Prediction for the risk of oral cancer based on the general external factors

BY

Rawnak Jahan Juthi
ID: 152-15-6044

A S M Noor Ud Doha Shawon
ID: 151-15-4921

Md Kamrul Hasan

ID: 153-15-6476

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

Supervised By
Zerin Nasrin Tumpa
Lecturer
Department of CSE
Daffodil International University

Co-Supervised By
Dewan Mamun Raza
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JULY 2020

APPROVAL

This Research based Project titled **“Identify the risk of oral cancer based on the most general factors**”, submitted by Rawnak Jahan Juthi (152-15-6044), A S M Noor Ud Doha Shawon (151-15-4921) and MD Kamrul Hasan (153-15-6476) to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering (B.Sc.) and approved as to its style and contents. The presentation has been held on 2th May 2020.

BOARD OF EXAMINERS

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

Dr. Fizar Ahmed **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Tarek Habib **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Mohammad Shorif Uddin **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that, this research project has been done by us under the supervision of **Zerin Nasrin Tumpa, Lecturer,Department of CSE** and Co-Supervised by **Dewan Mamun Raza Lecturer,Department of CSE** ,Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Zerin Nasrin Tumpa
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Dewan Mamun Raza
Lecturer
Department of CSE
Daffodil International University
Submitted by:

Shawon
A S M Noor Ud Doha Shawon
ID: 151-15-4921
Department of CSE
Daffodil International University

Rawnak Jahan
Rawnak Jahan Jithi
ID: 152-15-6044
Department of CSE
Daffodil International University

Kamrul

Md Kamrul Hasan
ID: 153-15-6476
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

We really grateful and wish our profound our indebtedness to **Zerin Nasrin Tumpa**, **Lecturer** and**,** Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor and co-supervisor in the field of "*Identify the risk of oral cancer based on the most general factors*" to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

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

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

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

ABSTRACT

The cancer of surface on mouth is grown in men as well as women. Oral fatality rate is an average of 13%. In the world about 657,000 oral cavity and pharynx new cases are appeared and about 330,000 deaths every year [1]. At the initial stage most of the time the primary lesions are not taken seriously and after a long time it may cause oral cancer and sometimes into serious disease. In south East Asia most of the patients who are suffering from oral cancer is female. The dominance of oral cancer is considered to be 10th reason of death. From a survey of WHO almost 49,000 oral cancer patients are suffering [2]. We undertook some factors that are responsible for oral cancer, are mentioned here for our experiment. This paper represents the risk of oral cancer based on the common factors in south East Asia and especially in Bangladesh. As the factors are almost nearest but most of the cases the identifying reason is to taking betel and betel nut in Bangladesh. 60% of oral cancer patients are regarding the age of above 40. In this paper we have made the result and our prediction with a trial of data from 856 patients [3]. We have explored the data sets including the risk factors which work as major cause for oral cancer. So we have worked to propound an exploration to identify the risk for oral cancer by analyzing the most common factors that presents. We have worked and find a simile for 8 diverse algorithm to make our results appropriate as our outcome. Our trial will sum up a new way to identify the risk for oral cancer. People who are invaded but the general symptoms can be predicted the risk percentage of future possibility for oral cancer. Surgical lack is very challenging because of behavior of cancer. Our study may help as a clinical prediction at a very early stage of the patients, people will be benefited and may avoid to be the sufferer of aggressive disease. As sometimes people bear the usual symptoms that are accountable for oral cancer in future it is necessary to identify the risk at an initial stage. Suppose a patient having the risk bears common habit or common symptoms from a certain period of time, the basic risk factors are considered here for this research. The motive of this work is to impel clinical practice for make the possible prediction of risk of this aggressive health issues. We took datasets based on that and using this it will be helpful to predict the risk by the most common factors. Our developed method will help to predict the risk for oral cancer at a very imminent period by undertaking the measure.

TABLE OF CONTENTS

LIST OF FIGURES

PAGE NO

LIST OF TABLES PAGE NO

CHAPTER 1
INTRODUCTION

1.1 Introduction

Oral cancer is considered as common cancer in the world now. Especially in the south East Asia region the percentage of this cancer is not negligible. At a very early stage oral cancer is not found out because sometimes a lesion in the surface of the mouth can cause oral cancer in the long run. Most of the people are not serious about the pre-cancerous stage. In the south East Asia most of the oral cancer patient are women. The common reason is taking betel and betel nut for a long period. People in different region have different habit and cause to hold the risk of this health issue. In the world the main reasons are smoking, tobacco use, and alcohol are mainly responsible. But this risk agent seems variant in the south East Asia and especially in Bangladesh. Here India, Nepal, Sri lank and many countries are facing this health risk [4]. Most of the time it causes for continuous habit and it's rare to have genetic mutation case in Bangladesh. At the time of carrying the symptoms like initial oral cavity, lesions on the soft palate on mouth is identically not predictable because it is deliberated as common issue. So a study of usual risk agent is needed to add a new opportunity to reduce the rate of oral cancer. An aim of development a practice of prognosis the risk for every general syndrome for every patient. The whole world count a major death rate for oral cancer in every part of the countries. We thought it as a frightening exposure for nations as well as whole world. In this study the most common factors that are responsible for oral cancer has been tested. Challenging fact that is mentioned here to predict the possibility to rise up oral cancer after a overlong time based on the common factors or basic symptoms. In the data mining field the possibility to find out the way to predict the upcoming cause of diseases is significant. In the previous time researchers tried to develop clinical tools to identify oral cancer at an early stage. They develop systems, feature, and identity common symptoms, the case of oral cancer for their internal and external factors. Our study has the strength to predict the chance to occur oral cancer after a long time as we undertook the most possible factors occur in south East Asia most commonly in Bangladesh. In this project we have explored the common external factors for future possibilities of oral cancer that is helpful for clinical betterment. We made a survey on patients who are suffering from oral cancer. By taking the symptoms that the patients face before oral cancer, we develop a system to identify the percentage of possibilities to face invasive disease after long or a certain period of years. Our system take the basic and general factors and make a prediction. This prediction make a sense to future diagnosis. Our experiment prove that in Bangladesh or south East Asia there are some usual practice to suffer from oral cancer. Researchers develop various systems to identify oral cancer by image analyzing, genetic exploration and their mutations. In some research paper it is mentioned that the external factors such as age, gender, some day to day habit, smoking, betel and betel nuts, genetic mutations, UV-Ray and sometimes environmental cause play a vital role for oral cancer [5]. It is possible to detect oral cancer but to detect oral cancer at initial stage is not possible as there no clinical tool for detect it before. Our study can implement the feasible prediction or that. We use data mining technique to predict the percentage of the risk. We use Naive Baise theorem for this prediction. This project works as a different way to predict oral cancer when the initial symptoms are arisen at an early stage.

Opportunities:

This project plays a vital role and create a new opportunity to make a prediction that can add a new dimension to detect oral cancer based on the external factors. In this research we have tried to organize a set of data to create a picked result to reduce the rate of this dangerous disease. A study of external factor can take a revolution in clinical practice. This approach can be helpful for factor analysis as the common change or symptoms are seen earlier sometimes. There is no dimension existing to mark the venture at an advance stage of health damage. So early prediction is most helpful to reduce the death rate and develop the system. It is graceful for medical sector to predict the chances to occur oral cancer that can help for the betterment of treatment. Uncourtly this study raise a recent degree for risk identification.

1.2Motivation

As Oral cancer is placed at 7-8% in the whole cancer in the world. So this is important to stay on a prediction that can give a result for possible cases. As the external factors which is most common are like betel and betel nut, gatka, smoking, tobacco use, oral hygiene etc. A wide number of people are not conscious about their oral hygiene [6]. Our considered key factors are taking betel and betel nut, tobacco, age, smoking. Majority of male is addicted to smoking. At the same time a huge number of male and female are taking betel. In Bangladesh a low portion of female is also involved in smoking. A study said that people at the of 40-49 is countable patients of oral cancer. In

Bangladesh the ratio of this habit is increasing. Comparing to raising the risk the less people are ignoring the common symptoms. As a result it creates a fatal disease. The death ration is increasing day by day in world by this cancer. Thought about the risk of oral cancer an early hypothesis is really necessary. We made a survey and see that most of the patients in our country and south East Asia are female and their age is above 40. As the main reason for oral cancer in this region is betel gatka that can be protected by avoiding this when a simple lesion arise in the soft or hard palates or any other places on mouth. Most of the cases initial lesions are considered as normal and not get is as a reason for future invasive tumor or disease. Patients may come with the very early symptoms but these are not taken as serious. As a result after the long time is can cause oral cancer. So our experiment can help to go out from this and make a possible prediction to detect the risk of oral cancer based on the external factors.

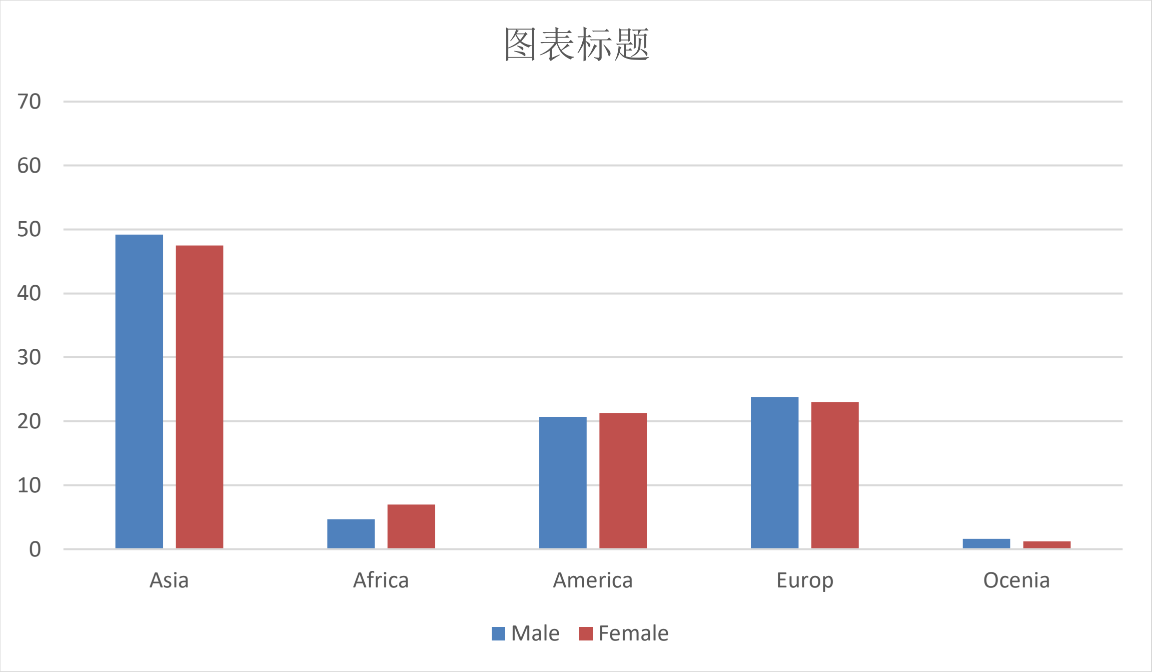


Figure 1.1: Graph for region the percentage of this cancer

1.3Rationale of study

In this project the most challenging case is to collect factors and analysis data. We have to undertake the real data and worked for it. To work with clinical data is tough as we have to identify external factors avoiding the internal cause. We get the result which ensure accuracy rate and better representation and we have to go through with various algorithms, their calculations and need to have the set of our Indisputable data. The fundamental principles of our study related the risk factors as much as possible. Gene mutations are supposed to take lightly because it happens rare. The growth of tissues or cells are abnormal and having genetic imbalance and that cannot predict before affecting. So others external part is analyzed here. The reason behind oral cancer are directly connected to the outer habit or lack of proper health care’s specially lack of mouth hygiene and bad practice so we took them as our datasets on hand.

1.4 Research Questions

We have collected data based on some questions. Patients who bear the common factors have been collected. This questions work as essential data for prediction. This research holds the title "A survey to identify the risk factors for oral cancer". Information about patients admitted or treated in the Bangabandhu Sheikh Mujib Medical University oral and maxillofacial surgery unit from (BSMMU) had been collected. Usual habit or reasons that are responsible for aggressive form of cancer had been included here. Questions work as level gaining of every future outcome. Questions are categorized by the risk agent. We included age of patients and consider it as a key factor for oral cancer. The age starts from 10 and mark to above 40. Another factor is gender because it has already been mentioned that in Bangladesh the ratio of oral cancer in women is larger than men. So gender is another multiplicand. Excessive use of alcohol is another main cause for oral cancer although this is rare in Bangladesh. Smoking is taken as another main multiplier as there is a high risk of mouth organ damage for smoking. In this data we have included smoking period also. People who are involved for a long period for smoking bear the risk for oral cancer. So smoking period 5 years to above 15 years has been considered. People have a habit for using tobacco. Tobacco is very harmful for our body especially for different parts of mouth. Tobacco is also a remarkable factor. Besides the habit of smoking in someone's family, lack of consciousness of oral hygiene, proper habit of diet or lack of nutrition, not taking human papillomavirus vaccine, bears long time premalignant lesions, gene mutations case, having sharp teeth’s are taken as granted factors. Our main key factor is taking betel and betel nuts. According to the explanation from experts, most of the oral cancer affected patients have a strong habit of taking betel and betel nuts.

1.5 Expected Output

A particular age, gender and a habit of taking betel, tobacco, and smoking are our key working factors. We considered this certain part because this are mainly suspected reason for oral cancer. Genetic, chemical or environmental cause is severely taken here as their participation is very low in this case. We need to analysis a certain amount of data to prepare datasets to predict how the general cause affected to create the risk for invasive diseases. Different algorithm give us different anticipation. 10 different algorithms that we choose to make the better output. We want to build a system to have better accuracy by applying Naive Basie theorem and another 9 algorithms to get the expected result. The ratio for each and every factors make the percentage for possibility. A certain algorithm from which we also able to generate a probable result. Our expected prediction has come when our systems can predict the result by taking the common risk factor. By identifying the risk for oral cancer it is possible to reduce the risk as early as possible. External multiplicands should strongly be analyzed because most of the time patients bear the common problems like precancerous lesions. Survival rate for patients, their treatment effectiveness, decomposition decoration, early detection and many Probabilistic will be created. Early detection is important because treatment is much more operative than later. A development of clinical treatment is possible in future for reducing oral cancer and can help to add a new dimension.

1.6Report Layout

Chapter 1: Introduction

This section is consist of the discussion on oral cancer, the reason behind this, how and why we consider the factor as key point and a short notes on overall appearance of this disease worldwide.

However the introduction, motivation of our work, rationale of the study, figure of the project, previous work on the identification of oral cancer are mentioned here. Furthermore our expected outcome and the survey questions has been added that followed with the report layout.

Chapter 2: Background

In this part we have speculated the background history of our thesis based project. We have described the about the trial of the work. Besides some related work on our project, overall summary, scope of the problem then the challenges we faced has been discussed here.

Chapter 3: Research Methodology

In this section we have discussed about the methodology of this work. We have speculated our datasets and the procedure. The required implementation analysis has also been prated here to reach our goal.

Chapter 4: Experimental results and discussion

In this chapter, deals with experiment outcome and its description of the process. Algorithms that are used and its functions are also included. However, we have a discussion about experimental result, descriptive analysis based on our project. Lately, we have included a summary of our project.

Chapter 5: Summary, Conclusion, recommendation and implication for future research

This is the final step where we have summarized the whole study and end up with conclusion. Lastly there is also have some recommendation and discussion about future work of the project.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Cancer is placed at 7th common cancer in the world. It is necessary to generate a way to predict the possibility of oral cancer. In the world the most usual cause is the habit of smoking. But the reason for suffering oral cancer is varied from one region to another in worldwide. Although in the whole world men are the main sufferer from oral cancer, it creates a variety of factors in the south Asia. Especially in Bangladesh oral cancer generally occurs among women and the reason is to taking betel nut and gatka. The major reason is taking betel and betel nut. As many people have the habit to take betel so they are not conscious about their oral hygiene. Most of the cases they ignore the earliest symptoms and in future it results into an incurable disease sometimes in tumor and cancer. Age is also a major factor for oral cancer.

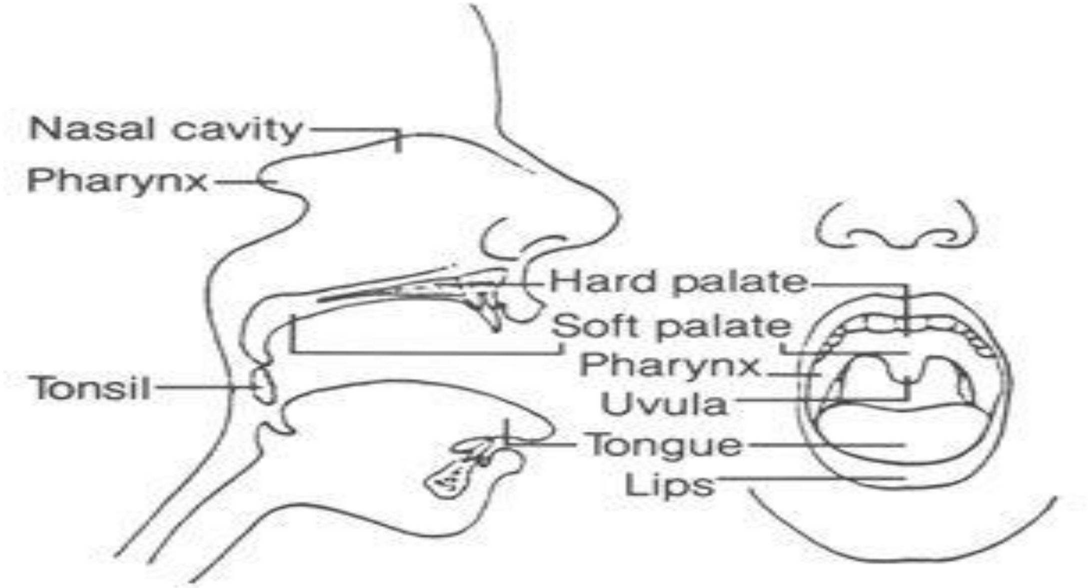


Figure 2.1: oral cancer affecting parts.

A short description of internal and external parts of mouth is shown in this figure. The picture shows a part nasal cavity that is a breeze filled area back to nose. Then pharynx is located back to the nasal cavity and mouth. Tonsil is a couple of limp tissue near to the pharynx. Mouth is consists of hard and soft palate. Hard palates are put in with two bones of fore skeleton. An extent organism is behind tongue is known as Uvula that is mentioned in figure. Tongue, lips are shown is the figure as parts of mouth. We set out mind to collect the real data from the hospitals that includes personal actuality of patients. We collected data and go through with different ages of people. Our aim is to make a general prediction for future possibility. We developed our systems to reach our goal. In the modern time people are addicted in smoking, being alcoholic, taking gatka, betel and so many things to get addicted. People are suffering from various types of cancers in the whole world. Oral cancer varies from person to person and mainly grow with a certain age, gender and their habit. The most common oral cancer types are mentioned here in the image. From all the types of oral cancer gene mutations case is rare. Oral cancer also have an impact of environmental cause like the painter professional have a high risk of having the disease. Sometimes lack of nutrition and food habit may turn into oral cancer. UV ray and any harmful ray emitted from sun is the cause of invasive situation of various parts of mouth [7].

2.2 Related Work

In the study of Data mining the diseases prediction is one of the most essential field. Many experts have worked hard to develop clinical tools or established methods to develop the systems to identify oral cancers from time to time. Researchers also develop techniques, models and processes.

Early detection of oral cancer:

In this paper the initial identity of oral cancer and the PMD (potential malignant disorder has been built up. P.J ford and C. S Farah used recent oral mucosal screening approach to detect oral cancer for future research and detection. They developed a promoted way to detect oral cancer based on general aspects [8]. It helps to identify oral cancer at exact time. For getting the excellent outcome they used the disease process and developed a five years survival rates that was not taken hand before. It helps to detect early stage of cancer and decreases decomposition.

Oral cancer detection using data mining technique:

Paranormal cells are efficient to attack the other cells. From oral cavity or any premalignant lesions results in carcinoma in situ and in the long run causes cancer. As most of the time oral cancer is detected at the last stage so it is possible to cure it at a rare case. Here data sets are collected from various patients and work with three different classification algorithm.

Bayesian classification model, a priory algorithm and SVM classification algorithm has been used here. These techniques are helpful for classification results. It helps doctors for clinical process of diagnosis.

Data mining technique for predicting oral cancer:

In this paper single tree, decision tree, and tree boost based on classification algorithm. Neha Sharma and Hari Om developed the better efficient model to identify oral cancer [9]. The impressibility of this model is 100%. Here the tree boost model is appeasement for the prediction.

2.3 Scope of the problem:

Some scope of this project are given below

1. Alleviate invasive diseases.
2. Reduce oral cancer rate.
3. Early awareness for oral cancer.
4. Adding new dimension for clinical purposes.
5. Identify risk of oral cancer.
6. Model of result to analysis common factors.

2.4 Challenges:

1. Difficult to collect data from patients.
2. Rare recorded of data in the hospital as there is no smart systems to save up data.
3. Tough to manage datasets.
4. Difficult process.
5. Differentiate the data that varies from age to age.
6. To work with a specific number of data.
7. Developing country like Bangladesh have lack of automated systems to store the record of patients. It is highly difficult to reach one patients to another to collect data in hospitals. For oral cancer maximum patients are not admitted in the hospitals so it is tough to reach them also. That make a barrier to collect data.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Research methodology includes the technique of research, and creates the way to appreciate data. It represents the methods, process, tools that are effective for research. By going through the previous work on oral cancer we found the part of our research. It is clear to us that the very early stage is out to destine till now. Lack of tools and methods of identification is the most dangerous cause for a huge number of death. The parameters are included smoking, taking betel and betel nuts, age, gender and the common keys that plays a vital role and responsible for oral cancer. So following the previous work our aim of the project was selected to mark out the risk of oral cancer a very initial stage. The process for our work comprehend with a number of steps. A set of data has been collected from hospitals as well as online resources. We use TCGA and ICGA data portal. Data sets has been prepared and applied algorithms to make the best possible sequel. The knowledge of this object is needed to analyzed. All outcomes are resolved separately based on different algorithms and methods. Although oral cancer is possible to identify but the challenge is to make sure the risk at first. Accuracy rate or prediction is the sign of authenticity of our work. The datasets are detached in the steps of applying algorithms. In our research we have use 10 different algorithms.

3.2 Research Subject and Instrumentation

Subject: Prediction for the risk of oral cancer based on the general external factors

Instrumentation:

- Google survey form.
- Question form.

Software Requirements:

- Core i5 laptop.
- Weka 3.9.

3.3 Data Collection Procedure:

- Generate survey question

- Information gathering
- Entry information into the google form
- Generate csv format for information
- Research and information of various data mining algorithm

For data collection we have to generate some basic questions that are the external factors for occurring oral cancer. We have to meet up with the patients. We reached about 746 patients to collect data. All the data has been stored in a google form as inputs of our research. Then we have to convert this as csv file for this inputs. Finally we have used different data mining algorithms and generates reports.

3.4 Statistical Analysis

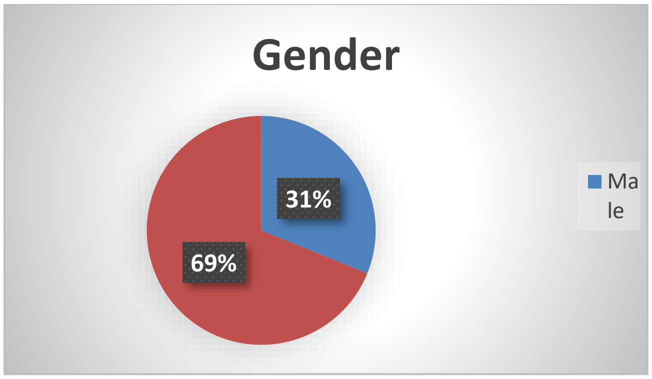
For the experiment we took total Seven Hundred and Forty five (745) data. There are fifteen (15) questions which shows how much an individual involves, interacts and in which way they operate their mobile phones. These fifteen (15) questions were same & mandatory for everyone. The pie graphs were indicating the percentage which makes the result understanding easier.

A survey to identify the risk factors for oral cancer

Information about patients admitted in the Bangabandhu Sheikh Mujib Medical University oral and maxillofacial surgery unit (BSMMU)

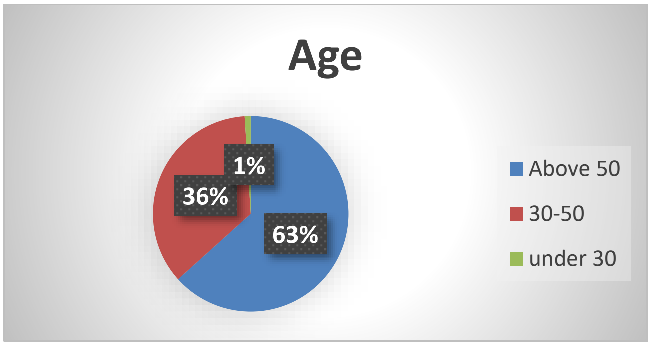
Name of patient:

- Gender-
- Male
- Female



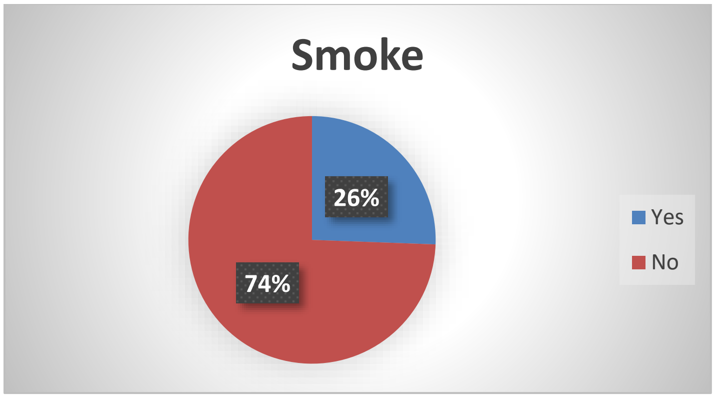
According to the statement from patients it was identified that the percentage of affected patients is including 31% male and 69% female.

- Age-
    - 10-30
    - 30-40
    - Above 40



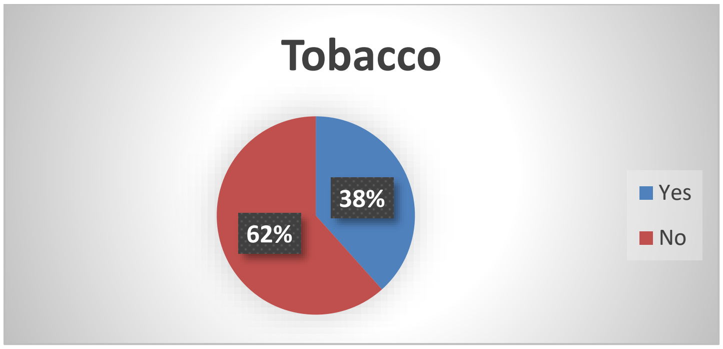
Major number of patients suffering from oral cancer we meet and took statement, their age were above 40 most of cases. The remarkable age of oral cancer diagnosed patients includes the age of under 30 years and ratio is 1%. At the age of 30-40 years holding the ratio of 63%. The high ratio is on the age of above 40.

- Do the patient involve with smoking?
    - Yes
    - No



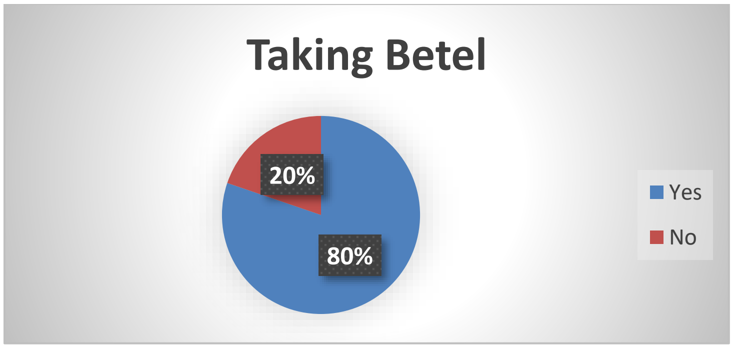
The reason smoking that holds the percentage of having addicted is 74% and non smoker is 26%. The patients whom we meet at the time of collecting their statements the number of patients was noted as per the ratio.

- Have the history of tobacco use?
  - Yes
  - No



At the time of taking the acknowledgement from patients at hospitals we came to know that the percentage of tobacco use patients is 62% and negetive case is 38%.

- Have the history of taking Betel?
  - Yes
  - No



Most of patients whom we talked had a strong habit of taking betel and betel nuts. The most responsible factor in is taking betel and betel nuts. Betel and betel nuts taking ratio is 80% and negative is 20%.

Working process:

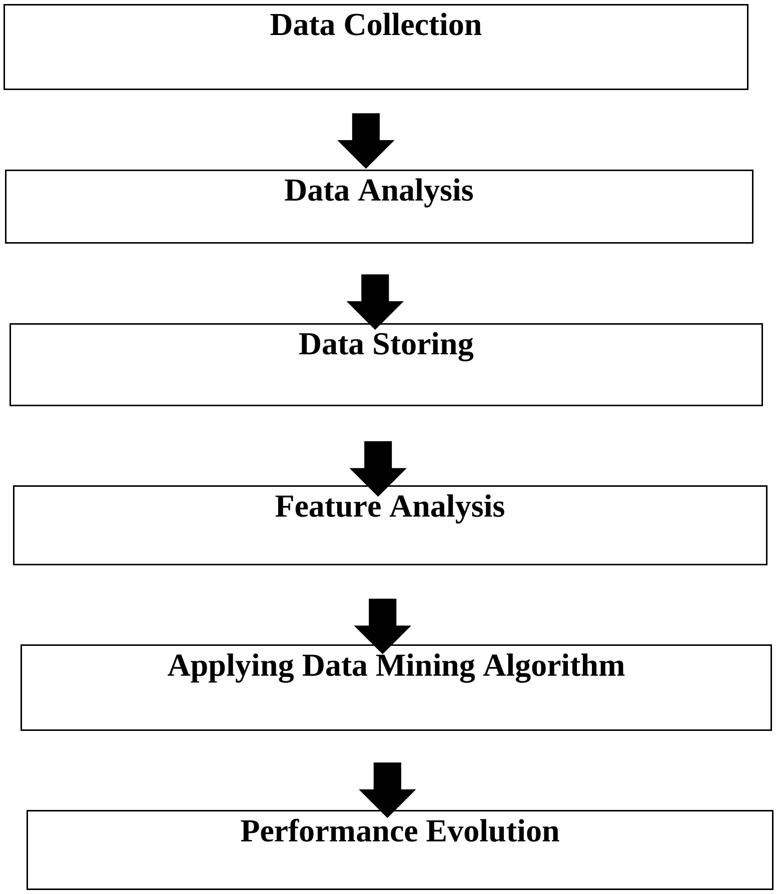


Figure 3.1: Flow of working process

3.5 Implementation Requirement:

To implement the work we used weka and different algorithms.

1. Participants: Collecting information from patients by a survey question.
2. Sorting and remove data: As it is necessary to sort data so after collecting data from hospital we sort those data and remove the irrelevant data.
3. Algorithms: We need some different algorithms to implement our work

CHAPTER 4

EXPERIMENTAL RESULTS & DISCUSSION

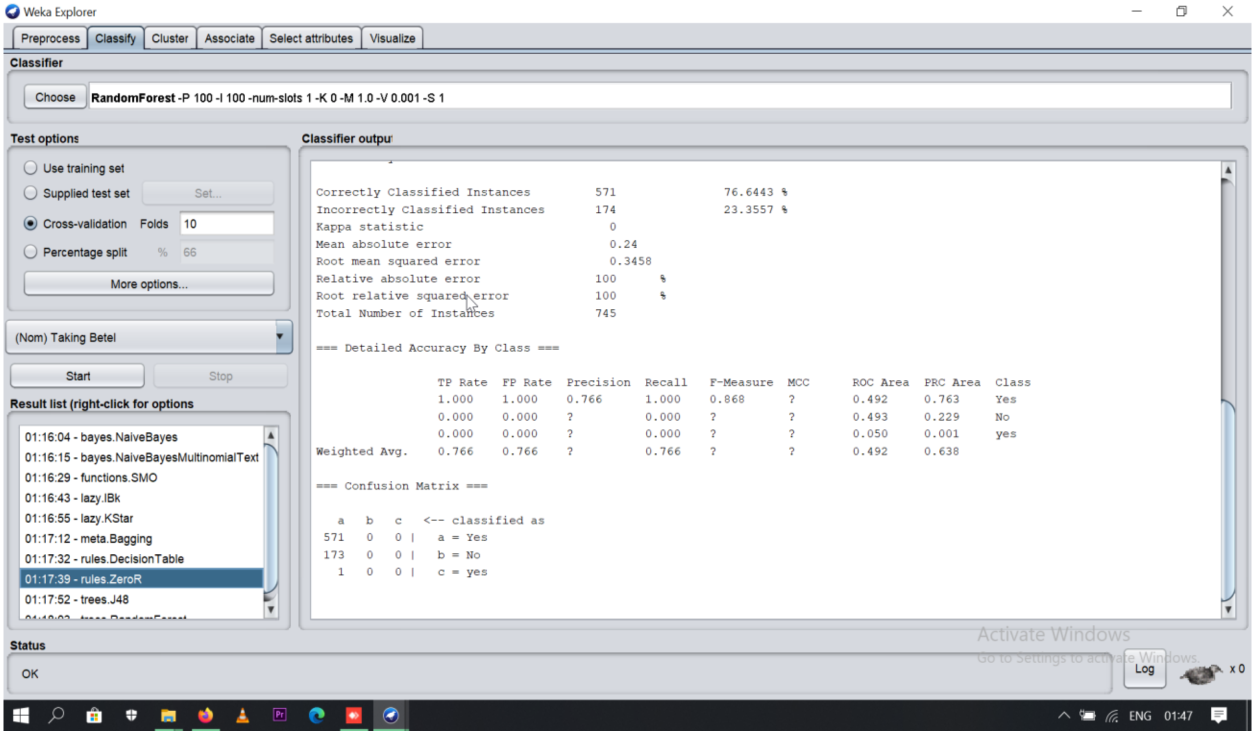
4.1 Introduction

For our research we have used the most efficient algorithm to get the best result as possible. The algorithm and methods are specifically mentioned in the research methodology in the certain section. In our experiment we keep focus on the datasets so that we can implement the better outcome for this research. By our work we are able to specify the challenges and the future possibilities of this work.

4.2 Experimental Results

We have used various types of algorithms and the best output generate from algorithm.

We have made a list of tables for the algorithms below.



RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Taking Betel

```
01:16:04 - bayes.NaiveBayes
01:16:15 - bayes.NaiveBayesMultinomialText
01:16:29 - functions.SMO
01:16:43 - lazy.IBk
01:16:55 - lazy.KStar
01:17:12 - meta.Bagging
01:17:32 - rules.DecisionTable
01:17:39 - rules.ZeroR
01:17:52 - trees.J48
```

```
Correctly Classified Instances         571               76.6443 %
Incorrectly Classified Instances       174               23.3557 %
Kappa statistic                          0
Mean absolute error                      0.24
Root mean squared error                  0.3458
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              745

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.766      1.000    0.868      ?        0.492     0.763     Yes
                 0.000    0.000    ?          0.000    ?          ?        0.493     0.229     No
                 0.000    0.000    ?          0.000    ?          ?        0.050     0.001     yes
Weighted Avg.    0.766    0.766    ?          0.766    ?          ?        0.492     0.638

=== Confusion Matrix ===

   a   b   c   <-- classified as
 571   0   0 |   a = Yes
 173   0   0 |   b = No
   1   0   0 |   c = yes
```

OK

Figure 4.1: Rules-zeroR Algorithm Result

We also construct a table for showing result for Ten (10) different algorithms.

The confusion matrix is a table to describe the performance of a classification model on a set of test data. Confusion matrix can define four terms:
True Positive (TP): we predicted result as disease which are actually disease.

True Negative (TN): we predicted result as not disease which are actually not disease.

False Positive (FP): we predicted disease, but these are not actually disease.

False Negative (FN): we predicted not disease, but these are actually disease.

Precision: precision is the piece of related instances among the retrieved instances. High precision means that an algorithm returned substantially more relevant results than irrelevant ones.

$$precision = \frac{tp}{tp + fp}$$

Recall: Recall is the piece of relevant instances that have been retrieved over the total amount of relevant instances. High recall means that an algorithm returned most of the relevant result.

$$Recall = \frac{tp}{tp + fn}$$

F-measure: f-score is a measure of test’s accuracy by considering both precision and recall. It is a harmonic average of precision and recall.

$$F - score = 2 * \frac{precision * recall}{precision + recall}$$

Accuracy: accuracy refers to the familiarity of the measured value to a known value.

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

False Positive Rate: Calculate the false positive rate by the given equation:

$$False\ positive\ rate = \frac{FP}{TN + Fp}$$

Table 4.1: Performance Metrics of Data Mining Algorithms

Algorithm name	TP Rate	FP Rate	Precision	Recall	F-measure	MCC	ROC Area
Lazy.IBk	0.541	0.507	0.530	0.541	0.532	0.036	0.500
Tree.J48	0.471	0.595	0.443	0.471	0.450	-0.135	0.416
Lazy. Star	0.494	0.541	0.488	0.494	0.491	-0.048	0.496
Naive Bayes	0.553	0.505	0.537	0.553	0.538	0.051	0.533
Trees.RandomForest	0.565	0.482	0.554	0.565	0.556	0.086	0.490
Bagging	0.600	0.478	0.588	0.600	0.571	0.142	0.595

Functions.SMO	0.494	0.593	0.448-	0.494	0.453	-0.117	0.451
Rules.ZeroR	0.576	0.576	0.576	0.576	0.731	-0.051	0.457
Rules.Decision Table	0.765	0.630	0.449	0.756	0.853	0.174	0.731
NaiveBayesMultinomialText	0.766	0.766	0.766	0.766	0.868		0.492

Algorithm table 4.1 shows the TP rate, FP rate, Precision, recall, F-measure, MCC and ROC-area values for 10 different algorithms’

Bar graph for Accuracy:

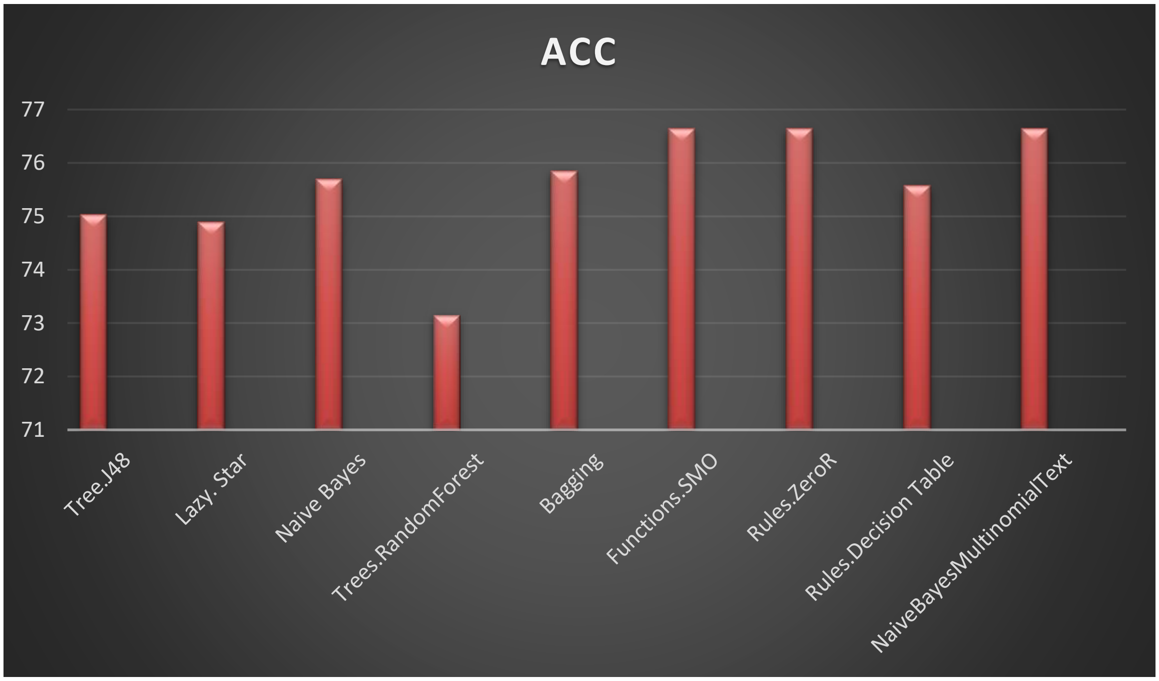


Figure 4.2: Bar Chart for ACC

Precision-recall (PRC) graph

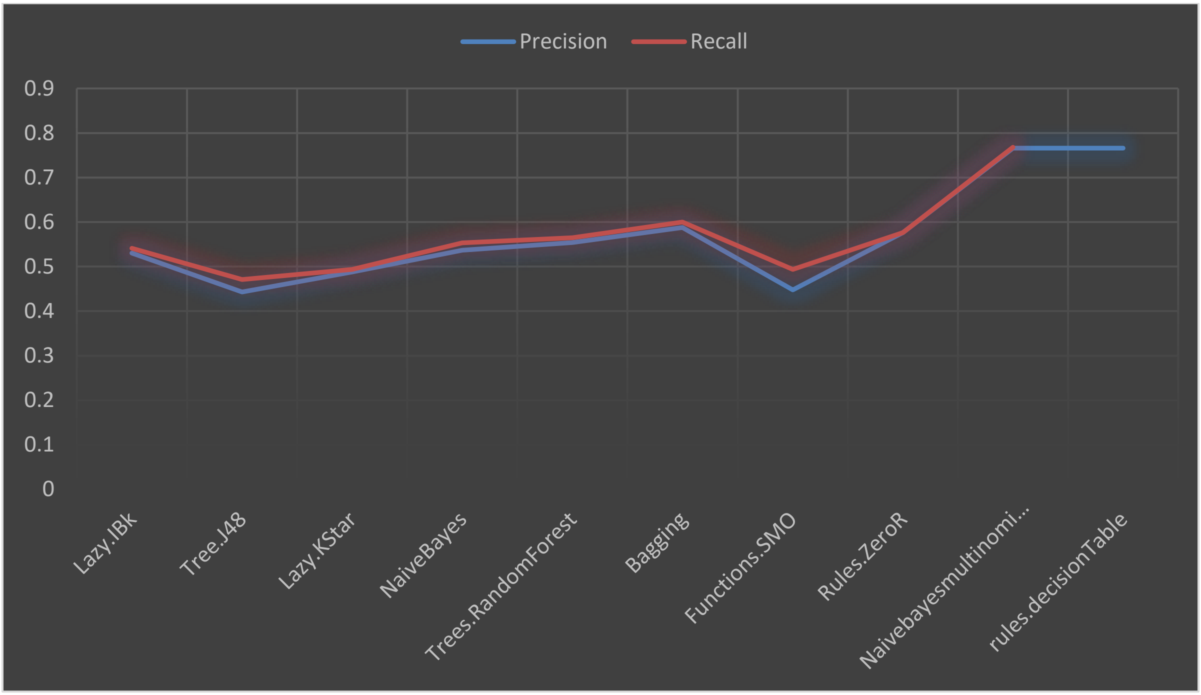
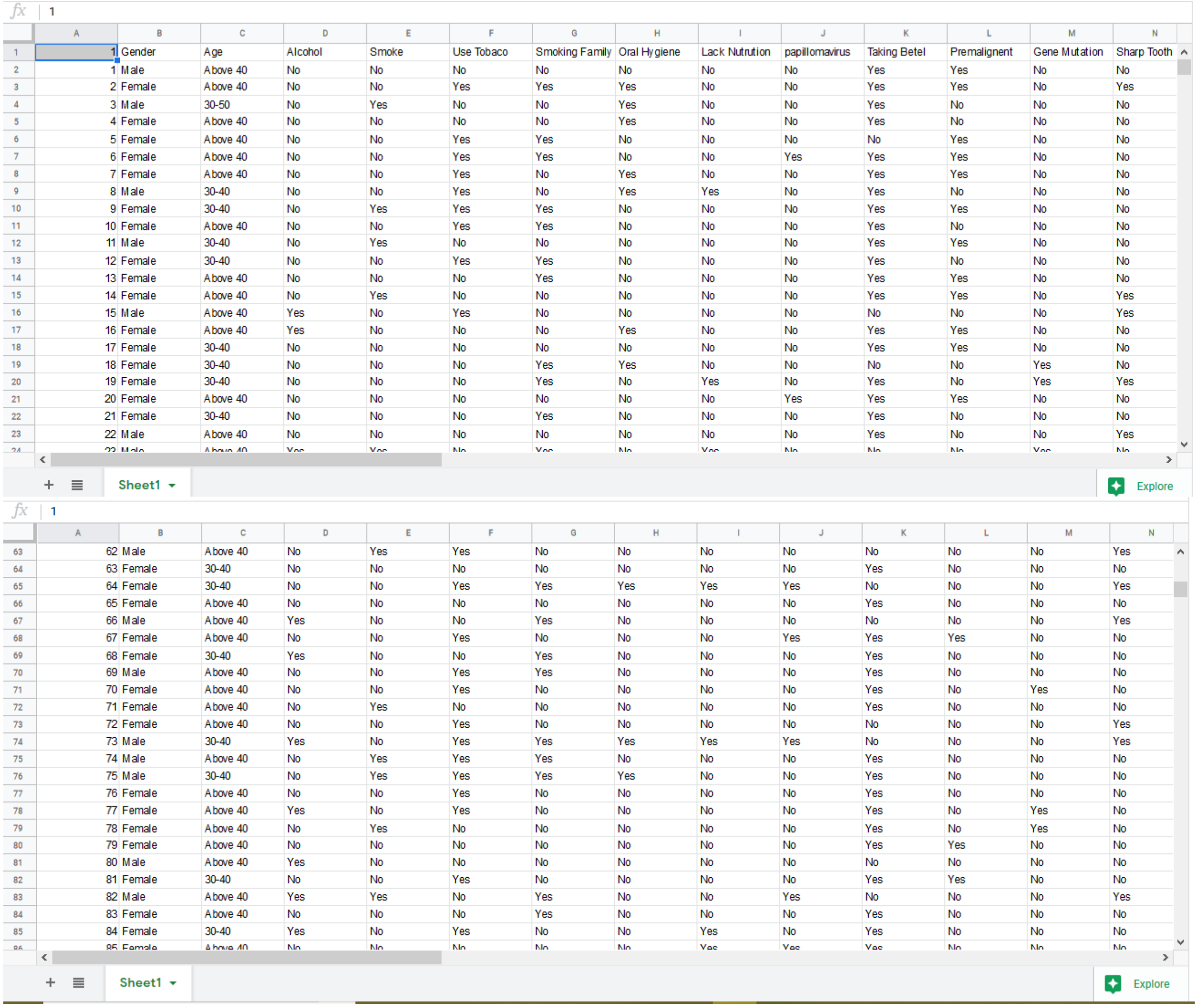


Figure 4.3: Line Graph for PRC

4.3 Descriptive Analysis

We collected the data from 745 patients. We set up our datasets with the real data. In the survey question we have 15 questions for factors identification. We remove the null values, missing values to get good accuracy. Accuracy level is varies from one algorithm to another. The best output generate from algorithm that was specified in the experimental result. The following datasets are given here that we generate from our data.

Picture of google form here



fx 1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	1	Gender	Age	Alcohol	Smoke	Use Tobaco	Smoking Family	Oral Hygiene	Lack Nutrution	papillomavirus	Taking Betel	Premalignent	Gene Mutation	Sharp Tooth
2	1	Male	Above 40	No	No	No	No	No	No	No	Yes	Yes	No	No
3	2	Female	Above 40	No	No	Yes	Yes	Yes	No	No	Yes	Yes	No	Yes
4	3	Male	30-50	No	Yes	No	No	Yes	No	No	Yes	No	No	No
5	4	Female	Above 40	No	No	No	No	Yes	No	No	Yes	No	No	No
6	5	Female	Above 40	No	No	Yes	Yes	No	No	No	No	Yes	No	No
7	6	Female	Above 40	No	No	Yes	Yes	No	No	Yes	Yes	Yes	No	No
8	7	Female	Above 40	No	No	Yes	No	Yes	No	No	Yes	Yes	No	No
9	8	Male	30-40	No	No	Yes	No	Yes	Yes	No	Yes	No	No	No
10	9	Female	30-40	No	Yes	Yes	Yes	No	No	No	Yes	Yes	No	No
11	10	Female	Above 40	No	No	Yes	Yes	No	No	No	Yes	No	No	No
12	11	Male	30-40	No	Yes	No	No	No	No	No	Yes	Yes	No	No
13	12	Female	30-40	No	No	Yes	Yes	No	No	No	Yes	No	No	No
14	13	Female	Above 40	No	No	No	Yes	No	No	No	Yes	Yes	No	No
15	14	Female	Above 40	No	Yes	No	No	No	No	No	Yes	Yes	No	Yes
16	15	Male	Above 40	Yes	No	Yes	No	No	No	No	No	No	No	Yes
17	16	Female	Above 40	Yes	No	No	No	Yes	No	No	Yes	Yes	No	No
18	17	Female	30-40	No	No	No	No	No	No	No	Yes	Yes	No	No
19	18	Female	30-40	No	No	No	Yes	Yes	No	No	No	No	Yes	No
20	19	Female	30-40	No	No	No	Yes	No	Yes	No	Yes	No	Yes	Yes
21	20	Female	Above 40	No	No	No	No	No	No	Yes	Yes	Yes	No	No
22	21	Female	30-40	No	No	No	Yes	No	No	No	Yes	No	No	No
23	22	Male	Above 40	No	No	No	No	No	No	No	Yes	No	No	Yes

fx 1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
63	62	Male	Above 40	No	Yes	Yes	No	No	No	No	No	No	No	Yes
64	63	Female	30-40	No	No	No	No	No	No	No	Yes	No	No	No
65	64	Female	30-40	No	No	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes
66	65	Female	Above 40	No	No	No	No	No	No	No	Yes	No	No	No
67	66	Male	Above 40	Yes	No	No	Yes	No	No	No	No	No	No	Yes
68	67	Female	Above 40	No	No	Yes	No	No	No	Yes	Yes	Yes	No	No
69	68	Female	30-40	Yes	No	No	Yes	No	No	No	Yes	No	No	No
70	69	Male	Above 40	No	No	Yes	Yes	No	No	No	Yes	No	No	No
71	70	Female	Above 40	No	No	Yes	No	No	No	No	Yes	No	Yes	No
72	71	Female	Above 40	No	Yes	No	No	No	No	No	Yes	No	No	No
73	72	Female	Above 40	No	No	Yes	No	No	No	No	No	No	No	Yes
74	73	Male	30-40	Yes	No	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes
75	74	Male	Above 40	No	Yes	Yes	Yes	No	No	No	Yes	No	No	No
76	75	Male	30-40	No	Yes	Yes	Yes	Yes	No	No	Yes	No	No	No
77	76	Female	Above 40	No	No	Yes	No	No	No	No	Yes	No	No	No
78	77	Female	Above 40	Yes	No	Yes	No	No	No	No	Yes	No	Yes	No
79	78	Female	Above 40	No	Yes	No	No	No	No	No	Yes	No	Yes	No
80	79	Female	Above 40	No	No	No	No	No	No	No	Yes	Yes	No	No
81	80	Male	Above 40	Yes	No	No	No	No	No	No	No	No	No	No
82	81	Female	30-40	No	No	Yes	No	No	No	No	Yes	Yes	No	No
83	82	Male	Above 40	Yes	Yes	No	Yes	No	No	Yes	No	No	No	Yes
84	83	Female	Above 40	No	No	No	Yes	No	No	No	Yes	No	No	No
85	84	Female	30-40	Yes	No	Yes	No	No	Yes	No	Yes	No	No	No

Figure 4.4: Dataset

Link of google form here

https://docs.google.com/spreadsheets/d/12f8tDiGyFGHWNbXqix9_Cp_2VvTqvPREszd4LTDSKrE/edit#gid=0

4.4 Summary

According to WHO oral cancer deaths reached 1.90% in total deaths in Bangladesh and world ranks in the 2nd.About 15,010 people die every year for oral cancer in Bangladesh. We undertook the general symptoms and maximum time we made the prediction correctly [10]. When a patient come with this symptoms our prediction can help to identify the risk of oral cancer that may usually turns into an invasive disease in future. It will help to identify initial risk of oral cancer patients.

As the regular habit of a person that make the way to take it into an aggressive damage of soft palate of mouth, tongue, teeth, lip and a serious disease may affect and most of the time the symptoms are ignored by the people who may suffer after a certain period so we have completed a way based on common prefix. Our work make a motivation to be conscious of avoiding from the habit that one should be avoided which are responsible behind oral cancer.

CHAPTER 5

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

5.1 Summary of the Study

In the world oral cancer is placed at 7th most common cancer. Every year a huge number of people die for oral cancer. It varies from country to country and continent to continent. In the south Asia region oral cancer is places 4th position among all cancers. It is difficult to identify oral cancer at an early stage. Especially in Bangladesh women are the main sufferer of this disease. The habit they absorb is taking betel and betel nut. There are various types of oral cancer and occurs in different parts of the mouth. It affects the soft palate of the tongue, berets esophagus, lip, teeth etc. Some external factors that are common for oral cancer in most of the cases are age, gender, daily habit, precancerous lesions, gene mutations, oral cavity and sharp tooth. The rare factor of oral cancer is gene mutations. It is rarely happened for oral cancer in future. It is possible to identify this disease but not in the imminent stage. Most of the time oral cancer is diagnosed at the stage when it is not possible to cure or very rare to improve the condition of patients. So a dimension is actually needed to identify the risk and that can helps to improve system to make a prediction of oral cancer [11]. In the modern world people are now a little bit conscious about their oral hygiene. As a result every year thousands of people die because of unconsciousness. When many of us pass by the lesions affected on mouth and consider it as a normal lesion. But this lesion may turn into a premalignant lesion or carcinoma in situ in future. Our aim to make a prediction to identify the risk so that it can be possible to overcome the danger and make it easier to help the pathological stage of oral cancer identification.

5.2 Conclusions

Many people of the world is now suffering from oral cancer besides all other cancer. In this research field we undertook a datasets from patients based on the most common factors that are responsible for oral cancer, carcinoma in situ and aggressive tumors. We used appreciable methods to make the possible accurate prediction for the risk. Diagnosis of cancer and its treatment is essential to minimize the death rate. Early prevention is most important for all cases. When oral cancer reached at high stage then the treatment can go wrong or not more efficient [12]. So early diagnosis of cancer is our aim in this study. People are suffering physically and mentally by affecting with this kind of fatal disease. So if we want to reduce the rate of oral cancer we have to conscious early. A prediction is essential to make a world better and make a prevention to oral cancer. Furthermore our work also make people conscious about their oral hygiene and keep healthy. According to our plan we used different types of algorithm that gives the best possible prediction.

5.3 Recommendations:

Our recommendations,

1. Real time data is needed as oral cancer affecting reason varies from region to region. Online data can be included.
2. The implementation methods must depends on the most responsible factor.
3. Development of data should be needed.
4. Increase conscious about oral hygiene and the basic reason for oral cancer.
5. Make the prediction easier to enlarge improvement of the diagnosis.

References

[1] W. G. O. H. Programme, "cancer," 2015.

[2] J. J. C. Oncol, "Cancer Control in Bangladesh," 2013.

[3] P. M. Yamini Ranchod, "What you should know about mouth cancer," 2019.

[4] A. S. Croat, "Challenges in Early Diagnosis of Oral Cancer: Cases Series," 2019.

[5] J. P. B. Sci, "A Clinicopathological Study of Various Oral Cancer Diagnostic Techniques," 2017.

[6] T. A. C. S. m. a. e. c. team, "Risk Factors for Oral Cavity and Oropharyngeal Cancers," 2018.

[7] D. E. H. L. P. B. Schmidt BL, "Risk Factors," 2004.

[8] R. M. Kim Attanasi, "Screening for Oral Cancer," 2015.

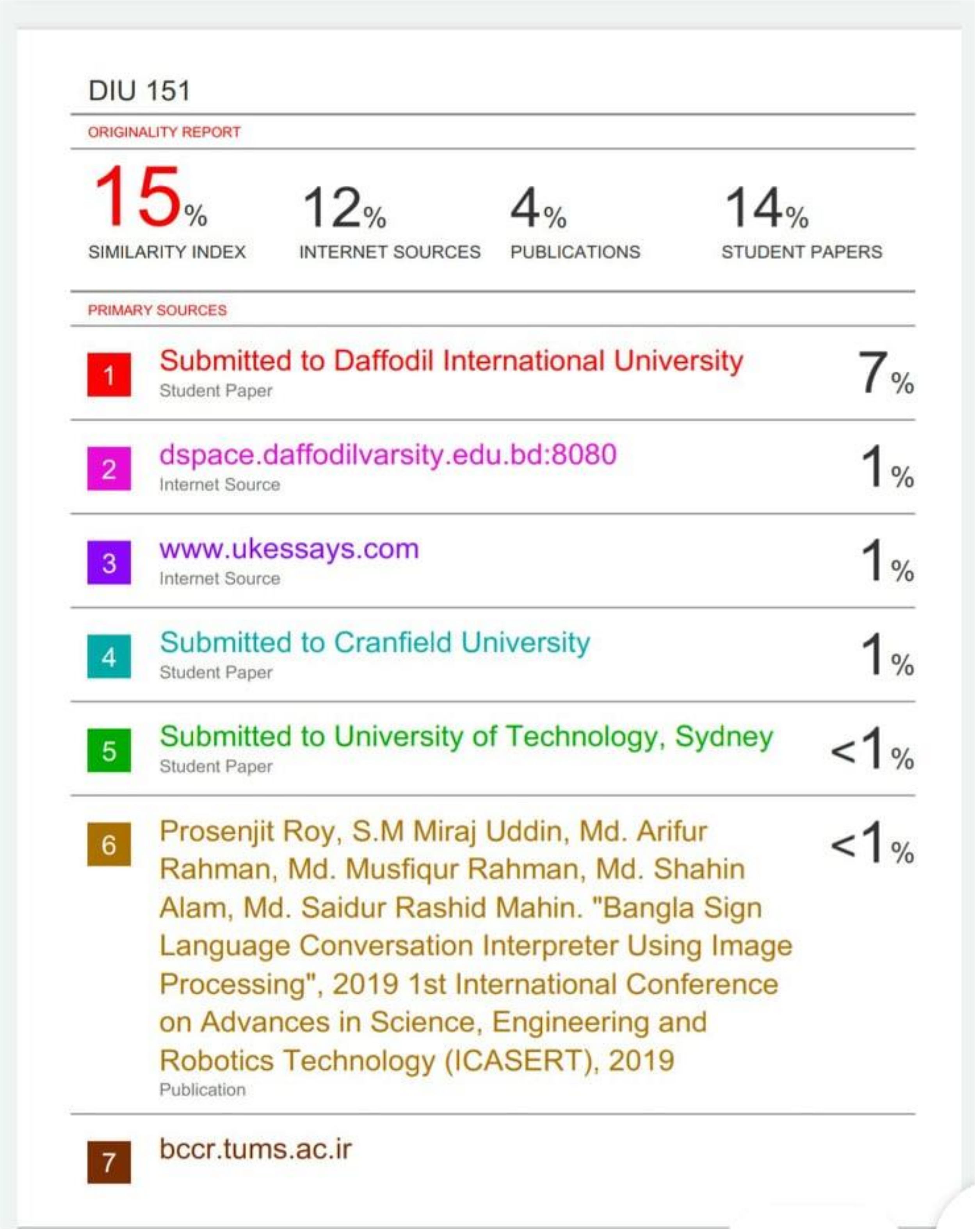
[9] N. Sharma, "Extracting Significant Patterns for Oral Cancer Detection Using Apriori Algorithm," 2014.

[10] B. H. Profile, "Leading Causes of Death Bangladesh," 2018.

[11] URMC, "Oral Cancer Overview".

[12] P. M. Yamini Ranchod, "Oral Cancers," 2020.

Plagiarism Report:



DIU 151

ORIGINALITY REPORT

15%	12%	4%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	7%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	www.ukessays.com Internet Source	1%
4	Submitted to Cranfield University Student Paper	1%
5	Submitted to University of Technology, Sydney Student Paper	<1%
6	Prosenjit Roy, S.M Miraj Uddin, Md. Arifur Rahman, Md. Musfiqur Rahman, Md. Shahin Alam, Md. Saidur Rashid Mahin. "Bangla Sign Language Conversation Interpreter Using Image Processing", 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), 2019 Publication	<1%
7	bccr.tums.ac.ir	