



Daffodil
International
University

**Smart Stomach Cancer Risk Prediction Application:
A Data Mining Approach**

By

MD. Rejaul Islam Royel

ID: 151-35-910

MD. Ajmanur Jaman

ID: 151-35-1100

A thesis submitted in partial fulfillment of the requirement for the degree of
Bachelor of Science in Software Engineering

**Department of Software Engineering
DAFFODIL INTERNATIONAL UNIVERSITY**

Fall – 2018

DECLARATION

We hereby declare that we have taken this thesis under the supervision of **Mr. Sayed Asaduzzaman, Lecturer, Department of Software Engineering, Daffodil International University**. We also declare that neither this thesis nor any part of this has been submitted elsewhere for award of any degree.

Royal
24/12-18

MD. Rejaul Islam Royel

ID: 151-35-910

Batch : 16th

Department of Software Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Ajmanur Jaman

MD. Ajmanur Jaman

ID: 151-35-1100

Batch : 16th

Department of Software Engineering
Faculty of Science & Information Technology
Daffodil International University

Certified by:

AS
24.12.2018

Mr. Sayed Asaduzzaman

Lecturer

Department of Software Engineering

Faculty of Science & Information Technology

Daffodil International University

APPROVAL

This Thesis titled “Smart Stomach Cancer Risk Prediction Application: A Data Mining Approach”, submitted by MD. Rejaul Islam Royel, ID: 151-35-910 & MD. Ajmanur Jaman, ID: 151-35-1100”, to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Software Engineering and approved as to its style and contents.

BOARD OF EXAMINERS




Dr. Touhid Bhuiyan
Professor and Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Chairman



Dr. Md. Asraf Ali
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1



Md. Maruf Hassan
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2



Prof Dr. Mohammad Abul Kashem
Professor
Department of Computer Science and Engineering
Faculty of Electrical and Electronic Engineering
Dhaka University of Engineering & Technology, Gazipur

External Examiner

Acknowledgment

First of all, we are expressing our gratitude to the Almighty Allah for giving us the ability to complete this thesis work. We will like to express our sincere gratitude to my honorable supervisor, Mr. Sayed Asaduzzaman, Lecturer, Department of Software Engineering. This thesis would not have been completed without his support and guidance. His constant encouragement gave us the confidence to carry out my work. We will like to give special gratitude to my teacher Md. Habibur Rahman, Lecturer, Department of Software Engineering. His proper direction and guidance help us to complete this thesis work without any difficulty.

We express our heartiest gratitude towards the entire Dept. of Software Engineering at DIU for providing good education and knowledge.

We also express our gratitude to all of our teachers Musfiquer Rahman, Bikash Kumer Paul, Department of Software Engineering, Daffodil International University. We would like to convey our sincere gratitude to Kawsar Ahmed, Assistant Professor; Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University. The knowledge that we have learned from the classes in our degree of Bachelor in Software Engineering level was essential for this thesis. In course of conducting the study, necessary information was collected through books, Journals, electronic media, and other secondary sources. We like to give thanks some of our friends for providing us with support and encouragement. Their suggestions helped us in countless ways.

Lastly but not the last, we will like to thank our parents and family members. Their optimism and encouragement have allowed me to overcome any obstacle that any faced.

TABLE OF CONTENT

Declaration.....	i
Approval	ii
Acknowledgment	iii
Table of Contents.....	iv
List of Tables	viii
List of Figures	ix
List of Abbreviations	x
Abstract.....	xi
CHAPTER 1: INTRODUCTION	1 – 5
1.1 Background	1
1.2 Motivation of the Research	2
1.3 Problem Statement	3
1.4 Research Questions	3
1.5 Research Objectives	4
1.6 Research Scope	4
1.7 Thesis Organization	5

CHAPTER 2: LITERATURE REVIEW	6- 8
2.1 Background	6
2.2 Preoperative Risk Factors	6
2.3 Protectable Ways	8
2.4 Summary	8
CHAPTER 3: RESEARCH METHODOLOGY	9 - 20
3.1 Data Collection and Preprocessing.....	10
3.1.1 Data Collection.....	10
3.1.2 Data Preprocessing (DPP)	10
3.2 Selecting analysis approach.....	11
3.2.1 Statistical Approach.....	11
3.2.1.1 ANOVA Test	11
3.2.1.2 Chi-Square Test.....	12
3.2.1.3 Odds Ratio Test.....	12
3.2.1.4 Probability Test.....	12
3.3 Data Mining Approach.....	13
3.3.1 Feature Selection.....	13
3.3.1.1 Correlation-based feature selection.....	13
3.3.1.2 Information Gain (IG) based Feature selection.....	14

3.3.1.3 Gain Ratio (GR) based Feature selection	14
3.3.1.4 Relief Based (RF) Feature Selection.....	14
3.3.1.5 Symmetrical Uncertainty (SU) Based Feature Selection.....	15
3.3.2 Association Rules by Predictive Apriori.....	15
3.4 Risk Score Calculation	16
3.4.1 Calculating Initial Score.....	17
3.4.1.1 Probability Calculation.....	17
3.4.1.2 Disease = Yes rules Calculation.....	17
3.4.1.3 Disease = No rules Calculation.....	18
3.4.2 Average Score Calculation.....	18
3.4.3 Elegant Score Calculation.....	19
3.4.4 Final Score Calculation.....	19
3.4.5 Designing Algorithm	19
3.5 Android Application Necessary Tools.....	20
3.6 Summary.....	20
CHAPTER 4: RESULTS AND DISCUSSION.....	21 - 44
4.1 Dataset Description.....	21
4.2 Statistical Analysis	22

4.2.1 ANOVA and Chi-square test	23
4.2.2 Odds Ratio test	26
4.2.3 Probability Test.....	28
4.3 Data mining results.....	30
4.3.1 Feature Selection.....	31
4.3.2 Yes rules from Apriori	33
4.3.3 No rules from Apriori	36
4.4 Score Calculation	39
4.5 Application Layout.....	44
4.6 Discussion.....	43
4.7 Summary.....	44
CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS.....	45 - 46
5.1 Findings and Contributions.....	45
5.2 Future Work	46

LIST OF TABLES

Table 3.1: Best Disease= Yes rules calculation formula.....	17
Table 3.2: Best Disease= No rules calculation formula.....	18
Table 3.3: Risk calculation results.....	19
Table 4.1: Frequency Distribution with P-value and Chi-square of Risk Factors Between Case Group and Control Group.....	23
Table 4.2: Odds Ratio with Confidence Interval of Predictor.....	26
Table 4.3: Probability table of all risk factors.....	28
Table 4.4: Feature selection techniques with respect to ranker method.....	31
Table 4.5: Best rules for Disease= Yes by Predictive Apriori algorithm.....	33
Table 4.6 Best rules for Disease = No by Predictive Apriori algorithm.....	36
Table 4.7 Score table for each sub-category.....	39

LIST OF FIGURES

Figure 3.1: Research Methodology.....	9
Figure 3.2: Risk Score calculation process.....	16
Figure 4.1: Top Feature based on feature selection.....	32
Figure 4.2: Support VS Confidence table with respect to lift for Disease = Yes rules.....	34
Figure 4.3: Visual relationship among factors for Disease = Yes.....	35
Figure 4.4: Support VS Confidence table with respect to lift for Disease = No rules.....	37
Figure 4.5: Visual relationship among factors for Disease = No.....	38
Figure 4.6: SC Risk Prediction Algorithm Flowchart.....	41
Figure 4.7A: application layout.....	42
Figure 4.7B: application Results.....	42

LIST OF ABBREVIATIONS

NICRH	National Institute of Cancer Research and Hospital
SC	Stomach Cancer
H. pylori	Helicobacter pylori
BMI	Body Mass Index
T2DM	Type 2 Diabetes Mellitus
AANHPI	Asian American, Native Hawaiian, and Pacific Islander
OR	Odds Ratio
DPP	Data Preprocessing
PRB	Probability
ANOVA	Analysis of variance
WEKA	Waikato Environment for Knowledge Analysis
SPSS	Statistical Package for the Social Sciences
XML	Extensible Markup Language
IG	Information Gain
GR	Gain Ratio
SU	Symmetrical Uncertainty

Abstract

Stomach Cancer is the 3rd conducting cause of cancer and the 5th most deadly disease among all diseases in world-wide. In this study, our aim is to find out all possible preoperative risk factors of SC and develop an android based application to predict the risk level of SC. From this perspective patient's data are collected from NICRH. To conduct this study statistical (ANOVA test, Chi-Square test, Odds Ratio, Probability test) and data mining (Feature Selection, Predictive Apriori Algorithm) approach has been used to get significant and highly related risk factors for SC. After that, a risk score algorithm has been designed based on an algorithm and finally developed the application. Experimenting 300 subjects' records (150 is affected and 150 is non-affected) with 33 risk factors we will get 25 statistically significant $P = (P < 0.05)$ risk factors and 18 top features. Where "Abdominal Pain" is the top preoperative risk factors of SC including ($P < 0.000$, $X^2=175.274$, and $OR = 66.769$) and "Nausea" including ($P < 0.000$, $X^2=152.261$, $OR = NA$) and "Skin Color Turn into Pale" including ($P < 0.000$, $X^2=138.240$, and $OR = 139.462$) respectively second and third most risk factors. Also, founded other high-risk factors are "Menetrier Disease = Yes", "Get Ill Too Much = Yes", "Previous Stomach Surgery = Yes", "Take Spicy and Salted Food = Yes", "Education Level = Less than high school", "Monthly Income= Less than 20 k", "Blood Group = A", "BMI= Severely Underweight or Overweight" and etc. This application will become very helpful and efficient for all researcher, doctors, and peoples from Bangladesh (individually low and middle-income people) to understand the risk factors of SC.

Key Word: Stomach Cancer, Data Mining, Statistical Analysis, Feature Selection, Risk Factors, Android Application.

CHAPTER 1

INTRODUCTION

1.1 Background

There are a few ailments those are deleterious for human life and malignant, cancer is one of them additionally it is the source of death. Last few decades stomach cancer is a critical sickness and general medical issue in around the world (Sitarz, et al., 2017). It starts when cells raise in the internal lining of our stomach and these cells can developed into a tumor which spreads to SC. This disease naturally grows slowly over many years, it is also known as gastric cancer. According to anatomic sites, SC can be classified into two subtypes like cardio and non-cardia (Kelley, et al., 2003). It is the 3rd conducting cause of cancer-related mortality and is also the 5th most deadly disease among all diseases in the world (Cheung, et al., 2018; Ferror, et al., 2014). There are some common risk factors of SC like as smoking, drinking alcohol, less physical activity and salt, including salt- continued foods, are probable causes of this cancer (Fang, et al., 2015; Torre, et al., 2015). Old age, female gender, and poor daily living were the factors most frequently linked with the morbidities of this disease (Kunisaki, et al., 2017). H. pylori is also correlated with an enhanced risk of SC (Amieva, et al., 2016; Cover, et al., 2016; Hansson, et al., 1996; Sitas, et al., 1991;). This study shows that the risk of SC with an increasing body mass index (BMI) was analyzed in men and women and obesity is a major contributing risk factor of SC indirectly (Latino-Loschmann, et al., 2017; Latino-Martel, et al., 2017). Diabetes has been documented as a risk factor of SC (Latino-Martel, et al., 2017). Rapid urbanization changed the environment and the lifestyle of peoples, this change creates an air–water–soil pollution, less physical activity,

electromagnetic radiation is also risk factors for SC (Liu, et al., 2016). According to GLOBOCAN cancer database estimates, about 14.1 million new cancer cases and 8.2 million deaths occurred in 2012 worldwide (Torre, et al., 2015). Last two decades, the death rate of cancer reduce to 26% and during this period 2.4 million peoples are death (Bray, et al., 2018). There are 13 to 15 lakh peoples are affected by cancer in Bangladesh and every year approximately 2 lakh peoples are newly diagnosed with cancer where SC is in the top five in Bangladesh perspective (Hussain, et al., 2013). In this world, over the last decades there is less mortality rate because of the decline in the prevalence of H. pylori infection and tobacco smoking, and to the improvements in food preservation and diet (In, et al., 2018). Generally, a diagnosis is made when the cancer is more advanced but early diagnosis is oppressive (Thrumurthy, et al., 2013). Because it can take some time to identify SC, only about 10% of people are diagnosed while it's still in the initial stage. For advanced SC, surgery is a must for a cure and its success rate is 50% but in its initial stage, SC can be rectified (Sitarz, et al., 2017). It should be said that to protect against SC a healthy lifestyle is very important (Liu, et al., 2016). There are lots of work to detect the risk factors of SC using population-based case-control study, algorithm and induction techniques. Apart from these, nowadays a most popular technique to predict SC risk is data mining technique. Using this new technology of risk prediction tool for cancer research may be hugely beneficial for the population-based research to prevent SC.

1.2 Motivation of the Research

In last 10 years, there are many cancer's related research paper were published in Bangladesh perspective like as Lung Cancer, Breast Cancer, Brain Cancer, Skin Cancer, Cervical Cancer, Oral Cancer (Ahmed, et al., 2015; Asaduzzaman, et al., 2015; Jesmin, et al., 2013; Kawsar, et al.,

2013;). Their success rate is attractive and these research papers are expressively and considerable to create public awareness. But we do not find any significant research work related with SC in Bangladesh perspective although SC is in the top five in Bangladesh perspective in all cancer diseases, that's why we were motivated to do our research work about SC in Bangladesh perspective.

1.3 Problem Statement

Day by day cancer diseases are grow up pernicious and death like all over the world. According to the Global Burden of Disease the incidence of cancer at 14.9 million cases, accounting for 8.2 million deaths and 196.3 million disability-adjusted life years (Grosso, et al., 2017). Cancer is the leading cause of death in Asian Americans, Native Hawaiians, and Pacific Islanders (AANHPI) (Torre, et al., 2016). SCs are the most demolishing and chronic forms among cancer and their treatment may be excessively complex and costly (Mahmoodi, et al., 2016). SC's influenced people are anxious for the family also the nations because they cannot contribute for the family economically. SC affected people become impotency and they are not come in useful.

1.4 Research Questions

- How to prepare an effective questionnaire for find out preoperative risk factors of SC in Bangladesh perspective?
- How to collect Case Group Data and Control Group Data?
- How to clean data for a meaningful research paper?
- How to make a serviceable dataset for the research work?
- How to analyze the data for effective research?
- How to use the analysis tools and extract significant results?
- How to design the risk prediction algorithm of SC?

- How to develop an attractive android application for checking the risk level of SC?

1.5 Research Objectives

- To find the preoperative risk factors of SC in Bangladesh perspective.
- To find the association among the preoperative risk factors.
- To develop an algorithm for risk level monitoring of SC.
- To design and implement an android application based on the risk algorithm.
- To create awareness among the people in Bangladesh of SC.

1.6 Research Scope

Data Mining Approach: To find significant preoperative risk factors two Data mining tools Orange and WEKA was used. Orange is a visual programming software package data mining toolkit for data visualization and this toolkit was used in our research for Probability Test, χ^2 – Test etc. WEKA was used for algorithm-based analysis. A search method Ranker is selected to rank all attributes regarding the evaluation results. This method treats the Missing value as a separate value for the attributes. WEKA was also used to find correlation among the factors using Apriori Algorithm. By these procedures, the significance level among the factors is explored on the Dataset.

Statistical Approach: Statistical approach has been used to find significance and correlation among preoperative risk factors. We have used SPSS V20.0 for ANOVA (P-Value) and Chi-square Test in our research work. By P value, the significant factors can easily be defined from the dataset. SPSS was also used for "odds ratio" and the odds that a case has been exposed to a risk factor is compared to the odds for a case that has not been exposed.

1.7 Thesis Organization

Chapter 1: Introduction: In this section, we will discuss why SC is accorded in older age and which the symptoms are responsible for identifying SC. Also have a short description for each risk factors.

Chapter 2: Literature Review: In this section, we discuss the previous work which is related to SC. We try to find the preoperative risk factors in the available literature and their suggestion which are influenced to occur SC directly and indirectly.

Chapter 3: Research Methodology: In this section, we discuss our useable methodology which we have been used in our analysis to search preoperative risk factors and we also trace the most 18 preoperative risk factors of 33 factors using different research methodology. We use two different approaches one is Data Mining Approach and other is Statistical Approach to make our research fruitful.

Chapter 4: Result and Discussion: In this section, we discuss all results in details. We take our most responsible preoperative risk factors and have to talk about them why these factors are liable to occur SC and which symptoms are accountable for the SC. We also recapitulate the reasonable preoperative risk factors of SC and ways to protect against the preoperative risk factors of SC.

Chapter 5: Conclusions and Recommendations: In this section, we will discuss what we find after completing result analysis. Also new finding is discussed properly and we provide some recommendation for all people in Bangladesh.

CHAPTER 2

LITERATURE REVIEW

2.1 Background

The occurrence of the SC varies throughout the world (Naylor, et al., 2006). SC indicates to tumors of the stomach that rise from the gastric mucosa, the connective tissue of the gastric wall, neuroendocrine tissue, or lymphoid tissue (Thrumurthy, et al., 2013). SC is anatomically divided into two categories, one is non-cardia cancers, which still comprises the majority of cases, and another is proximal gastric or cardia carcinomas (Van Cutsem, et al., 2011). According to AANHPI in 2016, there will be an estimated 57,740 new cancer cases and 16,910 cancer deaths. While AANHPIs have 30% to 40% lower incidence and mortality rates than non-Hispanic whites for all cancers combined, the risk of SCs is double (Torre, et al., 2016). SC is the fifth most common cancer and the third leading cause of cancer mortality in the world.

2.2 Preoperative Risk Factors

There are many well-known risk factors for SC, including diet, lifestyle, older age, gender, race, tobacco smoking, radiation, family history, *Helicobacter pylori* infection, low socioeconomic status, high intake of salty foods, low consumption of fruits and vegetables, obesity significantly associated with SC. Unhealthy dietary patterns were linked with higher BMI and energy intake, while healthy patterns were linked with higher education, physical activity, and less smoking (Grosso, et al., 2017). Processed meat and frequently used oil are significant factors of SC and the study shows that the diet high in salt and low in vitamins may be associated with an increase the risk of SC but a diet rich in vitamin C is protective (Ngoan, et al., 2002). The study also shows that SC mortality in the age groups 25–64 years to be double for lower educated compared to their

highly educated counterparts. People who are living in lower socioeconomic status generally are also more likely to have other lingering conditions, because of poor living conditions and lifestyle habits (De Vries, et al., 2015). SC is mostly found in male rather than female (Ngoan, et al., 2002). Positive associations between tobacco smoking and SC have been reported (Latino-Martel, et al., 2017). The risk of SC linked with bidi and cigarette smoking decreased with increased age at onset of smoking. The risk also seen between recent smokers was outstandingly unlike from that of ex-smokers. Those people who smoke cigarettes they have the double risk to affect SC rather than non-smoker (Gajalakshmi, et al., 1996). The study also shows that for SC the familial aggregation is unclear. Most of the SCs are scattered, approximately 10% show familial aggregation. While family history is an significant and consistently described risk factor for SC, the molecular basis for familial aggregation is unclear (Choi, et al., 2016). Nowadays *Helicobacter pylori* infection may be considered a cause of SC (Amieva, et al., 2016; Cover, et al., 2016; Hansson, et al., 1996; Sitas, et al., 1991). In US less than high school educated peoples are highly affected with cancer rather than Graduate or Ph.D. holders. It is also said that less educated people are affected with SC greater than 2 times higher on Graduate or Ph.D. holders (Mouw, et al., 2008). It is terminated that consumption of higher levels of vegetables and fruits are associated with a reduced risk of cancer at most sites. A wide variety of vegetables and fruits with some suggestion that raw forms are associated most successively with lower risk (Steinmetz, et al., 1991). All kinds of green vegetables are protective against SC. Highly significant vegetable consumption is 0.6 times safe rather than non-vegetable consumers and daily average consumption rate is > 80g. It is also found that taking fruits is significant when cigarette smoking is not in the account (Chyou, et al., 1990). The intake of fruits and fried vegetables can prevent SC. Vitamin C, such as oranges, lettuce, tomatoes, lemons, and citrus fruits, is protective against SC (Nomura, et al., 1990). Obesity is a

major contributing risk factor of SC indirectly. Diabetes has been documented as a risk factor of SC because of few studies have investigated the relationship between diabetes and SC (Latino-Martel, et al., 2017). Although some studies found diabetes was positively associated with risk of SC but two large prospective population-based cohorts suggest that type 2 diabetes mellitus (T2DM) is not associated with SC risk (Zu, et al., 2015).

2.3 Protectable Ways

A new invention to confirm the presence of a tumor, the name of that device 'MacSpec Pen'. 'MacSpec Pen' is effective for testing the presence of a tumor, which is prerequisite for SC to rectify (Sitarz, et al., 2017). It appears that environmental factors must act a major function to indemnify from SC (Polom, et al., 2016). Nowadays the rates of SCs are reduced due to in the prevalence of *H. pylori* infection and improvements in sanitation, preservation and storage of foods and other dietary factors like as high consumption of fruit and vegetables, reduced salt consumption and change the dietary patterns (Sierra, et al., 2016).

2.4 Summary

After reviewing the above literature, we notice that SC is the malignant disease of all cancers in many countries. Last 2 decades, there are many people are affected by this disease and it terrified the people all over the world. Soundly, the dietary patterns in different peoples in different areas are the central fact to be SC. In many countries, people have become aware of this disease and the government takes steps to reduce the percentage of SC affected people. But in our country, the people are not conscious about this disease and most of them doesn't know why it can grow in our stomach. We also found that there is no significant research work in our country to find out the preoperative risk factors of SC and lack of study on the reasonable risk factors of SC in Bangladesh perspective.

CHAPTER 3

RESEARCH METHODOLOGY

In any survey-based research collection of relevant data and selection of proper analytical process is always a crucial part. If, one of them is missed in any research study even proper guided research study also lost their goal. So, in this study to fulfill these two criteria, we will use an extended structured model to predict risk factors and the risk level of SC. Basically, this model is generated by observing some related research papers (Jesmin, et al., 2013; Ahmed, et al., 2013; Raihan, et al., 2016). This model is very easy and effective to generate the risk level of any disease. In here we will flow their data collection and analytical process and divided analytical section in two parts including statistical and data mining approach. Those are described below.

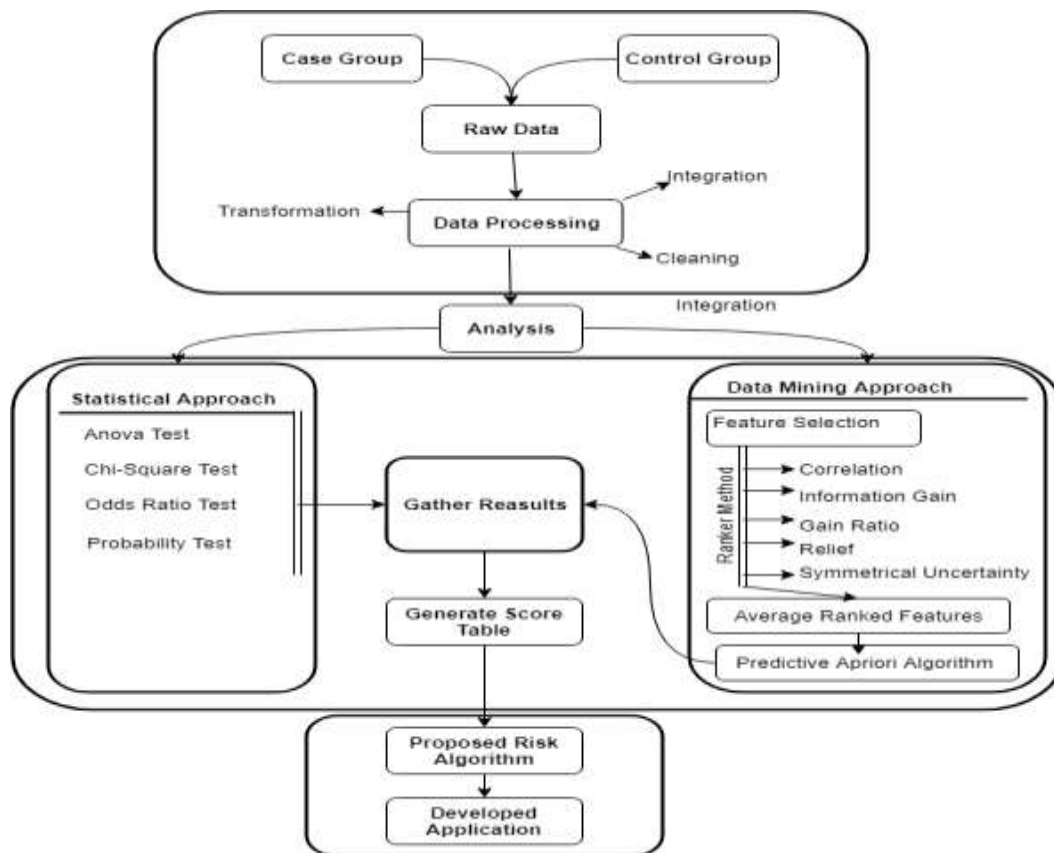


Figure 3.1: Research Methodology

Extracting information from any amount of raw data is a difficult task. Following a set of procedural operation is mandatory to extract hidden patterns of data. To, figure out most significant preoperative risk factors of SC in figure 3.1 we have collected affected (case group) and non-affected (control group) people's data from NICRH in Bangladesh. After collecting data all analysis is done with some statistical and data mining tools like IBM SPSS, Weka, Orange, and R.

3.1 Data Collection and Preprocessing

3.1.1 Data Collection

Collecting of data is very much an important task for any kind of survey-based research. In this research, all data are collected by a survey-based questionnaire. The question was designed by the study of different preoperative risk factors of the SC-based research paper. On behave those study most of the patient's age is greater than 30 years old. So, in this study, we only collect those people's data who are greater than or equal to 30 years old. Total collected individual's sample size of data is about 300 where case group is 150 and the control group is 150.

3.1.2 Data Preprocessing (DPP)

In real-world survey data is always incomplete, noisy and inconsistent. So, data preprocessing is much needed before conducting any analysis. It is true that in data mining without quality data there are no quality results. DPP is actually the combination of four tasks those are data integration, cleansing, transformation and reduction (Karegowda, et al., 2010). To get the good result, complementation of those tasks is mandatory. First of all, case and control groups of individual's data are integrated into one data set. Then, in the cleaning stage, all inconsistent data are corrected. To remove missing value all missing fields are filled by previous participant's history and also some binary method, clustering, and regression function is applied to handle noise in the dataset. Data transformation is done to get "BMI" from height and weight and "Duration of SC" is

categorized with some sub-categories. Data reduction is a technique to reduce the number of attributes and this process was discussed in the feature selection section.

3.2 Selecting analysis approach

In general, statistical analysis has been used to evaluate significant facts and test the hypothesis and data mining technology have been used to extract hidden risk factors of medical data like SC disease. In this study, our main motive is not only finding significant risk factors but also developed an android based application. From this perspective, we will use both statistical and data mining approach to get strong appropriate weight/score for each sub-category.

3.2.1 Statistical Approach

A different Statistical approach is used in medical research to influence the findings of relevant significance risk factors and also provide some recommendation (Sebastião, et al., 2018). Some important statistical approach like ANOVA test, Chi-Square test, Odds Ratio test, Probability test has been used to find out most significant risk factors for SC by statistical tool SPSS V21. Each test is discussed below.

3.2.1.1 ANOVA Test

ANOVA test has been used to find out the most significant risk factors for SC by observing the P-value of each factor.

$$\text{Sum of Square Between} = \sum N_A (\bar{X}_A - \bar{X})^2 \text{-----}(3.1)$$

Where the Sum of Square Between is two factors mean value. It is the total amount of distribution among the samples mean \bar{X}_A is an individual group mean value and \bar{X} is over all mean value (Shaphiro, et al., 1995).

3.2.1.2 Chi-Square Test

The human body's metabolomes are strongly interlinked with each other. Chi-Square Test has been used to find out associations among all factors except the dependent variable. It has been used to test the null hypothesis (There is no relation among categorical variables).

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} \text{-----}(3.2)$$

If the P value is $P > .05$ then it will accept the null hypothesis and if the P value is $P < .05$ to lower it will reject the null hypothesis and lower value shows the strong relationship among those categorical value (Rumsey, et al., 2015; Tallarida, et al., 1987).

3.2.1.3 Odds Ratio Test

After identifying all risk factors a question is come to a curious mind "Which factors are riskier for SC affected people than non-affected people?" To answer this question, the odds ratio test is performed, and it will give us the risk level ratio for case and control group. If, the value of the odds ratio is greater than or equal to 1. Then, case group is 1 time higher risky than the control group for this factor.

3.2.1.4 Probability Test

In an event probability is the ratio of the number of observations to the total numbers of the observations. The value of probability of an event is measured by the range of 0 to 1. Where 0 refers no probability and 1 refers to high probability (Kolmogorov, et al., 2018). Basic formula is

$$\text{Probability} = \frac{\text{Number of Favourable Outcome}}{\text{Total Number of Outcome}} \text{-----}(3.3)$$

3.3 Data Mining Approach

Data mining is a technology to extract hidden patterns of data. It will very helpful to figure out the most frequent risk factors for SC. In this approach, first of all, we have used different types of feature selection technique with respect to the ranker method and the average rank is used to getting top features. Because we do not find any study which said this feature selection technology is best for SC survey data.

3.3.1 Feature Selection

In machine learning and data mining field, analyzing of high dimensional data is a difficult task. Where feature selection provides an effective way to solve this problem by reducing irrelevant features (Cai, et al., 2018). Also, it will increase learning accuracy, improving result comprehensibility and enhance any machine learning model fitting capability (Cilia, et al., 2018). Feature selection can be classified into two types one is Filter method and another is Wrapper (method of classification) method (Blum, et al., 1997; Karegowda, et al., 2010). In this study, we will only use the filter method to extract best features.

3.3.1.1 Correlation-based feature selection

A correlation-based feature selection technic is evaluating the correlation within the subset of features that are extremely correlated with the class by using a greedy search strategy in a mode of Ranker search method (Oh, et al., 2009). So, if the ranks rate is high then subsets are extremely correlated with each other otherwise there are no strong relationships among them. ss

$$\text{Correlation, } C(A|B) = \frac{H(B) - H(B|A)}{H(B)} \text{-----(3.4)}$$

Where $C(A|B)$ is the correlation between A and B and $H(A)$, $H(B)$ is the entropy of respectively A and B and $H(B|A)$ is entropy of B given A (Hall, et al., 1997).

3.3.1.2 Information Gain (IG) based Feature selection

Information gain-based feature selection generates an expected number of output (Information) from the classification target attribute (Cooper, et al., 1992). When using that feature a score is calculated based on how much information is gained by the class. The information gain of feature A is defined as follows

$$\text{Information Gain (A)} = H(B) - H\left(\frac{B}{A}\right) \text{-----(3.5)}$$

Where $H(B)$ is entropy of B and $H\left(\frac{B}{A}\right)$ is conditional entropy of class B Given Feature A (Cilia, et al., 2018; Jantawan, et al., 2014).

3.3.1.3 Gain Ratio (GR) based Feature selection

Information Gain Ratio is a ratio and formulated by Information Gain. It will maximize the feature information gain while minimizing the number of its values. The formula is defined as

$$\text{Gain Ratio (A)} = \frac{\text{IG (A)}}{I(B)} \text{-----(3.6)}$$

Where the gain ratio of A is defined as the information gain of a divided by its own value B (Cilia, et al., 2018; Jantawan, et al., 2014; Karegowda, et al., 2010).

3.3.1.4 Relief Based Feature Selection

Relief algorithm first formulated by Kira and Rendered on by instance-based learning (Urbanowicz, et al., 2018). It is an instance-based searching to assign a new weight for an individual feature. The searching procedure is very simple, just search for nearest neighbors in training dataset, not depth search. For each sampled instance, the nearest sample match and not

match are found. Those matches and the not-matching ratio will update the weight of individual features and gives a rank (Cilia, et al., 2018).

3.3.1.4 Symmetrical Uncertainty (SU) Based Feature Selection

SU Evaluates the cost of a set attributes by measuring the SU with respect to another set of attributes. It defines as

$$SU = 2 \left[\frac{IG(A)}{H(X) + H(Y)} \right] \text{-----(3.7)}$$

Here, H(X) is entropy of X and H(Y) is entropy of Y (Cilia, et al., 2018; Yu, et al., 2003). Where it covers for information gain's bias toward features with more values and tempers its values to the range (0, 1). The value 1 indicates that attributes X and Y are completely hooked and 0 refers they are independent.

3.3.2 Association Rules by Predictive Apriori

Association rules are very useful to extract the hidden pattern of data. Some most popular algorithm is Apriori, Predictive Apriori, and Tertius (Nahar, et al., 2013). In this study Predictive Apriori was used to generate the best rules. It is also an advanced form of Apriori algorithm where we used minimum support 0.01, minimum confidence is 0.80 and the lift is greater than 1 (Mahmoodi, et al., 2017) and all related figures are developed by “arulesViz” packages in R.

3.4 Risk Score Calculation

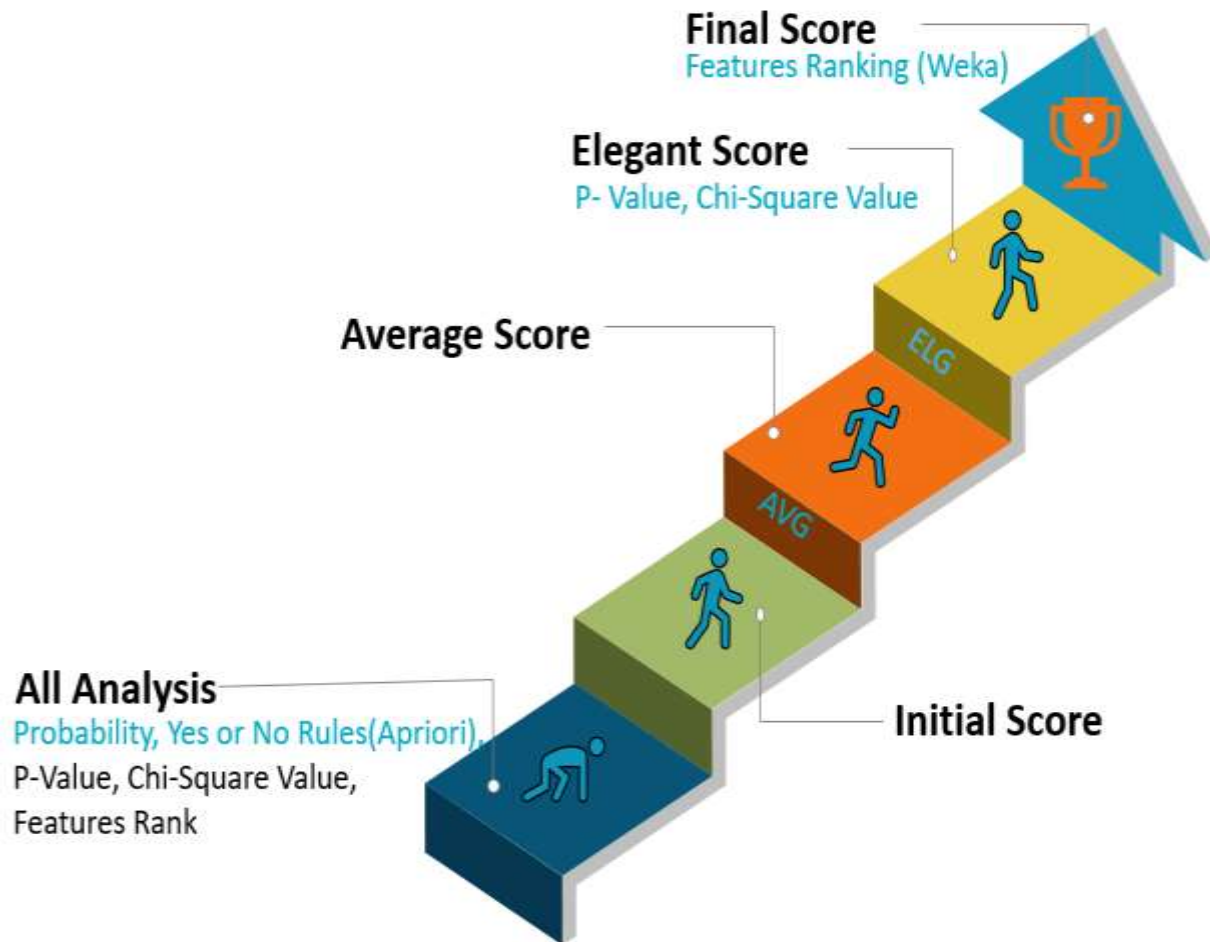


Figure 3.2: Risk Score calculation process

Figure 3.2 represents each sub-category risk score calculation process. In here, first of all, test results are gathered together and probability test, yes rules and rules by Apriori algorithm results are counts to calculate the initial score. Then simply get each average score. After that, we will elegant all our results by P-value and chi-square value and Finally features ranking has been applied to get final risk score.

3.4.1 Calculating Initial Score

The initial score is calculated by the probability test, Disease= Yes rules and Disease = No rules reports. Here, we will give priority first for probability test then Disease= Yes rules and finally Disease = No rules.

3.4.1.1 Probability Calculation

In our study, each factors subcategories occurrence probability is calculated by their probability ratio. In here, we have four different level of risk (Low, Moderate, High and Very High). So, if we divide probability equally in four categories then we get probability 0.1 to 0.25 is “Low”, 0.26 to 0.50 is “Moderate”, 0.51 to 0.75 is “High” and 0.76 to 1.00 is “Very High” risk.

3.4.1.2 Disease = Yes rules Calculation

Table 3.1: Best Disease= Yes rules calculation formula

Criteria	Support	Risk Score
max support	41	--
min support	14	--
Difference	27	--
Risk level	4	--
Per level will get	$31/4 = 6.75 \sim 7$	--
Low Risk	14 to 20	1
Moderate Risk	21 to 27	2
High Risk	28 to 34	3
Very High Risk	35 to 41	4

Table 3.1 is generated from the Apriori algorithm and we will measure risk criteria by its top supported rules. Top rules are selected by the methods of first come first serve. Here we get, supported value 14 to 20 is “Low Risk”, value 21 to 27 is “Moderate Risk”, value 28 to 34 is “High Risk” and value 35 to 41 is “Very High Risk”.

3.4.1.3 Disease = No rules Calculation

Table 3.2: Best Disease= No rules calculation formula

Criteria	Support	Risk Score
max support	42	--
min support	11	--
Difference	31	--
Risk level	4	--
Per level will get	$31/4 = 7.75 \sim 8$	--
Low Risk	35 to 42	1
Moderate Risk	27 to 34	2
High Risk	19 to 26	3
Very High Risk	11 to 18	4

Table 3.2 also calculated as like table 3.1. Top rules are selected by the methods of first come first serve. Here, high support value means this sub-category has strong evidence to have no disease and lower support value is denoted that this sub-category has been going to be the risk. We will get, the supported value between 11 to 18 is “Very High Risk”, between 19 to 26 is “High Risk”, between 27 to 34 is “Moderate Risk” and between 35 to 42 is “Low Risk”.

3.4.2 Average Score Calculation

After calculating the initial score for all sub-category then we will average all available score by their total number. Like, in the score table we will get Age = 30 to 49 get two scores (Yes rules = 3, and Probability = 2) so its average score is 2.5 on the other hand Age = Above 70 get only one score (Probability = 4) so, its average score is 4

3.4.3 Elegant Score Calculation

The elegant score is calculated by the P-value and the chi-square value of each factor. If P-value is highly significant and chi-square value is greater than 90 then we will assign 0.75 only for reasonable sub-category. In some cases, we will add or reduce score due to observing another sub-category in the same factors and overall literature review. Like “BMI = Normal” average score is 2.5, after eliminating this we will set it 0.5 and “skin color = Yes” average score 3.5 turns into 4.

3.4.4 Final Score Calculation

When we will get all lamented scores in hand, we will be applied feature selections average top features priority on each selected feature. In this case, all feature priority is count bottom to top and add another 0.05 value for each sub-category elegant score.

3.4.5 Designing Algorithm

Designing risk prediction algorithm is a major important task in this study. This algorithm is developed by the available value of risk score on table 4.7.

Table 3.3: Risk score calculation results

Criteria	Risk Score
Maximum Score	69.2
Minimum Score	30.35
Score Difference	38.85 ~ 39
Number of Risk level	04
Per level will get	$39/4 = 9.75$
Very High-Risk Score	Score ≥ 59.45
High Risk Score	Score ≥ 49.70
Moderate Risk Score	Score ≥ 39.95
Low-Risk Score	Score < 39.95

Here we will get, the maximum score is = 69.2 and the minimum score is 30.25. Score difference is 39 and our predicted risk level is 4 so every risk level will get 9.75 scores to get a new risk level.

3.5 Android Application Necessary Tools

- A. Java
- B. XML
- C. Android studio 3.2.1
- D. Other Some Dependency library

3.6 Summary

Procedural operational execution of any study is an important task. It will help any research study to go on the right path. In this study, our main motive is to develop an android based application which can predict SC risk factors mostly easily and effectively. That's why we are select statistical and data mining approach in the combine. Two different perspective operations will help us to take proper decision. We will select four statistical test those are (ANOVA, Chi-Square, Odds Ratio, and Provability test) to observe significant status, level and for testing hypothesis. And selected data mining approach was designed in two section one is feature selection and other is association rules mining. In the feature selection section, we will use five feature selection techniques (Correlation, Information gain, Gain Ratio, Relief, and Semantical Unsent) with ranker method to get rank between 0 to 100 for each feature. Then we will average them all features rank and get top 18 features among 33 features. Then, in the second stage association rule, miner Apriori algorithm was used to get top associated rules individually for Disease = Yes and Disease = No. Finally, we will calculate a score based on those tests. This strategy is very much easier and effective to mine medical data and develop a risk prediction algorithm and tools.

CHAPTER 4

RESULT ANALYSIS

4.1 Dataset Description

It is very important to understand research data before conducting any kind of research work. From this perspective, we like to explain our dataset briefly. In our dataset, all data represent the preoperative risk factors of SC those are organized from the literature review. Here two groups of people data have been collected one is Case group and another is the Control Group. So, in the table 3.1 here “Gender is categorized as (Male and Female)”, “Age is categorized as (Between 30 to 49, Between 50 to 59, Between 69 to 70 and Above 70)”, “BMI which is Body Mass Unit (Height, Weight) it categorized as (Normal, Underweight, Severely Underweight, Overweight and Obese)”, “Living Area (recently where subjects are living in) and it is categorized as (Ruler, Urban and Suburban)”, “Education Level (Subjects last completed educational degree) it is categorized as (Less Than High School, High School or College, University Graduate and Doctoral Degree) ”, “Working Status is categorized as (Unemployed, Private Sector, Business and Government Employee)” , “Monthly Income categorized as (Less than 20k, 20 to 30 k, 30 to 45k and above 45k)”, “Family Member (Number of people are lived together) categorized as (2 - 3, 4 - 5 and above 5)”, “Blood Group categorized as (A, B, O and AB)”, “Physical Activity categorized as (Regularly, Often, No)”, “Daily Food (Three times meal in a day) categorized as (Yes, No)”, “Spicy and Salted Food (Those who are habited to eat Spicy and Salted Food regularly) categorized as (Yes, No)”, “Green Vegetables (Who eat Green Vegetables three to five times in a week) categorized as (Yes, No) ”, “Yellow Fruits (Who eat Yellow Fruits three to five times in a week) categorized as (Yes, No) ”, “Tobacco Status categorized as (Yes Sometimes, Yes Excessive and No) ”, “Alcohol Status categorized as (One drink per day, Two drinks in a day, Occasionally and

NO)", "Duration of SC is categorized as (No, 1 to 6 month, 7 to 12 month and above 1 year) ", "Get Ill Too Much categorized as (Yes, No)", "Skin Color (Skin color turn into pale) categorized as (Yes, No)", "Abdominal Pain (Pain around navel) " categorized as (Yes, No)", "Nausea (Likely to be have vomiting) categorized as (Yes, No)", "Frequent Vomiting categorized as (Yes, No)", "Poor Appetite (There is no intention to eat foods) categorized as (Yes, No)", "Blood Vomiting categorized as (Yes, No)", "Tarry Stool (Bloods are come with stools)", "Breast Cancer Status (Cancer in female breast) categorized as (Yes, No)", "Previous Stomach Surgery categorized as (Yes, No)", "Stomach Lymphoma (Caused by *Helicobacter Pylori* bacteria symptoms is abdominal pain at night or after eating) categorized as (Yes, No)", "Menetrier Disease (massive growth of mucous cells in the stomach and those mucous are sometimes comes from mouth during sleep) categorized as (Yes, No)", "Type Two Diabetes categorized as (Yes, No)", "Another Cancer categorized as (Colon Cancer, None)", "Family History (Any family member are affected with SC or not) categorized as (Father, Mother, Brother, Sister, father Cast, Mother Cast and None)", and "Gastric Medicine (Those who take gastric medicine regularly) categorized as (Yes, No)".

4.2 Statistical Analysis

Statistical test has been analyzed to evaluate the significant level of each factors on respect to disease status and it also used to figure out which factor is much risky when subject is in case group rather than control group. In this section, four different popular statistical tests ANOVA, Chi-square, Odds Ratio and Probability are analyzed below.

4.2.1 ANOVA and chi-square test

Table 4.1 Frequency Distribution with P-value and Chi-square of Risk Factors Between Case Group and Control Group

Factor	Cancer status		P – value	X ² Value	Factor	Cancer status		P- value	X ² Value
	Affected N (%)	Unaffected N (%)				Affected N (%)	Unaffected N (%)		
Gender					Take Tobacco				
Male	108(72)	88(58.7)	0.015	5.887	Yes sometimes	49(32.7)	19(12.7)	0.787	44.846
Female	42(28)	62(41.3)			Yes excessive	41(27.3)	14(9.3)		
Age					No	60(40.0)	117(78.0)		
Between 30 to 49	54(36)	98(65.3)	0.000	32.664	Take Alcohol				
Between 50 to 59	49(32.7)	37(24.7)			One times in a day	2(1.3)	1(0.7)	0.096	3.347
Between 69 to 70	31(20.7)	13(8.7)			Two times in a day	0(0)	1(0.7)		
Above 70	16(10.7)	2(1.3)			Occasionally	0(0)	2(1.3)		
BMI					No	148(98.7)	146(97.3)		
Normal	74(49.3)	78(52)	0.000	112.147	Get Ill Too Much				
Underweight	33(22)	0(0)			Yes	99(66.0)	18(12.0)	0.000	91.929
Severely underweight	33(22)	1(7)			No	51(34.0)	132(88.0)		
Obese	6(4)	15(10)			Skin Color Turn into Pale				
Overweight	4(2.7)	56(37.3)			Yes	98(65.3)	2(1.3)	0.000	138.24
Living Area					No	52(34.7)	148(98.7)		
Rural	103(68.7)	76(50.7)	0.029	13.648	Abdominal Pain				
Urban	34(22.7)	64(42.7)			Yes	124(82.7)	10(6.7)	0.000	175.274
Suburban	13(8.7)	10(6.7)			No	26(17.3)	140(93.3)		
Education Level					Nausea				
Less than high school	117(78)	57(38)	0.000	66.969	Yes	101(67.3)	0	0.000	152.261
High school or college	31(20.7)	43(28.7)			No	49(32.7)	150(100)		
University graduate	2(1.3)	46(30.7)							
Doctoral Degree	0(0)	4(2.7)							

Factor	Cancer status		P value	X ² Value	Factor	Cancer status		P- value	X ² Value
	Affected N (%)	Unaffected N (%)				Affected N (%)	Unaffected N (%)		
Working Status					Frequent Vomiting				
Unemployed	16(10.7)	42(28)	0.000	23.052	Yes	45(30.0)	0	0.000	25.941
Private Sectors	95(63.3)	70(46.7)			No	105(70.0)	150(100)		
Business	32(21.3)	20(13.3)			Poor Appetite				
Govt Employee	7(4.7)	18(12)			Yes	6(4.0)	4(2.7)	0.522	0.414
Monthly Income					No	144(96.0)	146(97.3)		
Less than 20K	123(82)	89(59.3)	0.000	24.384	Bloody Vomiting				
20K to 30K	19(12.7)	26(17.3)			Yes	5(3.3)	0	0.024	5.058
30K to 45K	6(4)	18(12)			No	145(96.7)	150(100)		
Above 45K	2(1.3)	17(11.3)			Tarry Stools				
Family Member					Yes	9(6.0)	2(1.3)	0.032	4.624
2 to 3	8(5.3)	18(12)	0.000	14.337	No	141(94.0)	148(98.7)		
4 to 5	62(41.3)	83(55.3)			Breast Cancer Status				
Above 5	80(53.3)	49(32.7)			Yes	1(0.7)	1(0.7)	1	0
Blood Group					No	149(99.3)	149(99.3)		
A	61(40.7)	46(30.7)	0.000	17.972	Previous Stomach Surgery				
B	50(33.3)	42(28)			Yes	19(12.7)	2(1.3)	0.000	14.798
O	32(21.3)	31(20.7)			No	131(87.3)	148(98.7)		
AB	7(4.7)	31(20.7)			Stomach Lymphoma				
Physical Activity					Yes	26(17.3)	0	0.000	28.467
Regularly	107(71.3)	82(54.7)	0.032	12.232	No	124(82.7)	150(100)		
Often	31(20.7)	37(24.7)			Menetrier Disease				
No	12(8)	31(20.7)			Yes	26(17.3)	1(0.7)	0.000	25.438
Daily Food In time					No	124(82.7)	149(99.3)		
Yes	67(44.7)	115(76.7)	0.000	32.185	Type Two Diabetes				
No	83(55.3)	35(23.3)			Yes	15(10.0)	7(4.7)	0.077	3.139
					No	135(90.0)	143(95.3)		

Factor	Cancer status		P value	X ² Value	Factor	Cancer status		P- value	X ² Value
	Affected N (%)	Unaffected N (%)				Affected N (%)	Unaffected N (%)		
					Another Cancer				
Take Spicy and Salted Food					Colon cancer	1(0.7)	0	0.318	1.003
Yes	117(78.0)	66(44.0)	0.000	36.444	None	149(99.3)	150(100)		
No	33(22.0)	84(56.0)			Family History				
Take Green Vegetables					Father	2(1.3)	2(1.3)	0.877	1.737
Yes	95(63.3)	137(91.3)	0.000	33.545	Mother	2(1.3)	1(0.7)		
No	55(36.7)	13(8.7)			Brother	1(0.7)	3(2.0)		
Take Yellow Fruits					Sister	1(0.7)	1(0.7)		
Yes	67(44.7)	130(86.7)	0.000	58.681	Father Cast	8(5.3)	7(4.7)		
No	83(55.3)	20(13.3)			Mother Cast	1(0.7)	2(1.3)		
Take Gastric Medicine					None	135(90.0)	134(89.3)		
Yes	58(38.7)	28(18.7)	0.000	14.671					
No	92(61.3)	122(81.3)							

Table 4.1 represents the frequency distribution of SC patients (case group) and the (control group) with significant variation among risk factors of SC. Here clearly shows that the risk factors “Age” ($P<.000$), “BMI” ($P<.000$), “Education Level” ($P<.000$), “Working Status” ($P<.000$), “Monthly Income” ($P<.000$), “Family Member” ($P<.000$), “Blood Group” ($P<.000$), “Daily Food In time” ($P<.000$), “Take Spicy and Salted Food” ($P<.000$), “Take Green Vegetables” ($P<.000$), “Take Yellow Foods” ($P<.000$), “Duration of SC” ($P<.000$), “Get Ill Too Much” ($P<.000$), “Skin Color” ($P<.000$), “Abdominal Pain” ($P<.000$), “Nausea” ($P<.000$), “Frequent Vomiting” ($P<.000$), “Previous Stomach Surgery” ($P<.000$), “Stomach Lymphoma” ($P<.000$), “Menetrier Disease” ($P<.000$), and “Gastric Medicine” ($P<.000$), is highly associated with SC and also “Gender” ($P<.015$), “Living Area” ($P<.029$), “Physical Activity” ($P<.032$), “Blood Vomiting” ($P<.024$), “Terry Stool” ($P<.032$) are highly associated with SC.

Among these significant factors “Abdominal Pain” ($X^2 = 175.274$), “Nausea” ($X^2 = 152.261$), “Skin Color Turn Into Pale” ($X^2 = 138.240$), “BMI” ($X^2 = 112.147$), “Get Ill Too Much” ($X^2 = 91.929$), “Education Level” ($X^2 = 66.969$), “Take Yellow Fruits” ($X^2 = 58.681$) is extremely significant risk factors.

4.2.3 Odds Ratio test

Table 4.2 Odds Ratio with Confidence Interval of Predictor

Factors	Category	Sig.	Odds ratio	95% C.I.		Factors	Category	Sig.	Odds ratio	95% C.I.	
				Lower	Upper					Lower	Upper
Gender						Daily Food In time					
	Male	0.016	1.812	1.118	2.935		Yes	0.000	0.246	0.149	0.404
	Female	0.022					No	0.000			
Age						Spicy and Salted food					
	Between 30 to 49	0.000		0.334	0.599		Yes	0.000	4.512	2.728	7.463
	Between 50 to 59	0.000					No	0.000			
	Between 69 to 70					Green Vegetables					
	Above 70						Yes	0.000	0.164	0.085	0.317
							No	0.000			
BMI						Take Yellow Fruits					
	Normal	0.000	1.457	1.25	1.697		Yes	0.000	0.124	0.07	0.22
	Underweight	0.000					No	0.000			
	Severely underweight					Get Ill Too Much					
	Obese						Yes	0.000	14.235	7.834	25.866
	Overweight						No	0.000			
Living Area						Skin Color					
	Rural	0.030	1.497	1.039	2.155		Yes	0.000	139.462	33.202	585.798
	Urban	0.045					No	0.000			
	Suburban					Abdominal Pain					
Education Level							Yes	0.000	66.769	30.967	143.964
	Less than high school	0.000	4.414	2.953	6.598		No	0.000			
	High school or college	0.000									
	University graduate										
	Doctoral Degree										

Factors	Sub Category	Sig.	Odds ratio	95% C.I.		Factors	Sub Category	Sig.	Odds ratio	95% C.I.	
				Lower	Upper					Lower	Upper
Working Status						Nausea					
	unemployed	0.263	0.854	0.648	1.126		Yes	0.000	---	0	0
	Private Sectors	0.296					No	0.000			
	Business					Frequent Vomiting					
	Govt Employee						Yes	0.997	---	0	0
Monthly Income							No	0.997			
	Less than 20K	0.000	2.104	1.531	2.892	Bloody Vomiting					
	20K to 30K	0.000					Yes	0.999	---	0	0
	30K to 45K						No	0.999			
	Above 45K					Tarry Stools					
Family Member							Yes	0.050	4.723	1.003	22.242
	2 to 3	0.000	0.491	0.336	0.717		No	0.051			
	4 to 5	0.000				Previous Stomach Surgery					
	Above 5						Yes	0.002	10.733	2.453	46.954
Blood Group							No	0.002			
	A	0.001	1.492	1.187	1.877	Stomach Lymphoma					
	B	0.001					Yes	0.998	---	0	0
	O						No	0.998			
	AB					Menetrier Disease					
Physical Activity							Yes	0.001	31.242	4.18	233.509
	Regularly	0.033	1.351	1.025	1.78		No	0.001			
	Often	0.057				Take Gastric Medicine					
	No						Yes	0.000	2.747	1.623	4.648
							No	0.000			

Table 4.2 Represent the test results of Odds Ratio (OR) to compare different groups with 95% confidence interval (CI). Here from the table it is clearly depicted that “Gender”, “BMI”, “Living Area”, “Education Level”, “Monthly Income”, “Blood Group”, “Physical Activity”, “Take Spicy and Salted Food”, “Get Ill Too Much”, “Skin Color”, “Abdominal Pain”, “Tarry Stools”, “Previous Stomach Surgery”, “Menetrier Disease” have statistically significant relationships ($P < 0.05$) with SC.

“Skin Color” is a highly significant factor for SC and it is 139.462 times higher risk than those who did not have SC. Similarly, except the factor “Abdominal Pain”, “Menetrier Disease”, “Get Ill Too Much”, “Previous Stomach Surgery”, “Tarry Stools”, “Take Spicy and Salted Food”,

“Education Level”, “Monthly Income”, “Gender”, “Living Area”, “Blood Group”, “BMI”, “Physical Activity” are respectively 66.769, 31.242, 14.235, 10.733, 4.723, 4.512, 4.414, 2.104, 1.812, 1.497, 1.492, 1.457, 1.351 times higher than those who are not related with those factors.

4.2.4 Probability Test

Table 4.3 Probability table of all risk factors

Attribute	Subcategory	Affected Status		Error	Remark	Attribute	Subcategory	Affected Status		Error	Remark
		Yes	No					Yes	No		
Gender						Daily Food in time					
	Female	0.404	0.596	±0.094	Medium		No	0.703	0.297	±0.082	High
	Male	0.551	0.449	±0.07	Medium		Yes	0.368	0.632	±0.07	Medium
Age						Spicy and Salted food					
	30 to 49	0.355	0.645	±0.076	Medium		No	0.282	0.718	±0.082	Medium
	50 to 59	0.57	0.43	±0.105	Medium		Yes	0.639	0.361	±0.07	High
	60 to 70	0.705	0.295	±0.135	High	Green Vegetables					
	Above 70	0.889	0.111	±0.145	Very High		No	0.809	0.191	±0.093	Very High
BMI							Yes	0.409	0.591	±0.063	Medium
	Normal	0.487	0.513	±0.079	Medium	Yellow Fruits					
	Obese	0.286	0.714	±0.193	Low		No	0.806	0.194	±0.076	Very High
	Overweight	0.067	0.933	±0.063	Low		Yes	0.34	0.66	±0.066	Medium
	Severely Underweight	0.971	0.029	±0.057	Very High	Abdominal Pain					
	Underweight	---	---	---	---		No	0.157	0.843	±0.055	Low
Living Area							Yes	0.925	0.075	±0.044	Very High
	Rural	0.575	0.425	±0.072	Medium	Tarry stools					
	Urban	0.347	0.653	±0.094	Medium		No	0.488	0.512	±0.058	Medium
	Suburban	0.565	0.435	±0.072	Medium		Yes	0.818	0.182	±0.228	Very High
Education						Skin color					
	Less than high school	0.672	0.328	±0.07	High		No	0.26	0.74	±0.061	Low
	High school or College	0.419	0.581	±0.112	Medium		Yes	0.98	0.02	±0.027	Very High
	University graduate	0.042	0.958	±0.057	Low	Blood Vomiting					
	Doctoral Degree	---	---	---	---		No	0.492	0.508	±0.057	Medium
							Yes	---	---	---	---

Attribute	Subcategory	Affected Status		Error	Remark	Attribute	Subcategory	Affected Status		Error	Remark
		Yes	No					Yes	No		
Working Status						Get ill too much					
	Business	0.615	0.385	±0.132	High		No	0.279	0.721	±0.065	Low
	Govt. employee	0.28	0.72	±0.176	Low		Yes	0.846	0.154	±0.065	Very High
	Private sector	0.576	0.424	±0.075	Medium	Nausea					
	Un employed	0.276	0.724	±0.115	Low		No	0.246	0.754	±0.06	Low
Monthly Income							Yes	---	---	---	---
	Less than 20K	0.58	0.42	±0.066	Medium	Frequent vomiting					
	20K - 30K	0.422	0.578	±0.144	Medium		No	0.412	0.588	±0.06	Medium
	30K - 45K	0.25	0.75	±0.173	Low		Yes	---	---	---	---
	Above 45K	0.105	0.895	±0.138	Low	Previous stomach surgery					
Family Member							No	0.466	0.534	±0.058	Medium
	2 to 3	0.308	0.692	±0.177	Medium		Yes	---	---	---	---
	4 to 5	0.428	0.572	±0.081	Medium	Stomach Lymphoma					
	Above 5	0.62	0.42	±0.084	High		No	0.453	0.547	±0.059	Medium
Blood Group							Yes	---	---	---	---
	A	0.57	0.43	±0.094	High	Menetrier Disease					
	B	0.543	0.457	±0.102	High		No	0.454	0.546	±0.059	Medium
	O	0.508	0.492	±0.123	High		Yes	0.963	0.037	±0.071	Very High
	AB	0.184	0.816	±0.123	Low	Gastric Medicine					
Physical Activity							No	0.43	0.57	±0.066	Medium
	No	0.279	0.721	±0.134	Low		Yes	0.674	0.326	±0.099	High
	Often	0.456	0.544	±0.118	Medium		Yes	0.674	0.326	±0.099	High
	Regularly	0.566	0.434	±0.071	High						

Table 4.3 describe the likelihood of an individual sub-category occurred ratio. It will help to understand the probability of a sub-categories action could motivate to having a disease or not. If, positive probability (BMI= Severely Underweights affected *probability ~ 0.971*) is moderate or higher then, this factors probability will motivate to having the disease. On the other hand, if negative probability (Nausea = NO *~ 0.754*) is moderate or higher then, its factors probability will motivate to having no disease. Here observed Very High risky sub-category is Age= Above 70 (*Probability ~ 0.889*), BMI= Severely Underweights (*Probability ~ 0.971*), Green Vegetable = NO (*Probability ~ 0.809*), Yellow Fruits = No (*Probability ~ 0.806*), Abdominal Pain = Yes (*probability ~ 0.925*), Tarry stools = Yes (*Probability ~ 0.818*), Skin Color = Yes (*Probability ~*

0.98), Get Ill too Much = Yes (*Probability ~ 0.846*) and Menetrier Disease = Yes (*Probability ~ 0.963*).

Also there are some high risky sub-category has found those are Age = 60 to 70 (*Probability 0.705*), Education = Less than high school (*Probability ~ 0.672*), Working Status = Business (*Probability ~ 0.615*), Family Member = Above 5 (*Probability ~.0.62*), Physical Activity = Regularly (*Probability ~ 0.566*), Daily Food in time = NO (*Probability ~ 0.703*), Spicy and Salted food = Yes (*Probability ~ 0.639*) and Gastric Medicine = Yes (*Probability ~ 0.674*). Except for the rest of them are moderate or low risky subfactors.

4.3 Data mining results

Data mining mainly used to discover knowledge from data. It will find out hidden knowledge from row data in a short and smart way. Where feature selection and predictive Apriori algorithm is a smart procedure to discover knowledge.

4.3.1 Feature Selection

Table 4.4 Feature selection techniques with respect to ranker method

Features	Correlation	GR	IG	Relief	SU
Abdominal Pain	0.764	0.764	0.764	0.764	0.764
Nausea	0.712	0.505	0.466	0.459	0.485
Skin Color	0.679	0.437	0.402	0.288	0.419
Get Ill Too Much	0.554	0.246	0.238	0.170	0.242
Yellow Foods	0.442	0.160	0.149	0.180	0.154
Frequent Vomiting	0.420	0.277	0.169	0.075	0.210
Spicy Salted Food	0.349	0.093	0.090	0.143	0.092
Green Vegetables	0.334	0.111	0.085	0.122	0.096
Daily Food	0.328	0.082	0.079	0.041	0.080
Tobacco Status	0.325	0.080	0.111	0.110	0.093
Education Level	0.323	0.129	0.189	0.159	0.154
Stomach Lymphoma	0.308	0.218	0.093	0.045	0.130
Menetrier Disease	0.291	0.172	0.075	0.030	0.104
Previous Stomach Surgery	0.260	0.195	0.066	0.008	0.099
Gastric Medicine	0.221	0.041	0.036	0.061	0.038
Age	0.211	0.050	0.083	0.046	0.062
Monthly Income	0.211	0.049	0.063	0.056	0.055
BMI	0.185	0.176	0.341	0.156	0.232
Living Area	0.182	0.026	0.033	0.066	0.029
Family Member	0.168	0.026	0.035	0.021	0.030
Working Status	0.164	0.034	0.057	0.067	0.043
Physical Activity	0.145	0.023	0.030	0.031	0.026
Gender	0.140	0.015	0.014	0.065	0.015
Blood Vomiting	0.130	0.138	0.017	0.000	0.030
Tarry Stools	0.124	0.053	0.012	0.002	0.020
Diabetes	0.102	0.020	0.008	0.014	0.011
Blood Group	0.087	0.024	0.046	0.101	0.032
Another Cancer	0.058	0.104	0.003	0.013	0.006
Alcohol Status	0.048	0.064	0.011	0.007	0.019
Poor Appetite	0.037	0.005	0.001	-0.002	0.002
Family History	0.012	0.006	0.004	0.041	0.005
Breast Cancer Status	0.000	0.000	0.000	-0.001	0.000

Table 4.4 represent some most popular feature selection techniques (Correlation, Gain Ratio, Information Gain, Relief and Symmetrical Uncertainty) results which are mostly used to filter medical data. All attribute evaluator is filtered with ranker method and expected rank 1.00 is highly correlated with target variable (Disease) and 0.00 refers there is no relationship among them. From this perspective, it is observed that “Abdominal Pain”, “Nausea”, “Skin Color”, “Get Ill Too Much” is highly correlated with Disease but there are also some features like “Diabetes”, “Another Cancer”, “Alcohol Status”, “Poor Appetite”, “Family History”, and “Breast Cancer Status” those are less or negatively related with Disease.

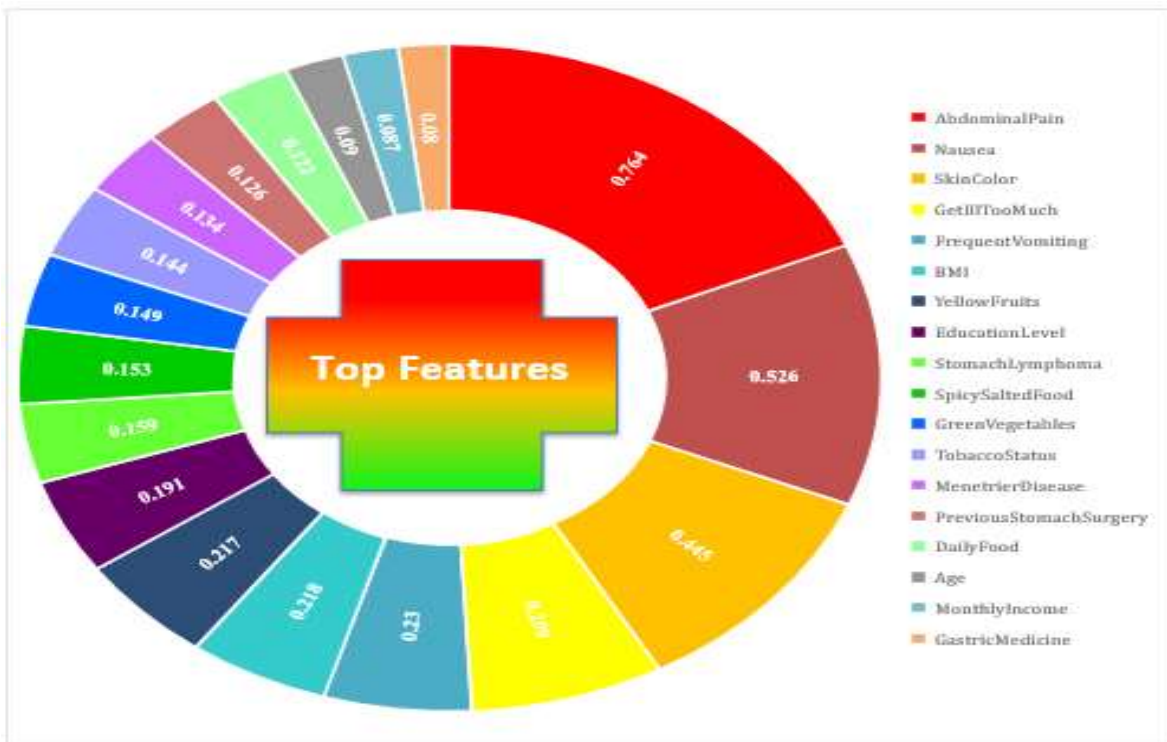


Figure 4.1: Top Feature based on feature selection

It is not clear that which filtering methods are works better on SCs dataset. So, in this perspective we will average of all individual filters ranks from table 4.4, the result will represent a solution to find out top features. Here figure 4.1 Prove the average importance of a single factor and get top eighteen preoperative features which were significantly correlated with SC. Those are, Abdominal Pain (0.764), Nausea (0.526), Skin Color (0.445), Get Ill Too Much (0.299), Frequent Vomiting

(0.13), BMI (0.218), Yellow Fruits (0.217), Education Level (0.191), Stomach Lymphoma (0.159), Spicy and Salted Food (0.153), Green Vegetables (0.149), Tobacco Status (0.144), Menetrier Disease (0.134), Previous Stomach Surgery (0.126), Daily Food (0.122), Age (0.09), Monthly Income (0.89) and Gastric Medicine (0.08).

4.3.2 Yes rules from Apriori

Table 4.5 Best rules for Disease= Yes by Predictive Apriori algorithm

No	LHS	RHS	Support
1	{Age=50 to 59, SkinColor=Yes}	Disease=Yes	0.413
2	{BMI=Severely Underweight, AbdominalPain=Yes}	Disease=Yes	0.31
3	{Age=30 to 49, Nausea=Yes}	Disease=Yes	0.307
4	{BMI=Underweight}	Disease=Yes	0.297
5	{BMI=Normal, SkinColor=Yes}	Disease=Yes	0.29
6	{Education=Less than high school, TobaccoStatus=Yes excessive}	Disease=Yes	0.287
7	{MonthlyIncome=Less than 20k, FrequentVomiting=Yes}	Disease=Yes	0.263
8	{GetIllTooMuch=Yes, SkinColor=Yes}	Disease=Yes	0.253
9	{SpicySaltedFood=Yes, SkinColor=Yes}	Disease=Yes	0.23
10	{DailyFood=No, Nausea=Yes, YellowFoods=No}	Disease=Yes	0.22
11	{GreenVegetables=No, SkinColor=Yes, YellowFoods=No}	Disease=Yes	0.21
12	{TobaccoStatus=Yes Sometimes, AbdominalPain=Yes, Nausea=Yes}	Disease=Yes	0.203
13	{DailyFood=No, AbdominalPain=Yes, PreviousStomachSurgery=No}	Disease=Yes	0.18
14	{SpicySaltedFood=Yes, GetIllTooMuch=Yes, SkinColor=Yes, Nausea=Yes, StomachLymphoma=No}	Disease=Yes	0.137
15	{GreenVegetables=Yes, GetIllTooMuch=Yes, SkinColor=Yes, Nausea=Yes, MenetrierDisease=No}	Disease=Yes	0.137

Table 4.5 represent top fifteen rules to have SC. Where it is examined that “Age = 50 to 59”, “Skin Color = Yes”, “BMI = Underweight”, “BMI =severely underweight”, “Nausea = YEs”, “Education

= Less Than High School”, “Tobacco Status = Yes Excessive” and so on is extremely supported risk level to have SC and “Daily Food =No”, “Stomach Lymphoma=No”, “Menetrier Disease=No”, “Previous Stomach Surgery = No” show low-risk level.

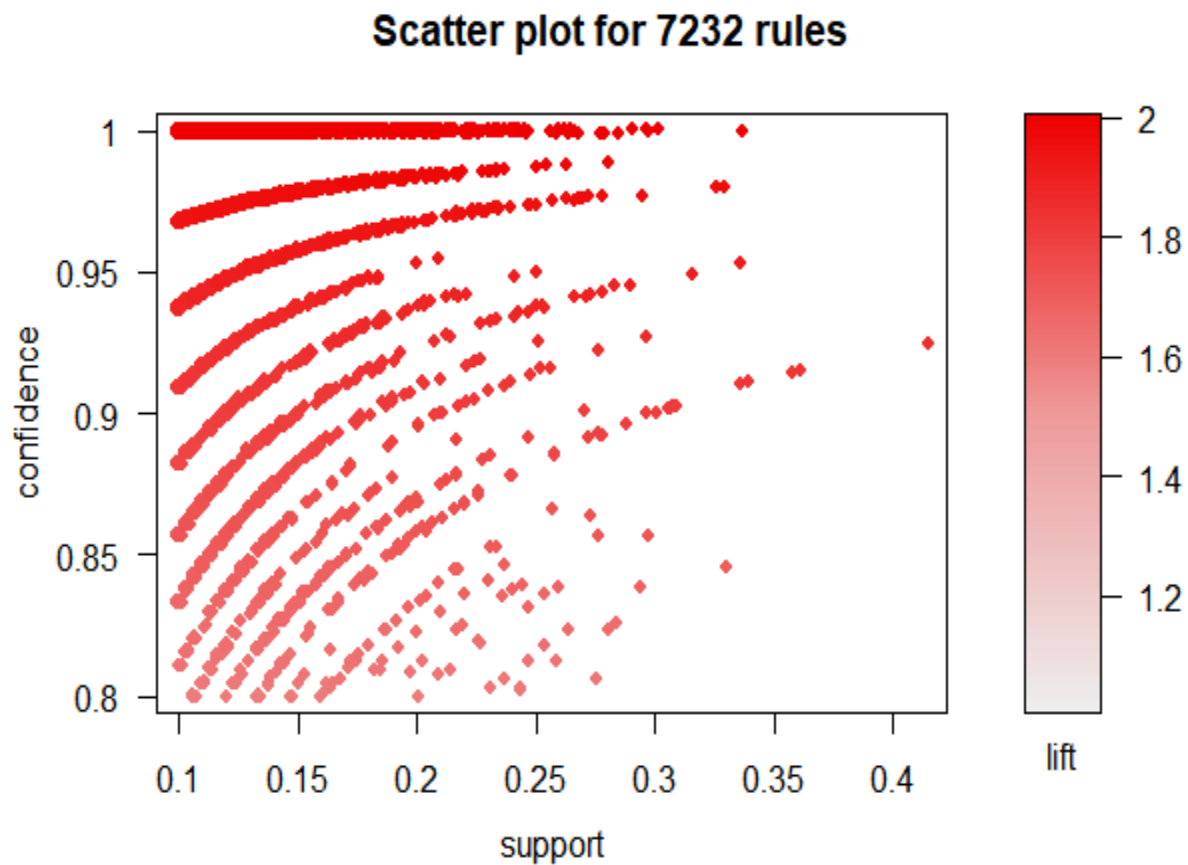


Figure 4.2: Support VS Confidence table with respect to lift for Disease = Yes rules

Figure 4.2 Represent support VS confidence with respect to lift for Disease = Yes rules (N = 7232). Support is count from 0.10 to 0.40, confidence is 0.8 to 1.00 and Lift is count from 1 to 2. Strong rules are always indicated when all parameters values are rising to top value. Here we observe that most of the high rules are available when support between 0.1 to 0.3 and respectively confidence is 1 and lift is 2.

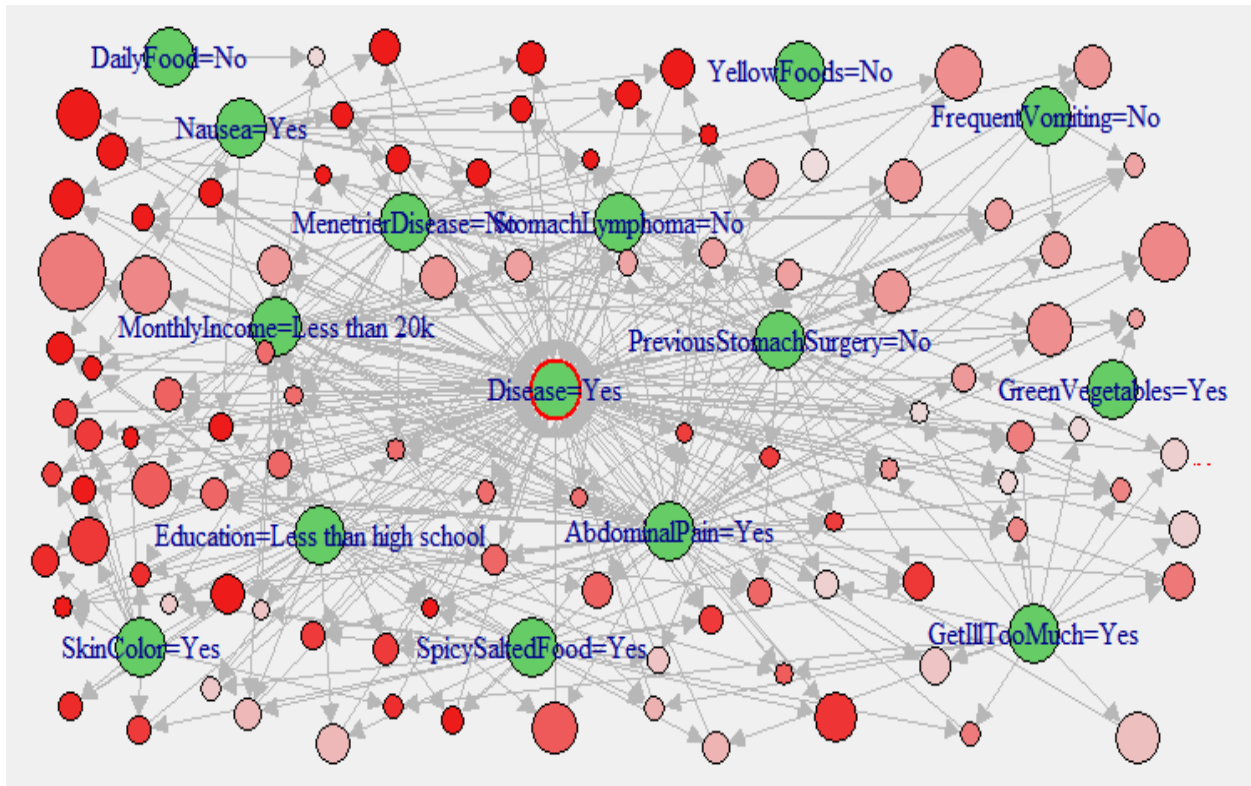


Figure 4.3: Visual relationship among factors for Disease = Yes

Figure 4.3 is very helpful to understand the relationship among some top factors those are responsible to have a disease. Here, bubble size is representing the support and color is represent the confidence. If bubble size goes to bigger than its support level will increase and if color shows dark red then it shows pretty high confidence on this relationship. Here we get, if someones get abdominal pain, have stomach lymphoma, having nausea, education level is less than high school, monthly income is less than 20000 takas, get ill too much, do not take daily food properly, do not eat yellow fruits and vegetables every day, also habited to eat spicy and salted food, and overall subjects skin color turn into pale then he/she could be affected with SC. And some factors like no frequent vomiting, no previous stomach surgery also indicated to have SC.

4.3.3 No rules from Apriori

Table 4.6 Best rules for Disease = No by Predictive Apriori algorithm

NO	LHS	RHS	Support
1	{Education=University graduate,AbdominalPain=No}	Disease=No	0.42
2	{Education=University graduate,StomachLymphoma=No}	Disease=No	0.417
3	{Age=30 to 49,BMI=Overweight}	Disease=No	0.403
4	{Education=High school or college,AbdominalPain=No}	Disease=No	0.4
5	{DailyFood=Yes,SpicySaltedFood=NO} => {Disease=No}		0.4
6	{AbdominalPain=No,MenetrierDisease=No}	Disease=No	0.397
7	{TobaccoStatus=No,YellowFoods=Yes}	Disease=No	0.397
8	{BMI=Normal,MonthlyIncome=Less than 20k, SpicySaltedFood=NO,TobaccoStatus=No,Nausea=No}	Disease=No	0.143
9	{DailyFood=Yes,TobaccoStatus=No,AbdominalPain=No,Nausea=No, FamilyMember=Above 5}	Disease=No	0.137
10	{BMI=Normal,GreenVegetables=Yes,FrequentVomiting=No, StomachLymphoma=No}	Disease=No	0.137
11	{Age=30 to 49,DailyFood=Yes,TobaccoStatus=No, PreviousStomachSurgery=No}	Disease=No	0.133
12	{BMI=Overweight,SpicySaltedFood=NO,SkinColor=No,Nausea=No, FrequentVomiting=No,MenetrierDisease=No}	Disease=No	0.116
13	{DailyFood=Yes,TobaccoStatus=No,SkinColor=No,AbdominalPain=No, Nausea=No,FrequentVomiting=No}	Disease=No	0.113
14	{DailyFood=No,GreenVegetables=Yes,GetIllTooMuch=No,SkinColor=No,Nausea=No, FrequentVomiting=No,StomachLymphoma=No}	Disease=No	0.113
15	{TobaccoStatus=No,GetIllTooMuch=No,AbdominalPain=No,Nausea=No, FrequentVomiting=No,MenetrierDisease=No}	Disease=No	0.113

Table 4.6 also represent top fifteen rules to have no appendicitis. Where it is examined that “Education = University Graduate”, “Abdominal Pain = NO”, “BMI = Overweight”, “Daily Food = Yes”, “Abdominal Pain = No”, “Menetrier Disease= No”, “Spicy Salted Food=NO and so on is

highly supported to have no SC disease and “Tobacco Status=No”, “Age = 30 to 49”, “Stomach Lymphoma=No”, “Skin Color = No” and so are shown low supported value to have SC.

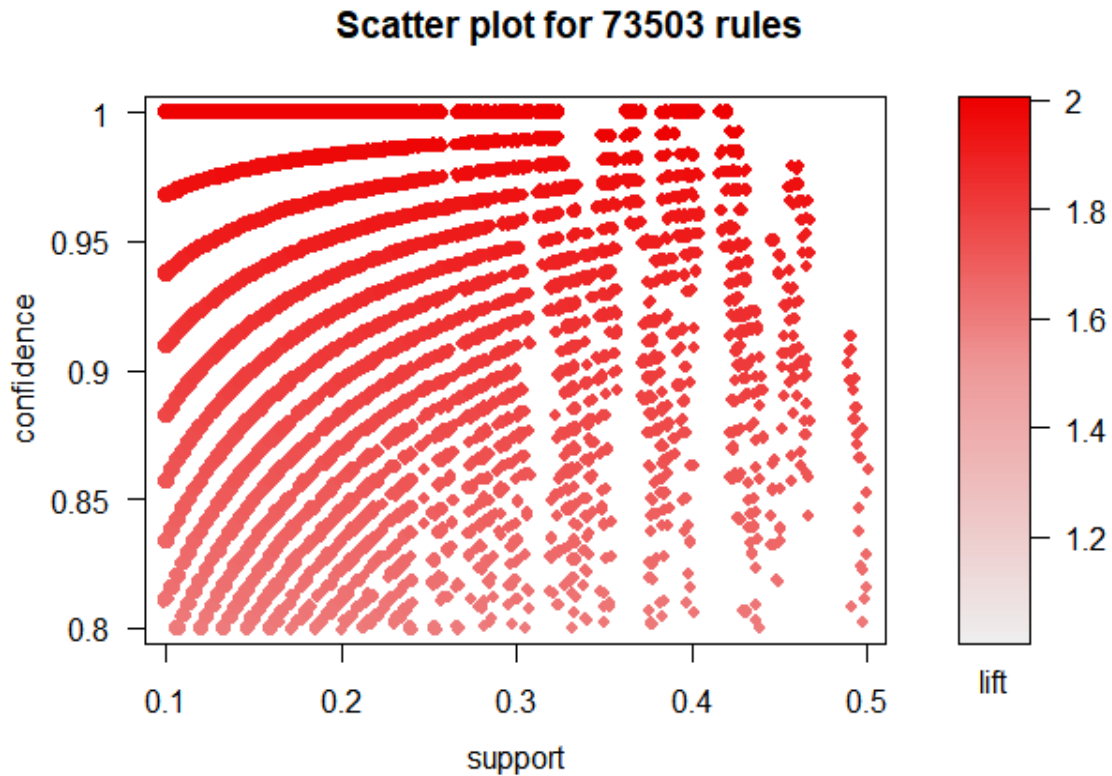


Figure 4.4 Support VS Confidence table with respect to lift for Disease = No rules

Figure 4.4 represent support VS confidence with respect to lift for Disease = No (N= 73503) rules. Support is count from 0.10 to 0.50, confidence is 0.8 to 1.00 and Lift is count from 1 to 2. Strong rules are indicated when all parameters values are rising to top value. Here we observe that highly rules are available when support between 0.1 to 0.4, confidence is 0.95 to 1 and lift is 2.

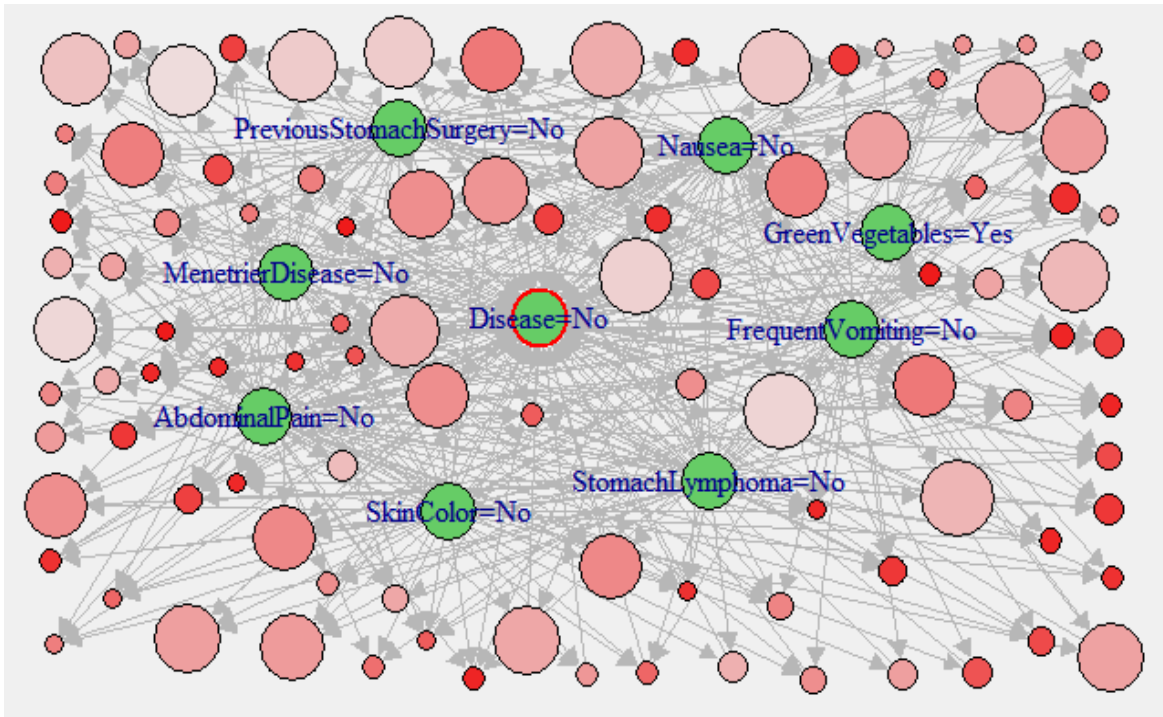


Figure 4.5: Visual relationship among factors for Disease = No

Figure 4.5 is a visual relationship among some top factors those are gathering evidence to have no disease. At the same conditional parameters as figure 4.5 as bubble size represents the support and color is represent the confidence. Bigger bubble size indicates high support level and color dark red shows pretty high confidence among those relationships. Here we get, if someone will not have any stomach lymphoma, menterier disease, nausea, frequent vomiting, abdominal pain, not changed skin color and eats fresh green vegetables every day then he/she will be remaining safe from SC.

4.4 Score Calculation

Table 4.7 Score table for each sub-category

Attribute	Initial Score			Average score	Elegant Score		Final Score
	Yes rules	No rules	probability		P value, X^2 Value		
Age							
30 to 49	3	---	2	2.5	2.5		2.75
50 to 59	4	---	2	3	3		3.3
60 to 70	---	---	3	3	3		3.3
Above 70	--	---	4	4	4		4.35
BMI							
Normal	3	---	2	2.5	0.5		1.05
Obese	---	---	1	1	1		1.95
Overweight	3	--	1	2	2.5		3.45
Severely Underweight	3	---	3	3	3.75		4.7
Underweight	3	---		3	3.75		4.7
Education							
Less than high school	3	---	3	3	3		3.5
High school or College	1	---	2	1.5	1.5		2.4
University graduate	1	---	1	1	1		1.85
Doctoral Degree	---	---	---	0	0		0.8
Monthly Income							
Less than 20K	2	4	2	3	3		3.2
20K - 30K	---	---	2	2	2		2.15
30K - 45K	---	---	1	1	1		1.1
Above 45K	---	---	1	1	1		1.1
Daily Food in time							
No	2	---	3	2.5	2.5		2.9
Yes	---	---	2	2	1.5		1.5
Spicy and Salted food							
No	---	4	2	3	3		1
Yes	2	---	3	2.5	2.5		3.15
Green Vegetables							
No	2	---	4	3	3		3.6
Yes	1	---	2	1.5	1.5		1.5

Attribute	Score			Average score	Elegant Score	Final Score
	Initial Score					
	Yes rules	No rules	probability			
				P value, X^2 Value		
Tobacco Status						
No	---	1	2	1.5	1.5	1.5
Yes sometimes	1	---	4	2.5	2.5	3.05
Yes excessive	3	---	4	3.5	3.5	4.1
Get ill too much						
No	---	4	1	2.5	2.5	2.5
Yes	2	---	4	3	3.75	5.8
Skin color						
No	---	4	1	2.5	2.5	2.5
Yes	3	---	4	3.5	4	6.05
Abdominal Pain						
No	---	1	---	1	1	3.15
Yes	3	---	---	3	3.75	6.15
Nausea						
No	2	---	1	1.5	1.5	3.55
Yes	3	---	---	3	3.75	6.1
Frequent vomiting						
No	---	4	2	0.5	0.5	0.5
Yes	2	---	---	1	1	2
Previous stomach surgery						
No	1	---	2	1.5	1.5	1.95
Yes				0	0	1.45
Stomach Lymphoma						
No	1	---	2	0.5	0.5	0.5
Yes	1	---		1	1	1.7
Menetrier Disease						
No	1	---	2	1.5	1.5	1.5
Yes	1	---	4	2.5	2.5	3
Yellow fruits						
No	2	---	4	3	3	3.9
Yes	---	1	2	1.5	1.5	1.5
Gastric Medicine						
Yes	---	---	3	3	3	3.05
No	---	---	2	2	2	2

Table 4.7 represents the overall sub-factors score in a single table. It has a total of 18 factors with 46 sub-factors individuals score. First of all, the initial score is calculated the average score, elegant score and finally, we get the final score. Each sub-factor score is defined by their importance or impact on the disease. Like Lowest score is Stomach Lymphoma = No (0.5) and Highest score is Abdominal Pain = Yes (6.15). Finally, this table will help to generate risk flow chart.

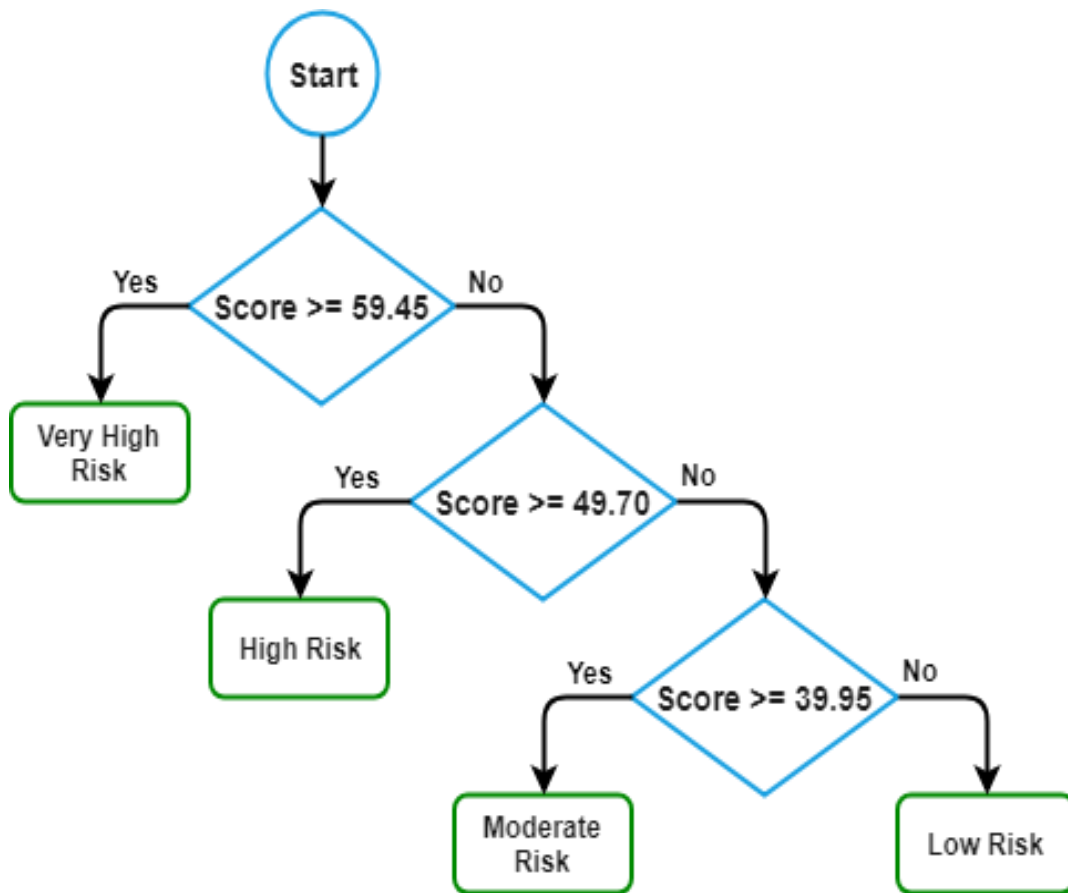


Figure 4.6: SC Risk Prediction Algorithm Flowchart

Figure 4.6 is a conditional flowchart. Our applications algorithm will work based on this flowchart. It will clearly show that if any individual subjects risk score is $\text{Score} \geq 59.45$ then he is in Very High Risk, if $\text{Score} \geq 49.70$ then he/ she is in “High Risk”, if $\text{Score} \geq 39.95$ then he/she is in “Moderate Risk” other wiles subject is in “Low” risk to have SC.

4.5 Application Layout

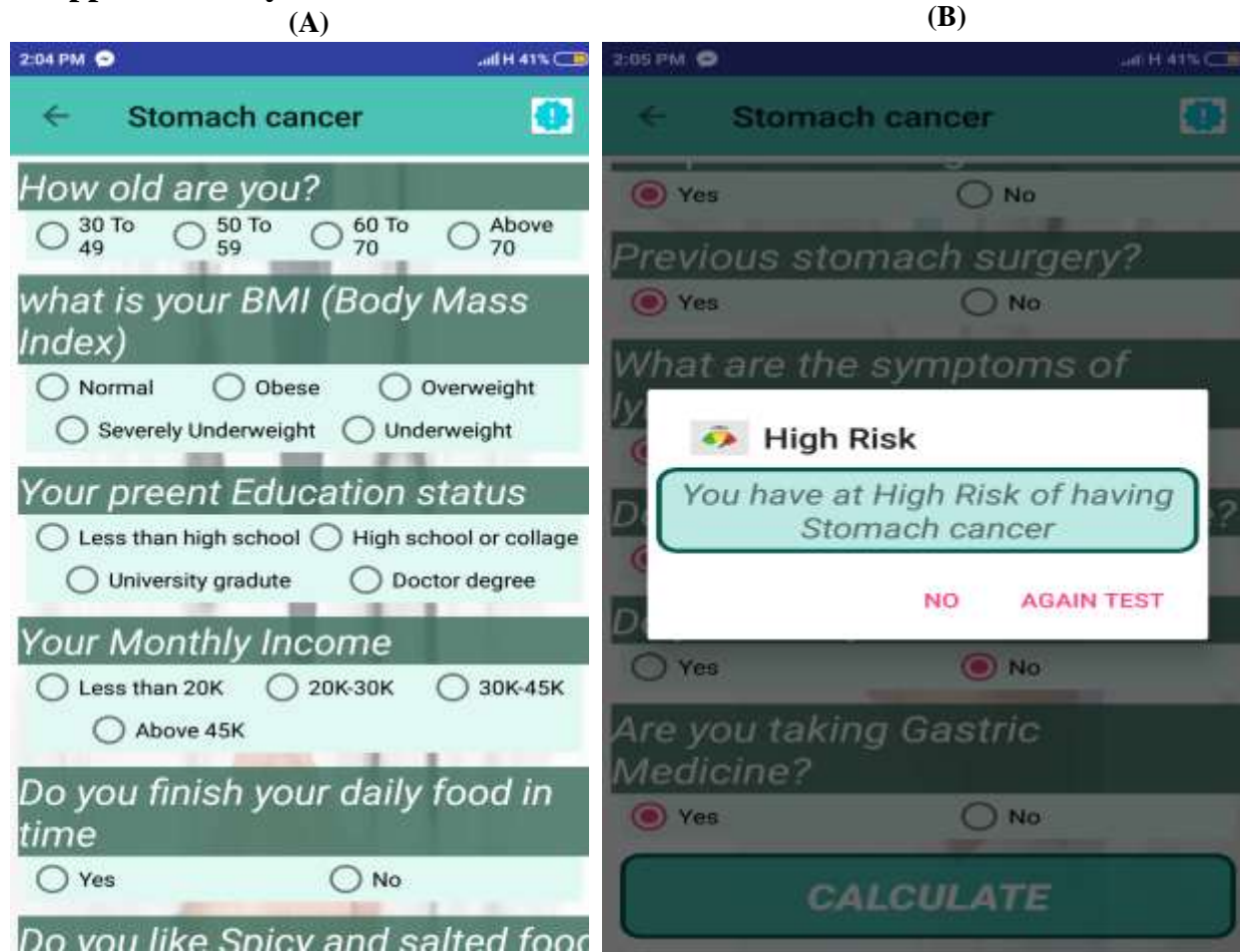


Figure 4.7: (A) application layout and (B) application Results

Developing an application is the final task of this study. Numerous users can easily test them by using this application within a short moment. Only a single Android-based smartphone is enough for them. First of all, the user has to get his application and install it. Then, when they run the application the will found eighteen related questions about SC. User have to answer all the question carefully shows in figure 4.7A. After selecting all questions user just have to press calculate button shows in at the bottom of figure 4.7B. When user press calculates a pop-up screen will come out with user's risk record. Then he/she has the opportunity to test another person or none.

4.6 Discussion

Naturally, disease prevention power decreases by increasing individual's age. When people are reached to older ages, they have been misfortunately affected with one or many deadly diseases like SC. But it is a good thing that the death rate of SC has been decreasing. In here all experimental results are tested with the dependent variable "disease status". From those results, we get, having "Abdominal Pain" is the first most risk factor for SC and it is 66.769 times higher in the case group and also it is found as a significant risk factor of SC in another study (Torpy, et al., 2010). Also having "Nausea" founded as a second most risk factors for SC and "Skin color turn Into Pale" is third most risk factors where it is 139.462 times higher in case group rather than the control group. Mainly it happened for acute anemia (rapid blood loss from the stomach) which was caused by a lack of iron and vitamin B-12 (paleness from <https://www.healthline.com/health/paleness> on 23 Nov 2018). There are also some high-risk factors are observed those are "Menetrier Disease", "Get Ill Too Much", "Previous Stomach Surgery", "Tarry Stools", "Take Spicy and Salted Food", "Education Level", "Monthly Income", "Gender", "Living Area", "Blood Group", "BMI", "Physical Activity" which are significantly found in another research study (Behrens, et al., 2014; Bray, et al., 2018; Cover, et al., 2016; Gelband, et al., 2016) Taking Tobacco, drinking alcohol, Poor appetite, Breast cancer for female, relative cancer as colon cancer, type two diabetes, and family history are not shown as a risk factor in this study. It doesn't mean that these do not risk factor for SC, it risks level could be very low among Bangladeshi peoples because there are many studies shows that these factors are also the most identifiable risk factor for SC (Behrens, et al., 2014; Choi, et al., 2016; Ellison-Loschmann, et al., 2017; Huang, et al., 2018; Suh, et al., 2013). And we will suggest trying to take proper dietary components, nutrition with vitamin A, C, and E every day. Nutrition is protective against stomach lymphoma and vitamin A, C and E are very

protective against stomach disease (Amieva, et al., 2016; Huang, et al., 2018; Nomura, et al., 1990; Ngoan, et al., 2002). Also taking some fresh green vegetable and yellow fruits every day and avoid smoking because this study and including some other study shows that, adequate eating vegetables and fruits can prevent SC significantly and sometimes it could reduce risk level 4-5% (Vingeliene, et al., 2016; Ngoan, et al., 2017).

Remember that, physical exercise is very helpful to prevent any disease, it will produce disease protective hormone in your body naturally and it could be a risk factor of SC if you do not perform any physical activity (Behrens, et al., 2014; Ngoan, et al., 2017).

We believe that prevention is better than cure. Bangladesh is a developing and unhealthy country, most of the people's monthly income is very low (less than 20000 thousand) and they are mostly not educated. Besides this SC prognosis is very low and difficult to identify preoperative symptoms and also its diagnosis is expensive. So, this study will be very helpful for those peoples who want to prevent SC in the initial stage.

4.7 Summary

In the result section, we can observe different result are has been generated by statistical test and data mining operation. All results will indicate that this one factor is related to SC or not. If the factor is related to disease, then how much it is related. To, get this answer we will get, if any person will get abdominal pain, having nausea, education level is less than high school, have stomach lymphoma, monthly income is less than 20000 taka, get ill too much, do not take daily food properly, do not eat yellow fruits and vegetables every day, also habited to eat spicy and salted food, and overall subjects skin color turn into pale then he/she could be affected with SC. And if someone will not have any stomach lymphoma, Menetrier disease, nausea, frequent vomiting, abdominal pain, not changed skin color and eats fresh green vegetables every day then he/she will be remaining safe from SC. Finally, those conditions are implemented on the application and it works pretty well to predict SC risk level.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 Findings and Contributions

Man is mortal, throughout the lifetime he/she will fall into disease randomly several times in every year. It is a natural thing. Disease affected ratio is increased when he/she have become older like SC, it is mainly found whose age is greater than 30-year-old. No one can stop it happens but can control disease affection ratio. It is a very good thing that, prevention is better than cure. All disease is occurred in our body due to having a shortage of some kind of antibody and hormones. And all disease symptoms are shows in our body mostly in a long time period. Sometime it could show one to three years. SC symptoms are like that, it risks factor are not clear enough to normal people, so they don't even think that he/she could be at risk of SC. According to our study, we found that most of the affected people are taking gastric medicine in a long time period, they do not think long time gastric could be a risk factor of SC. Geographical location (East Asian, Russia, Asian Americans, Native Hawaiians etc.) and working sectors (Cole industry, gold ornaments maker, lade, silver, shisha factory) s peoples are in first aid risk factor of SC. Also, there are some riskiest factors are founded those are related to any people in the world. Main risk factors are "Abdominal Pain= Yes", "Skin Color Turn into Pale = Yes", "Nausea = Yes", "Menetrier Disease= Yes", "Get Ill Too Much = Yes", "Previous Stomach Surgery = Yes", "Tarry Stools = Yes", "Take Spicy and Salted Food = Yes", "Education Level =Less than high School", "Monthly Income = Less than 20000", "Gender =Male", "Living Area = Rural", "Blood Group = A", "BMI = Underweight and Severely underweight", "Physical Activity = No", "Taking fruits and vegetables = No" are founded risk factor of SC . Where "Abdominal Pain= Yes", "Skin Color Turn into Pale = Yes", "Nausea = Yes", is the top most risk factor of SC, those are also significantly

founded risk factors in only some studies. But we do not find Tobacco and alcohol as a significant risk factor which is founded as a risk factor in other studies. Good thing is that from last two decades SC death rates are decreasing. It could be prevented if ones take action in the initial stage of cancer. We will recommend all people try to take proper nutrition, vitamin A, E and C, >80g vegetable and fruits every day, perform physical exercise regularly and try to avoid tobacco. Vitamin, nutrition's, fruits and vegetables are very much protective against SC but tobacco decries this prevention power. If it is possible to do that, the low-income country like Bangladesh will be kept safe from SC and capable to save a huge amount of money for the future.

5.2 Future Work

Stomach cancer symptoms have been shown all over the world. In recent years, stomach cancer is increasing at an alarming rate in a developing country like Bangladesh. In this study, we just collect data from NICRH in Bangladesh, not including other government or private hospital and outside countries hospital. The total collected sample size is 300 that was not sufficient to identify all preoperative risk factors correctly and applied technology could have some lacking to take proper calculation. In feature, we will try to collect data from all popular hospitals including private and government if possible, to collect data from other countries in Asia and we will try to use deep learning to predict risk level of SC. In this paper we mainly discuss about Risk Prediction Algorithm and developing implementation smart mobile apps. After this we will add more attributes (Risk Factors) to make vast analysis as well as research on stomach cancer.

References

- Sitarz, R., Polkowski, W. P., Maciejewski, R., & Offerhaus, G. J. A. (2017). Risk of peritoneal dissemination in stomach cancer. *Current Issues in Pharmacy and Medical Sciences*, 30(4), 184-186.
- Kelley, J. R., & Duggan, J. M. (2003). Gastric cancer epidemiology and risk factors. *Journal of clinical epidemiology*, 56(1), 1-9.
- Cheung, K. S., Chan, E. W., Wong, A. Y., Chen, L., Wong, I. C., & Leung, W. K. (2018). Long-term proton pump inhibitors and risk of gastric cancer development after treatment for *Helicobacter pylori*: a population-based study. *Gut*, 67(1), 28-35.
- Ferro, A., Peleteiro, B., Malvezzi, M., Bosetti, C., Bertuccio, P., Levi, F., ... & Lunet, N. (2014). Worldwide trends in gastric cancer mortality (1980–2011), with predictions to 2015, and incidence by subtype. *European journal of cancer*, 50(7), 1330-1344.
- Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., & Jemal, A. (2015). Global cancer statistics, 2012. *CA: a cancer journal for clinicians*, 65(2), 87-108.
- Fang, X., Wei, J., He, X., An, P., Wang, H., Jiang, L., ... & Min, J. (2015). Landscape of dietary factors associated with risk of gastric cancer: A systematic review and dose-response meta-analysis of prospective cohort studies. *European Journal of Cancer*, 51(18), 2820-2832.
- Kunisaki, C., Miyata, H., Konno, H., Saze, Z., Hirahara, N., Kikuchi, H., ... & Mori, M. (2017). Modeling preoperative risk factors for potentially lethal morbidities using a nationwide Japanese web-based database of patients undergoing distal gastrectomy for gastric cancer. *Gastric Cancer*, 20(3), 496-507.
- Hansson, L. E., Nyrén, O., Hsing, A. W., Bergström, R., Josefsson, S., Chow, W. H., ... & Adami, H. O. (1996). The risk of stomach cancer in patients with gastric or duodenal ulcer disease. *New England Journal of Medicine*, 335(4), 242-249.
- Cover, T. L. (2016). *Helicobacter pylori* diversity and gastric cancer risk. *MBio*, 7(1), e01869-15.
- Sitas, F. (2016). Twenty five years since the first prospective study by Forman et al.(1991) on *Helicobacter pylori* and stomach cancer risk. *Cancer epidemiology*, 41, 159-164.

- Amieva, M., & Peek Jr, R. M. (2016). Pathobiology of *Helicobacter pylori*-induced gastric cancer. *Gastroenterology*, *150*(1), 64-78.
- Kim, H. J., Kim, N., Kim, H. Y., Lee, H. S., Yoon, H., Shin, C. M., ... & Kim, Y. H. (2015). Relationship between body mass index and the risk of early gastric cancer and dysplasia regardless of *Helicobacter pylori* infection. *Gastric Cancer*, *18*(4), 762-773.
- Latino-Martel, P., Cottet, V., Druesne-Pecollo, N., Pierre, F. H., Touillaud, M., Touvier, M., ... & Ancellin, R. (2016). Alcoholic beverages, obesity, physical activity and other nutritional factors, and cancer risk: a review of the evidence. *Critical reviews in oncology/hematology*, *99*, 308-323.
- Ellison-Loschmann, L., Sporle, A., Corbin, M., Cheng, S., Harawira, P., Gray, M., ... & Pearce, N. (2017). Risk of stomach cancer in Aotearoa/New Zealand: A Māori population based case-control study. *PloS one*, *12*(7), e0181581.
- Liu, J., Li, X., Lin, T., Dai, L., Zhang, G., Zhang, C., ... & Zhao, Q. (2016). Spatial analysis of gastric cancer morbidity in regions of rapid urbanization: a case study in Xiamen, China. *Stochastic environmental research and risk assessment*, *30*(2), 713-723.
- Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*.
- Hussain, S. A., & Sullivan, R. (2013). Cancer control in Bangladesh. *Japanese journal of clinical oncology*, *43*(12), 1159-1169.
- In, H., Langdon-Embry, M., Gordon, L., Schechter, C. B., Wylie-Rosett, J., Castle, P. E., ... & Rapkin, B. D. (2018). Can a gastric cancer risk survey identify high-risk patients for endoscopic screening? A pilot study. *Journal of Surgical Research*, *227*, 246-256.
- Thrumurthy, S. G., Chaudry, M. A., Hochhauser, D., & Mughal, M. (2013). The diagnosis and management of gastric cancer. *Bmj*, *347*, f6367.

- Ahmed, K., Asaduzzaman, S., Bashar, M. I., Hossain, G., & Bhuiyan, T. (2015). Association assessment among risk factors and breast cancer in a low income country: Bangladesh. *Asian Pac J Cancer Prev*, 16(17), 7507-12.
- Jesmin, T., Ahmed, K., Rahman, M. Z., & Miah, M. B. A. (2013). Brain cancer risk prediction tool using data mining. *International journal of computer applications*, 61(12).
- Ahmed, K., Jesmin, T., & Rahman, M. Z. (2013). Early prevention and detection of skin cancer risk using data mining. *International Journal of Computer Applications*, 62(4).
- Ahmed, K., Kawsar, A. A., Kawsar, E., Emran, A. A., Jesmin, T., Mukti, R. F., ... & Ahmed, F. (2013). Early detection of lung cancer risk using data mining.
- Asaduzzaman, S., Ahmed, K., Chakraborty, S., Hossain, G., Bashar, M. I., Bhuiyan, T., & Chandan, S. S. (2015). Anticipation of the significance of risk factors in cervical cancer for low incoming country: Bangladesh perspective. *International Journal of Scientific & Engineering Research*, 6(11), 876-881.
- Grosso, G., Bella, F., Godos, J., Sciacca, S., Del Rio, D., Ray, S., ... & Giovannucci, E. L. (2017). Possible role of diet in cancer: Systematic review and multiple meta-analyses of dietary patterns, lifestyle factors, and cancer risk. *Nutrition reviews*, 75(6), 405-419.
- Torre, L. A., Sauer, A. M. G., Chen Jr, M. S., Kagawa-Singer, M., Jemal, A., & Siegel, R. L. (2016). Cancer statistics for Asian Americans, Native Hawaiians, and Pacific Islanders, 2016: Converging incidence in males and females. *CA: a cancer journal for clinicians*, 66(3), 182-202.
- Mahmoodi, S. A., Mirzaie, K., & Mahmoudi, S. M. (2016). A new algorithm to extract hidden rules of gastric cancer data based on ontology. *SpringerPlus*, 5(1), 312.
- Naylor, G. M., Gotoda, T., Dixon, M., Shimoda, T., Gatta, L., Owen, R., ... & Axon, A. (2006). Why does Japan have a high incidence of gastric cancer? Comparison of gastritis between UK and Japanese patients. *Gut*, 55(11), 1545-1552.

- Van Cutsem, E., Dicato, M., Geva, R., Arber, N., Bang, Y., Benson, A., ... & Grothey, A. (2011). The diagnosis and management of gastric cancer: expert discussion and recommendations from the 12th ESMO/World Congress on Gastrointestinal Cancer, Barcelona, 2010. *Annals of oncology*, 22(suppl_5), v1-v9.
- Grosso, G., Bella, F., Godos, J., Sciacca, S., Del Rio, D., Ray, S., ... & Giovannucci, E. L. (2017). Possible role of diet in cancer: Systematic review and multiple meta-analyses of dietary patterns, lifestyle factors, and cancer risk. *Nutrition reviews*, 75(6), 405-419.
- Ngoan, L. T., Mizoue, T., Fujino, Y., Tokui, N., & Yoshimura, T. (2002). Dietary factors and stomach cancer mortality. *British journal of cancer*, 87(1), 37.
- de Vries, E., Uribe, C., Pardo, C., Lemmens, V., Van de Poel, E., & Forman, D. (2015). Gastric cancer survival and affiliation to health insurance in a middle-income setting. *Cancer epidemiology*, 39(1), 91-96.
- Gajalakshmi, C. K., & Shanta, V. (1996). Lifestyle and risk of stomach cancer: a hospital-based case-control study. *International journal of epidemiology*, 25(6), 1146-1153.
- Choi, Y. J., & Kim, N. (2016). Gastric cancer and family history. *The Korean journal of internal medicine*, 31(6), 1042.
- Mouw, T., Koster, A., Wright, M. E., Blank, M. M., Moore, S. C., Hollenbeck, A., & Schatzkin, A. (2008). Education and risk of cancer in a large cohort of men and women in the United States. *PloS one*, 3(11), e3639.
- Steinmetz, K. A., & Potter, J. D. (1991). Vegetables, fruit, and cancer. I. Epidemiology. *Cancer Causes & Control*, 2(5), 325-357.
- Chyou, P. H., Nomura, A. M., Hankin, J. H., & Stemmermann, G. N. (1990). A case-cohort study of diet and stomach cancer. *Cancer research*, 50(23), 7501-7504.
- Nomura, A., Grove, J. S., Stemmermann, G. N., & Severson, R. K. (1990). A prospective study of stomach cancer and its relation to diet, cigarettes, and alcohol consumption. *Cancer research*, 50(3), 627-631.

- Xu, H. L., Tan, Y. T., Epplein, M., Li, H. L., Gao, J., Gao, Y. T., ... & Xiang, Y. B. (2015). Population-based cohort studies of type 2 diabetes and stomach cancer risk in Chinese men and women. *Cancer science*, *106*(3), 294-298.
- Polom, K., Marrelli, D., Pascale, V., Roviello, G., Voglino, C., Rho, H., ... & Roviello, F. (2016). High-risk and low-risk gastric cancer areas in Italy and its association with microsatellite instability. *Journal of cancer research and clinical oncology*, *142*(8), 1817-1824.
- Sierra, M. S., Cueva, P., Bravo, L. E., & Forman, D. (2016). Stomach cancer burden in Central and South America. *Cancer epidemiology*, *44*, S62-S73.
- Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, *2*(2), 271-277.
- Sebastião, Y. V., & Peter, S. D. S. (2018, December). An overview of commonly used statistical methods in clinical research. In *Seminars in pediatric surgery* (Vol. 27, No. 6, pp. 367-374). WB Saunders.
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality. *Biometrika*, *52*(3), 591-611.
- Tallarida, R. J., & Murray, R. B. (1987). Chi-square test. In *Manual of Pharmacologic Calculations* (pp. 140-142). Springer, New York, NY.
- Rumsey, D. J. (2015). *U Can: statistics for dummies*. John Wiley & Sons.
- Kolmogorov, A. N. (2018). *Foundations of the Theory of Probability: Second English Edition*. Courier Dover Publications.
- Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, *300*, 70-79.
- Cilia, N. D., De Stefano, C., Fontanella, F., & di Freca, A. S. (2018). A ranking-based feature selection approach for handwritten character recognition. *Pattern Recognition Letters*.

- Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1-2), 245-271.
- Oh, J. H., Al-Lozi, R., & El Naqa, I. (2009, December). Application of machine learning techniques for prediction of radiation pneumonitis in lung cancer patients. In *Machine Learning and Applications, 2009. ICMLA'09. International Conference on* (pp. 478-483). IEEE.
- Hall, M. A., & Smith, L. A. (1997). Feature subset selection: a correlation based filter approach.
- Cooper, G. F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4), 309-347.
- Jantawan, B., & Tsai, C. F. (2014). A comparison of filter and wrapper approaches with data mining techniques for categorical variables selection. *International Journal of Innovative Research in Computer and Communication Engineering*, 2(6), 4501-4508.
- Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: introduction and review. *Journal of biomedical informatics*.
- Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)* (pp. 856-863).
- Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P. (2013). Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems with Applications*, 40(4), 1086-1093.
- Torpy, J. M., Lynn, C., & Glass, R. M. (2010). Stomach cancer. *Jama*, 303(17), 1771-1771.
- <https://www.healthline.com/health/paleness> on 23 Nov 2018
- Gelband, H., Sankaranarayanan, R., Gauvreau, C. L., Horton, S., Anderson, B. O., Bray, F., ... & Gupta, S. (2016). Costs, affordability, and feasibility of an essential package of cancer control interventions in low-income and middle-income countries: key messages from Disease Control Priorities. *The Lancet*, 387(10033), 2133-2144.

- Behrens, G., Jochem, C., Keimling, M., Ricci, C., Schmid, D., & Leitzmann, M. F. (2014). The association between physical activity and gastroesophageal cancer: systematic review and meta-analysis. *European journal of epidemiology*, 29(3), 151-170.
- Huang, Q., & Lew, E. (2018). Epidemiology and risk factors. In *Gastric Cardiac Cancer* (pp. 39-49). Springer, Cham.
- Suh, M., Choi, K. S., Lee, Y. Y., & Jun, J. K. (2013). Trends in cancer screening rates among Korean men and women: results from the Korean National Cancer Screening Survey, 2004-2012. *Cancer research and treatment: official journal of Korean Cancer Association*, 45(2), 86.
- Amieva, M., & Peek Jr, R. M. (2016). Pathobiology of *Helicobacter pylori*-induced gastric cancer. *Gastroenterology*, 150(1), 64-78.
- Vingeliene, S., Chan, D. S., Aune, D., Vieira, A. R., Polemiti, E., Stevens, C., ... & Norat, T. (2016). An update of the WCRF/AICR systematic literature review on esophageal and gastric cancers and citrus fruits intake. *Cancer Causes & Control*, 27(7), 837-851.
- Raihan, M., Mondal, S., More, A., Sagor, M. O. F., Sikder, G., Majumder, M. A., ... & Ghosh, K. (2016, December). Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design. In *Computer and Information Technology (ICCIT), 2016 19th International Conference on* (pp. 299-303). IEEE.