule mining algorithm. In our study we used the one of the most popular one Apriori. From various rule mining on our dataset we have seen some interesting finding to understand the relation of different factor for starting smoking

6.2 Conclusion

This study is only for academic purpose. We have found that the algorithm works well on our dataset for other researches to apply Apriori on their study. Analyzing our rules we believe it will help the society to reduce the number of smoker.

6.3 Implication for Future Research

In the result section we have present various factors rule mining result that will be only helpful if we mass people can use our research. It can be a mobile app and web base service to predict one’s chances of starting smoke so that he/she can more aware about it. Also out dataset worked well, so we have a plan to do a further study to predict smoker based on the factors.

References

[1] C. A. Rose, J. Henningfield, M. J. Hilton, and D. T. Sweanor, “Smoking Definition, Types, Effects, History, & Facts Britannica,” *Britannica Encyclopedias*, 2018. https://www.britannica.com/topic/smoking-tobacco (accessed Oct. 06, 2020).

[2] “Tobacco-Related Mortality | CDC.” https://www.cdc.gov/tobacco/data_statistics/fact_sheets/health_effects/tobacco_related_mortality/index.htm (accessed Sep. 21, 2020).

[3] N. Nargis *et al.*, “Prevalence and patterns of tobacco use in Bangladesh from 2009 to 2012: Evidence from International Tobacco Control (ITC) study,” *PLoS One*, vol. 10, no. 11, Nov. 2015, doi: 10.1371/journal.pone.0141135.

[4] “Tobacco causes over 1 lakh deaths in Bangladesh every year: study | The Daily Star.” https://www.thedailystar.net/city/news/tobacco-causes-over-1-lakh-deaths-bangladesh-every-year-study-1706461 (accessed Sep. 21, 2020).

[5] “WHO | Smoking-attributable mortality in Bangladesh: proportional mortality study.” https://www.who.int/bulletin/volumes/91/10/13-120196/en/ (accessed Sep. 21, 2020).

[6] ACS Medical Content and News Staff, “Why People Start Using Tobacco, and Why It’s Hard to Stop,” *Am. Cancer Soc.*, pp. 1–7, 2015, [Online]. Available: https://www.cancer.org/cancer/cancer-causes/tobacco-and-cancer/why-people-start-using-tobacco.html#references.

[7] M. J. Zaki and C. Hsiao, “CHARM: An efficient algorithm for closed association rule mining,” *Proc. SDM*, vol. 2002, p. 20, 2002, [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.17.2956&amp;rep=rep1&amp;type=pdf.

[8] N. Vijayalakshmi and M. UmaMaheshwari, “Data Mining To Elicit Predominant Factors Causing Infertility in Women,” *Interantional J. Comput. Sci. Mob. Comput.*, vol. 5, no. 8, pp. 5–9, 2016.

[9] R. Agrawal, T. Imieliński, and A. Swami, “Mining Association Rules Between Sets of Items in Large Databases,” *ACM SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, Jan. 1993, doi: 10.1145/170036.170072.

[10] R. and R. S. Agrawal, “Fast Algorithms for Mining Association Rules in Large Databases. in Proceedings of the 20th International Conference on Very Large Data Base,” 1994. https://dl.acm.org/doi/10.5555/645920.672836 (accessed Sep. 21, 2020).

[11] P.-N. Tan, M. Steinbach, and V. Kumar, “Introduction to Data Mining, (First Edition),” p. 769, 2005, doi: 10.1016/j.cll.2007.10.008.

[12] G. Atluri, R. Gupta, G. Fang, G. Pandey, M. Steinbach, and V. Kumar, “Association analysis techniques for bioinformatics problems,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5462 LNBI, pp. 1–13, 2009, doi: 10.1007/978-3-642-00727-9_1.

[13] P. Tanna and D. Y. Ghodasara, “Using Apriori with WEKA for Frequent Pattern Mining,” *Int. J. Eng. Trends Technol.*, vol. 12, no. 3, pp. 127–131, Jun. 2014, doi: 10.14445/22315381/ijett-v12p223.

[14] P. Giudici and G. Passerone, “Data mining of association structures to model consumer behaviour,” *Comput. Stat. Data Anal.*, vol. 38, no. 4, pp. 533–541, 2002, doi: 10.1016/S0167-9473(01)00077-9.

[15] F. Bodon, “A Fast Apriori Implementation Hungarian Academy of Sciences,” *Fimi*, vol. 3, p. 63, 2011, [Online]. Available: http://www.cs.bme.hu/~bodon/kozos/papers/bodon_trie.pdf.

[16] K. Meena and N. Vijayalakshmi, “Analysis of factors causing infertility in women using statistical analysis and association rule mining,” *Indian J. Public Heal. Res. Dev.*, vol. 6, no. 2, pp. 112–117, 2015, doi: 10.5958/0976-5506.2015.00084.4.

[17] K. M. Mehedi Hasan Sonet, M. Mustafizur Rahman, P. Mazumder, A. Reza, and R. M. Rahman, “Analyzing patterns of numerously occurring heart diseases using association rule mining,” in *2017 12th International Conference on Digital Information Management, ICDIM 2017*, Jun. 2017, vol. 2018-Janua, pp. 38–45, doi: 10.1109/ICDIM.2017.8244690.

[18] J. Nahar, T. Imam, K. S. Tickle, and Y. P. P. Chen, “Association rule mining to detect factors which contribute to heart disease in males and females,” *Expert Syst. Appl.*, vol. 40, no. 4, pp. 1086–1093, 2013, doi: 10.1016/j.eswa.2012.08.028.

[19] G. Singh and S. Jassi, “A Review Paper: A Comparative Analysis on Association Rule Mining

Algorithms,” *Int. J. Recent Technol. Eng.*, vol. 6, no. 2, pp. 1–3, 2017.

[20] A. Khanum, “General Pseudocode for Association Rule Mining | Download Scientific Diagram.” https://www.researchgate.net/figure/General-Pseudocode-for-Association-Rule-Mining_fig8_44236182 (accessed Oct. 06, 2020).

[21] J. Han and I. Pei, “Hangars - an overview | ScienceDirect Topics.” https://www.sciencedirect.com/topics/engineering/hangars (accessed Oct. 06, 2020).

Plagiarism Report

ORIGINALITY REPORT

27%	22%	13%	16%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Fundación Universitaria del Area Andina Student Paper	2%
2	jfmt.indianjournals.com Internet Source	1%
3	Submitted to Monash University Student Paper	1%
4	journals.plos.org Internet Source	1%
5	Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava. "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", IEEE Access, 2019 Publication	1%
6	en.wikipedia.org Internet Source	1%
7	Palvi Soni, Sheveta Vashisht. "Exploration on Polycystic Ovarian Syndrome and Data Mining Techniques", 2018 3rd International Conference	1%