Kingdom of Saudi Arabia
Majmaah University
Ministry of Higher Education
College of Science Al Zulfi

المملكة العربية السعودية
جامعة المجمعة
وزارة التعليم العالي
كلية العلوم الزلفي

WEB USAGE MINING FOR RECURRENT BREAST CANCER

Student Affairs System
For College of science Al Zulfi
Department of Computer Science and Information

Submitted in partial fulfillment of the requirements for the award of
Bachelor degree of the Majmaah University
(Semester 1, 2018-19)

BY

STUDENT NAME: SARA FALEH AHMED AL SUGHAEIR

STUDENT ID:351205273

Supervised by:

MRS. NASREEN SULTANA QUADRI

DATE:13/3/1440

ABSTRACT:

Breast cancer is one of the most leading causes of deaths between women and even though it is rare, but still exists in men too. Medicine field have been improved the past years, but the death rate due to breast cancer is increasing day by day. The problem is that after initial treatment is done the breast cancer may come back again and that is called recurrent breast cancer. This paper aim is to determine the best classification algorithm to detect recurrent breast cancer based on its symptoms. This paper is using a Wisconsin data set and analysis it using WEKA tool by data mining techniques in specific it's concentrating on using classification algorithms such as Logistic model tree (LMT), Decision Tree Algorithm, Bayesian Network, Naive Bayesian Classification, K-Nearest Neighbor (KNN), Support Vector Machine, etc. it is expected to find the best classification algorithm to determine the tumors.

Key words: Breast cancer, Data Mining, Classification algorithms, UCI Wisconsin dataset, WEKA

Acknowledgments:

First I thank Allah for everything that made me capable of accomplishing this academic research project. Then to my parents and family who support and help me during this time, I am thankful to you. My supervisor for sharing expertise and helpful guidance. To my partners who helped me and cheers me. I am very thankful to teacher noura who guided my project. Finally, to the doctor who provided the important information I needed in this project. I am very thankful to all of you.

Special acknowledgment for Jiawei, Micheline & Jian the authors of *"Data Mining Concepts and Techniques*" book for their amazing and helpful book. A lot of information provided from this book have been very helpful for this project to complete. thank you

MAJMAAH UNIVERSITY,
COLLEGE OF SCIENCE AL ZULFI,
DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION

(CERTIFICATE BY STUDENT)

This is to certify that the project titled **"web usage mining for recurrent breast cancer"** submitted by me (**Sara Faleh Ahmed Alsughaeir**, **351205273**) under the supervision of **MRS. NASREEN SULTANA QUADRI** for award of Bachelor degree of the Majmaah University carried out during the Semester 1, 2018-19 embodies my original work.

Signature in full: ----------

Name in block letters: SARA FALEH AHMED ALSUGHAEIR

Student ID: 351205273

Date: 21/11/2018

Table of contents:

List of Tables:

List of figures:

List of symbols:

SMO	Sequential Minimal Optimization
KNN	K-Nearest Neighbor
BF Tree	Best first tree
FF	Farthest First
HCM	Hierarchical Cluster Method
EM	Expectation Maximization
SVM	Support Vector Machine
LMT	Logistic model tree
IBK	Instance based classifier
LBR	lazy Bayesian rules

List of abbreviations:

Recurrent breast cancer	Breast cancer tumor that comes back after initial treatment.
Data mining	The process of discovering patterns from huge amounts of data.
Data Analysis	The process of applying statistical and/or logical techniques to describe and illustrate, summary, and evaluate data.
Weka	It is a software that consist of collection of machine learning algorithms for data mining tasks.

Chapter 1: Introduction

1.1 Overview :

The amount of data over years has been in continuous growing simultaneously with the increasing in use of electronic devices, and data in all electronic format (text, image, audio, videos, …) have been used in many ways to extract important information from it, that helped to increase the knowledge in many industries. From this point Data mining concept comes.

> "We are living in the information age" is a popular saying; however, we are actually living in the data age. Terabytes or petabytes of data pour into our computer networks, the World Wide Web (WWW), and various data storage devices every day from business, society, science and engineering, medicine, and almost every other aspect of daily life. This explosive growth of available data volume is a result of the computerization of our society and the fast development of powerful data collection and storage tools. Businesses worldwide generate gigantic data sets, including sales transactions, stock trading records, product descriptions, sales promotions, company profiles and performance, and customer feedback. For example, large stores, such as Wal-Mart, handle hundreds of millions of transactions per week at thousands of branches around the world. Scientific and engineering practices generate high orders of petabytes of data in a continuous manner, from remote sensing, process measuring, scientific experiments, system performance, engineering observations, and environment surveillance. Global backbone telecommunication networks carry tens of petabytes of data traffic every day. The medical and health industry generates tremendous amounts of data from medical records, patient monitoring, and medical imaging. Billions of Web searches supported by search engines process tens of petabytes of data daily. Communities and social media have become increasingly important data sources, producing digital pictures and videos, blogs, Web communities, and various kinds of social networks. The list of sources that generate huge amounts

> of data is endless. This explosively growing, widely available, and gigantic body of data makes our time truly the data age. Powerful and versatile tools are badly needed to automatically uncover valuable information from the tremendous amounts of data and to transform such data into organized knowledge. This necessity has led to the birth of data mining. (Jiawei, Micheline and Jian ,2012, p 1-2)

We can define data mining as follows **"**Data mining is the process of discovering interesting patterns and knowledge from *large* amounts of data" (Jiawei, Micheline and Jian ,2012, p 8). Data mining uses different techniques for that.

In this project we are going to concentrate on recurrent breast cancer. Breast cancer is one of the most leading causes of deaths among women. Although death rates have been decreasing over the past years because of the improvement in the medicine field, there is still numbers of death because of breast cancer in world. In order to help saving lives the world should continue to find different ways to predict and diagnose of breast cancer, as early detection improves the chances for survival as well as the tumor may appear again and then the chances for lives decrease. So, here we're going to introduce a data mining model that compares different classification algorithms by using web data to identify the key factors that can help to diagnosis tumors.

1.2 Problem definition:

The computer science developing helped with solving many problems in different industries, from these industries is the medicine and health industry. One of the problems that scare people around the world is breast cancer. Recurrent breast cancer is a danger disease that causing death around the world specially women even with the different treatment methods developing over years. so we are going to try to find a way that may help with this problem.

Recurrent breast cancer is the tumors that comes back after initial treatment. Although the initial treatment is aimed at eliminating all cancer cells, a few may have evaded treatment and survived. These undetected cancer cells multiply, becoming recurrent breast cancer. Recurrent breast cancer may occur months or years after initial treatment. The cancer may come back in the same place as the original cancer (local recurrence), or it may spread to other areas of the body (distant recurrence). Treatment may eliminate local or distant

recurrent breast cancer. Even if cure is hard or hopeless, treatment may control the disease for long periods of time. (Mayo clinic, 2014, para 1)

the technique we are going to use to detect the existence of recurrent breast cancer is classification.

Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical class labels. For example, a medical researcher wants to analyze breast cancer data to predict which one of three specific treatments a patient should receive. the data analysis task is classification, where a model or classifier is constructed to predict class (categorical) labels, such as "treatment A" "treatment B" or "treatment C" for the medical data. A general approach to classification as a two-step process. In the first step, we build a classification model based on previous data. In the second step, we determine if the model's accuracy is acceptable, and if so, we use the model to classify new data. (Jiawei, Micheline and Jian, 2012)

The techniques used for classification of data that are commonly used in the field of data mining: (Deeba and Amutha, 2016)

1. Decision Tree Algorithm

   Decision tree algorithm create a tree structure from given data set where each node represents attributes test or conditions and final leaf node represents the test results or classes. The construction of decision tree is done by divide and conquer strategy.

2. Bayesian Network

   Bayesian network is a kind of probabilistic method of rule generation. This method will derive a directed acyclic graph that describes the dependency relationship among the variables. Directed acyclic graph consists of set of nodes that represent the random variables and edges connecting these nodes represent the probabilistic dependency between the variables

3. Naive Bayesian Classification

   The naive Bayesian is a type of probabilistic classifier. This method uses Bayes' theorem and also assumes that each and every features in a class are highly independent, that is the appearance of a feature in a particular category is not connected with to the presence of any other feature.

4. K-Nearest Neighbor (KNN)

   KNN is a classification method works is simple. The training sample consists of set of tuples and class labels. This algorithm works for random number of classes. KNN uses distance function to map the samples with classes.

Classification process of KNN will find the distance between the given test instance X with that of existing samples Y1, Y2, ..., YK. The nearest neighbors are found and based on the voting of neighbors, the common neighborhood class is assigned to the test samples.

5. Support Vector Machine (SVM)

   SVM is a classification model based on supervised training method for binary classification. Here the training samples belong to any one of two classes. Based on the training data samples, SVM builds a prediction model that will classify the new sample properly to any one of the two classes.

1.2.1 Goals

This project is aiming to:

1. identify a data mining model that accurately predicts the presence of breast cancer
2. using web data to find different issues in recurrent cancer.
3. Comparing and analyzing different data mining algorithms.

1.2.2 Objectives:

1. get a database of information related to breast cancer from Patients / doctors or web data (online data)
2. predicting the presence of breast cancer malignant/benign
3. Finding the efficient algorithm for predicting presence of tumor

1.2.3 Critical success factors:

1. Accurate data set.
2. Different classification algorithm.

1.3 Feasibility study:

The feasibility study is an analysis of the ability to complete a project successfully. A feasibility study allows project managers to investigate the possible negative and

positive outcomes of a project before investing too much time and money (Fincher, Sally, Marian Petre, & Martyn Clark 2001).

In the survey we did to use in order to build our project on (see appendix 1), we have got 108 responds. There was a question is "have you heard About Recurrent Breast Cancer" and there are 22 respond out of 108 said no, the point is that even though breast cancer is widely known there are still people who does not know anything about. These past years the computer science developed in many industries and it makes our life simpler in many ways so I believe that this project might be useful to people to get to know about recurrent breast cancer and diagnosis process be simpler.

Chapter 2: Literature review

2.1 Introduction:

In this chapter we are going to talk about statistics on breast cancer in Saudi Arabia and review some articles in the same field as our paper.

2.2 Statistics on breast cancer in Arab countries and the world:

Breast cancer is the most Pervasively cancer in the world in women, with 22% of all cases. In 2000, the number of cases was estimated at 1,050,346 cases in advanced countries with an average of 55.2%.

In Saudi Arabia, the number of new cases of cancer was 2741 cases, 19.9% of cancer in women was breast cancer taking the first place. the difference between the Arab countries, including Saudi Arabia and the United States is in terms of age and disease when discovered. In the United States, 50% of new breast cancer cases occur in women over 65 years old, while in Arab countries including Saudi Arabia at age 52. In terms of disease, in advanced countries, the disease is detected at earlier stage, while other countries large numbers are still diagnosed at later stages.

(Ministry of health,1436)

2.3 Literature review:

For our paper I am going to concentrate on recurrent breast cancer we can summary the recurrent breast cancer by the cancer that comes back after initial treatment. the initial treatment is aimed at removing all cancer cells, but a few of cancer cells may didn't get the treatment and survived. These undetected cancer cells multiply becoming recurrent breast cancer. We will start getting the data using web by google forms and publish it in our university and try to spread it to other places using web and social media. Then we are going to use WEKA tool to compare all classification algorithm we can to get the best algorithm.

the authors focus on taking the breast cancer with its two basic type that if it is benign then it is not cancer or malignant to be cancer, then they list the risk factors for example Gender, Age, Genetic risk factors, Family history…etc. Then after that it start comparing the performance of different classification techniques. They get The data for the paper for breast cancer from the Wisconsin dataset from UCI machine learning data with a total 683 rows and

10 columns. They put their aim to be developing accurate prediction models for breast cancer using data mining techniques. They compared three classification techniques in Weka tool software the results were that SMO has the higher prediction accuracy for 96.2% than KNN and BF Tree method (Vikas and Saurabh, 2014)

I think that it well be more efficient to compare more than three algorithms since it may give totally different results.

the authors concentrate on early diagnosis of the breast cancer. They used clustering data mining algorithm to detect breast cancer. The dataset is used form the UCI web data repository. They toke four clustering data mining techniques k-means clustering algorithm, FF algorithm, HCM algorithm and EM they think that This research would become very helpful to doctors and patients for early diagnosis of breast cancer. The authors took attribute based on tumors itself aside from age they took menopause (the period in a woman's life when menstruation ceases), Tumor-size, Deg-malig indicate the Stage of breast cancer…etc. The final results show that k-means and FF farthest first clustering algorithms are helpful to early diagnosis of the breast cancer. But HCM hierarchical cluster method algorithm has high error rate. And last EM expectation maximization technique cannot diagnosis 36% patients. (Jahanvi, Rinal and Jigar 2014)

This paper used clustering algorithms to early diagnosis breast cancer but on our paper I am going to use classification techniques because I believe It would give more accurate results.

The author focuses on Bayesian Classification to predict and detect anomalies in breast cancer for determining the higher prediction accuracy. They used the Naïve Bayes Algorithm to develop this methodology. They took attribute as Clump Thickness, Uniformity of cell size, Uniformity of cell shape…etc. the dataset is from Wisconsin breast cancer from UCI. (Souad Demigha, 2016)

This paper above used Bayesian algorithm to diagnosis breast cancer and the author explain it in details but I think comparing more than algorithms of classification would give us a better result because we might find higher prediction algorithm for recurrent breast cancer.

Ahmed and Ayman presented an automatic diagnosis system for detecting breast cancer based on unsupervised pre-training phase followed by a supervised back propagation neural network phase (DBN-NN). That has achieved higher classification accuracy in comparison to a classifier with just one supervised phase. The enhancement of overall neural network accuracy is reaching 99.68% with 100% sensitivity and 99.47% specificity breast cancer case. (Ahmed and Ayman ,2016)

Chapter 3: System Analysis

3.1 Introduction:

Systems analysis describes in detail the "what" that a system must do to satisfy the need or to solve the problem that may face the system. It's providing the tools and techniques to the developer, so the developer can understand the need (business need), capture the vision, define a solution, communicate the vision and the solution. we can conclude that Systems analysis is consisting of those activities that enable a person to understand and specify what the system should accomplish. (John, Robert and Stephen, 2012.)

3.2 Description of Data Flow Diagram (DFD):

Somerville (2011) said Data-flow diagrams (DFDs) are system models that show a functional perspective where each transformation represents a single function or process. DFDs are used to show how data flows through a sequence of processing steps. For example, a processing step could be the filtering of identical records in a customer database. The data is transformed at each step before moving on to the next stage. These processing steps or transformations represent software processes or functions where data-flow diagrams are used to document a software design.

3.2.1 Context Diagram:

1. A context data flow diagram (DFD), also known as a level 0 DFD, gives a broad overview of an information system and the way it interacts with external entities. (Lucid chart Inc, n.d.)

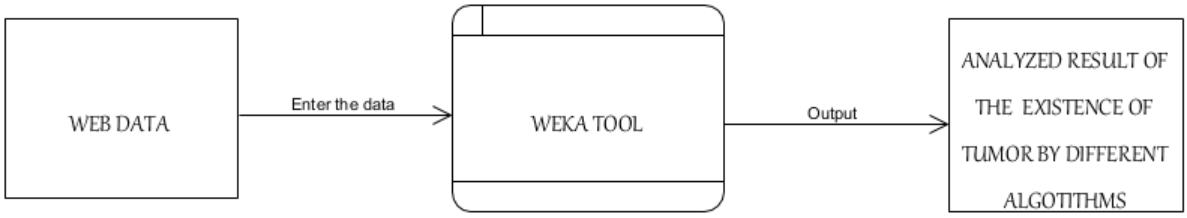


Figure 1:context diagram

We have as our external entity the web data that we got using web that data will be entered to the WEKA tool, we will use WEKA tool to analyze the data using different algorithms to give us the result of the existence of tumors.

3.3 Use Case Diagram:

use case diagram can summarize the details of your system's users (also known as actors) and their interactions with the system. To build one, you'll use a set of specialized symbols and connectors. (Lucid chart Inc, n.d.)

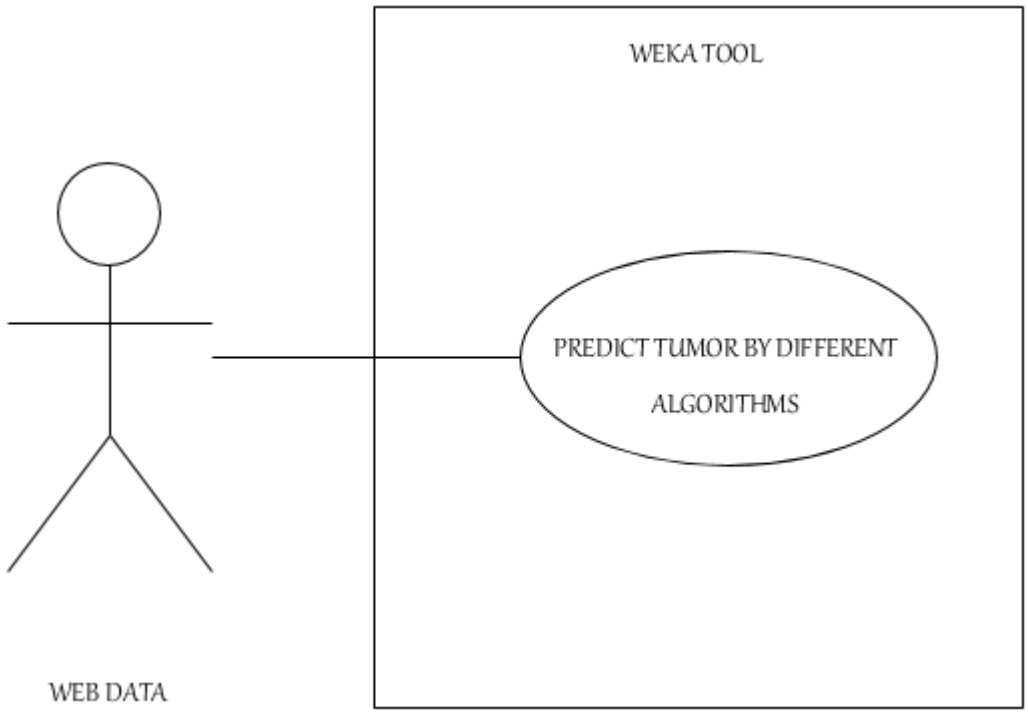


Figure 2:use case

The above use case diagram illustrates that a web data will be entered to the WEKA tool to give us the result of predicting tumor using different algorithm.

3.4 Sequence Diagram:

Sequence diagram is a type of interaction diagram because it describes how and in what order a group of objects works together. These diagrams are used by software developers and business professionals to understand requirements for a new system or to document an existing process. Sequence diagrams are sometimes known as event diagrams or event scenarios. (Lucid chart Inc, n.d.)

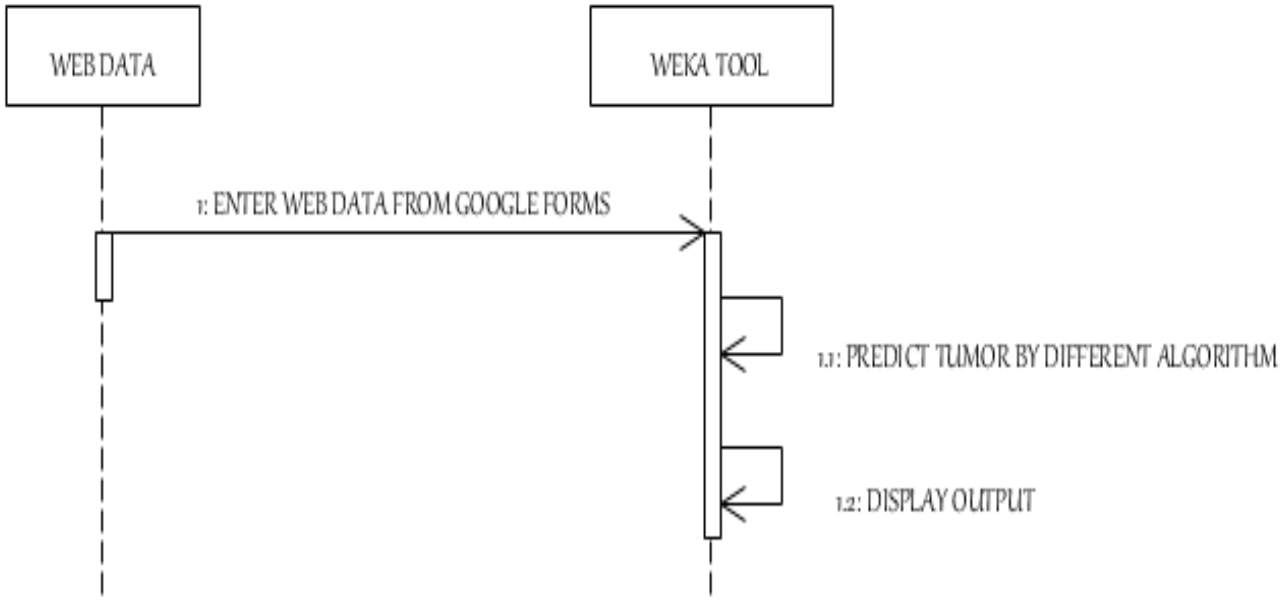


Figure 3:sequence diagram

First enter the data that has been collected through web that in our case we collected it by google forms in WEKA tool, then the results will be displayed after analyzing data to give us the predict of existence of tumors.

Chapter 4: System Design

4.1 Introduction:

Systems design consists of those activities that enable a person to describe in detail the system that solves the need. systems design describes how the system will work. It specifies all the components of the solution system and how they work together to provide the desired solution. (John, Robert and Stephen, 2012.)

4.2 Description of procedures and function:

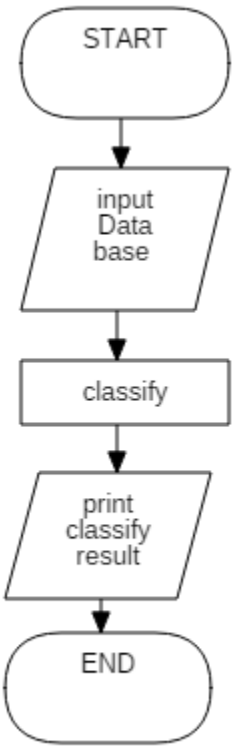


Figure 4 Flow Chart for clssification model

1. first we have to enter the data into WEKA tool
2. second it will classify the data
3. finally, the result will be printed

4.3 Hardware and software Requirements

4.3.1 Hardware Requirements:

We only need a computer device.

4.3.2 Software Requirements:

To this project we need one program that is a WEKA tool, Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. (The University of Waikato, n.d., para.1)

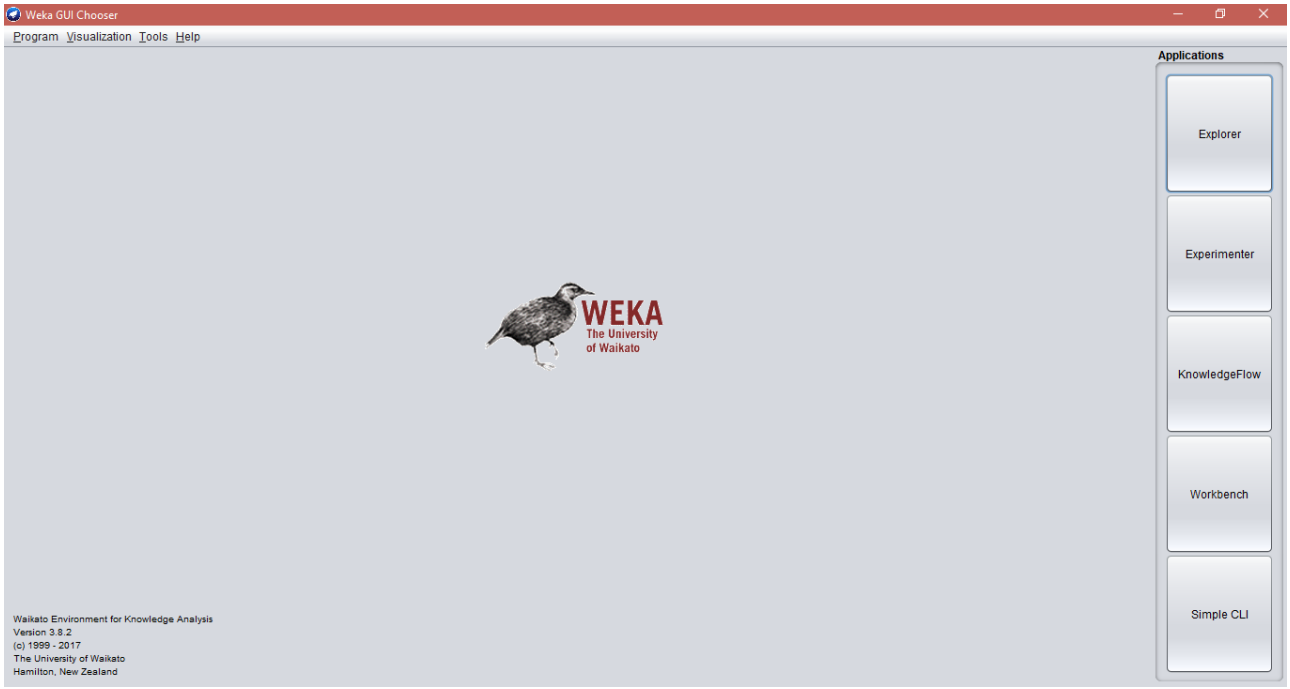


Figure 5 WEKA tool main page

Retrieved from: WEKA tool.

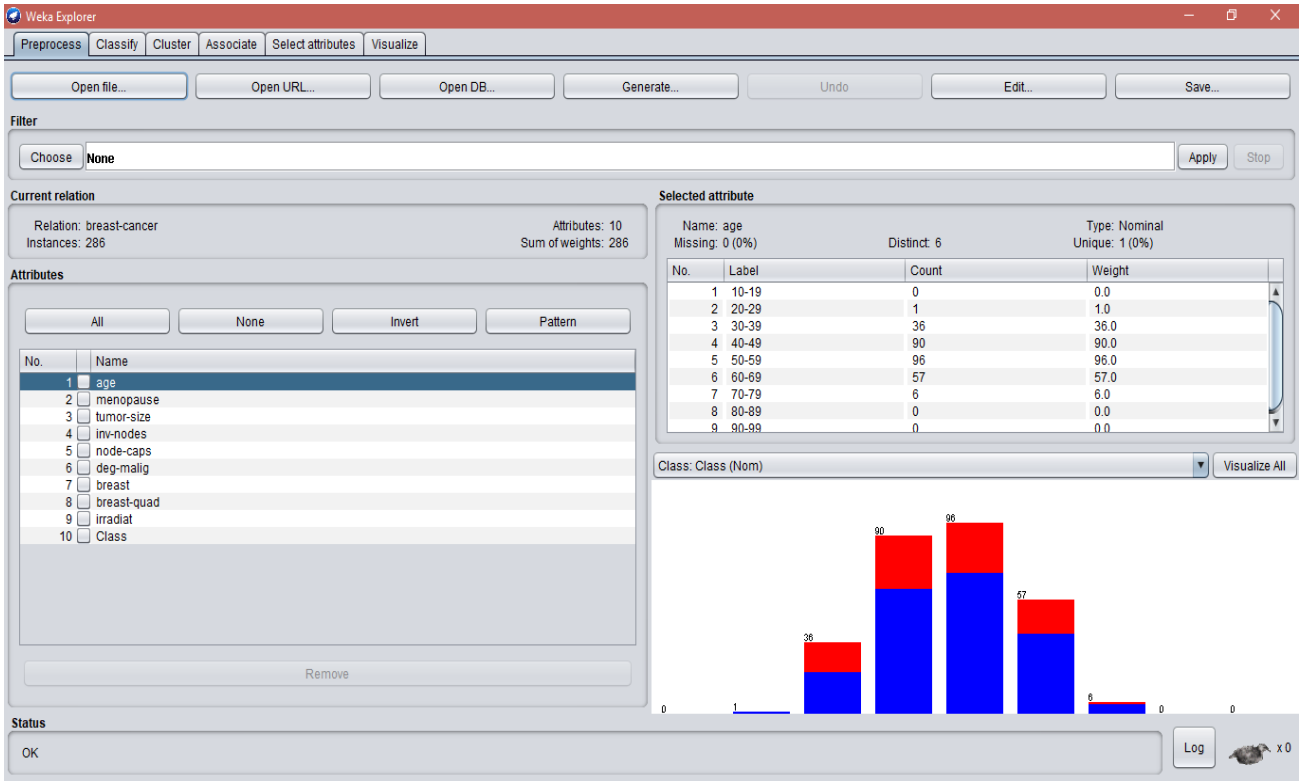


Figure 6 Sample on WEKA using

Retrieved from: WEKA tool.

In this figure (see figure 6) it shows how WEKA tool works, we used a sample dataset already exists in the program files, and as it shows (see figure 6) it gives back in detail all information regarding the dataset. Then you can apply the algorithms you want that the program provides.

Chapter 5: Implementation

5.1 Experimental Methodology

Since this project is depending on the analysis of real world data we are using Experimental approach for analyzing data. Experiment shows that the experiments extract results from real world implementations.

We have Three main steps:

1. Collect data
2. Preprocessing
3. Classification

5.1.1 Collect Data

From the survey we did before in graduation project 1 we have got in total 108 responds with totally 17 variables and with that we build a database (see appendix 1).

The variables are represented as questions as follows:

1- Are you a doctor?

2-gender?

3-have you heard About Recurrent Breast Cancer?

4-have you ever felt a mass (lump) in your breasts?

5-Is the mass hard?

6-Is the mass immovable?

7-Is it single mass?

8-Is the mass painful or tender?

9-what is the size of the mass?

10-Do you recently have continuous back and or leg pain?

11-Do you recently have continued abdominal pain, nausea or vomiting?

12-Did you notice any yellowish discoloration of your eyes or skin?

13-Do you recently have shortness of breath or cough?

14-Did you notice any redness over your breast skin that does not disappear?

15-Did you notice any skin changes of your breast for example thickening or dimpling of the skin?

16-Do you have bloody discharge from your nipples?

17-which treatment is best for breast cancer?

5 out of 108 are female doctors, the first question is to see how doctors will respond to this survey, for the last question one female doctor answer that Radiation therapy is best but two other female doctors answered with hormone therapy, the fourth prefer surgery and the last one answered with chemotherapy. In total for last question 44 answered surgery, 24 for chemotherapy,15 targeted therapy drugs, 14 radiations and 11 for hormone therapy. This make surgery the most prefer treatment in Saudi Arabia. We analyze that Mass is the important variable in finding the presence of tumor and if mass is present the size of mass decide the stage of breast cancer. other variable depends on the mass if mass is present then patient might felt other symptoms. Totally 22 answered with yes that they feel mass in their breast and between 6 to 16 answered the following question with yes. The questions after mass question define the presence of breast cancer they represent symptoms. Symptoms are categorized as general symptoms as is the mass hard, immovable, single, painful or tender, continuous back and or leg pain, continued abdominal pain, nausea or vomiting, yellowish discoloration of eyes or skin, shortness of breath or cough, redness over breast skin that does not disappear, skin changes of breast for example thickening or dimpling of the skin and bloody discharge from your nipples.

Figures below shows the responds to the survey

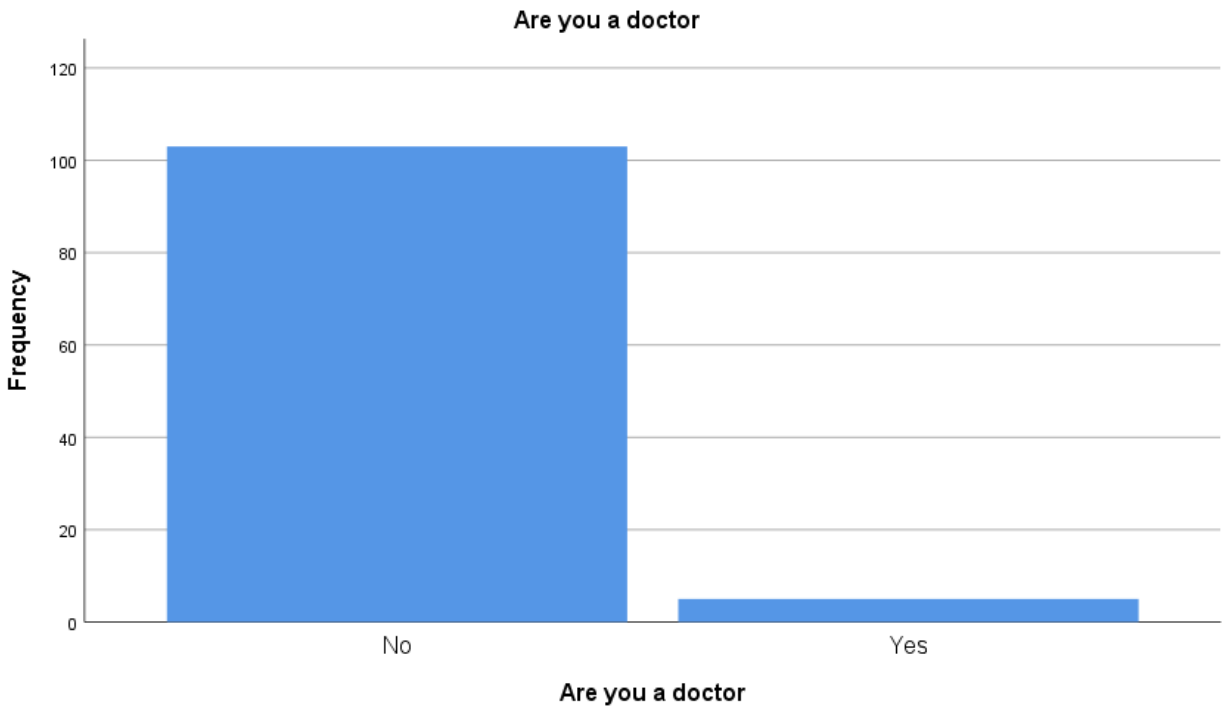


Figure 7 5.1.1.1

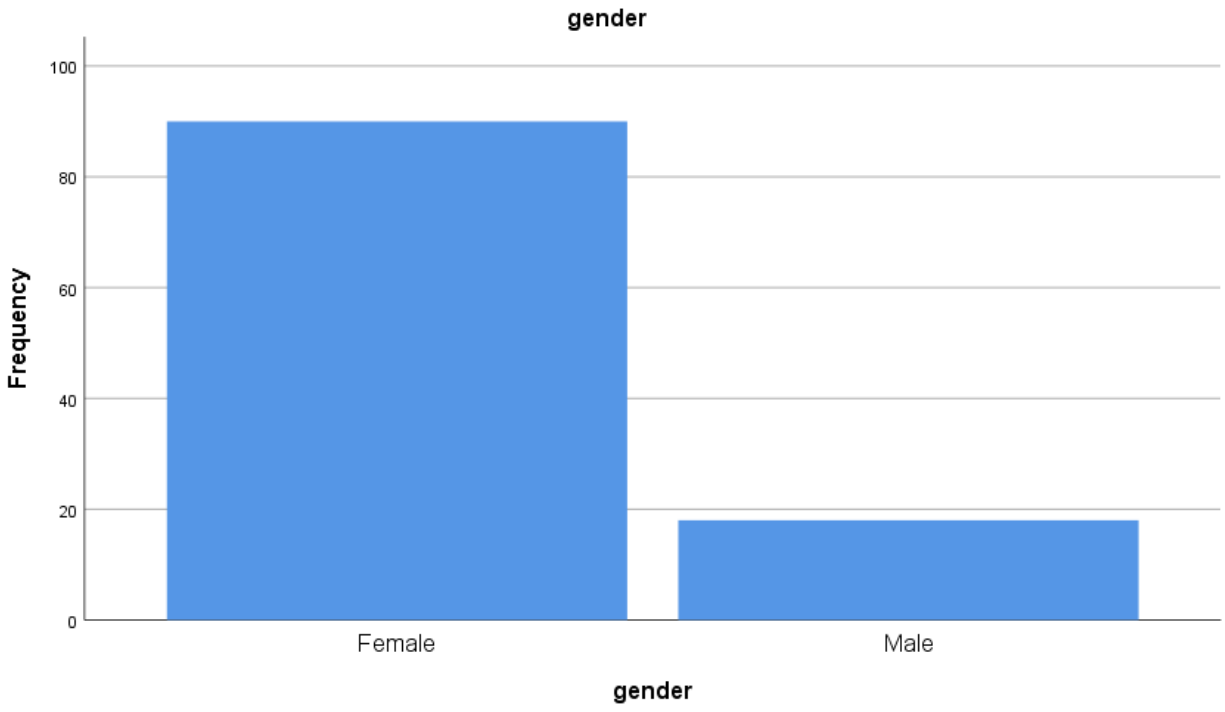


Figure 8 5.1.1.2

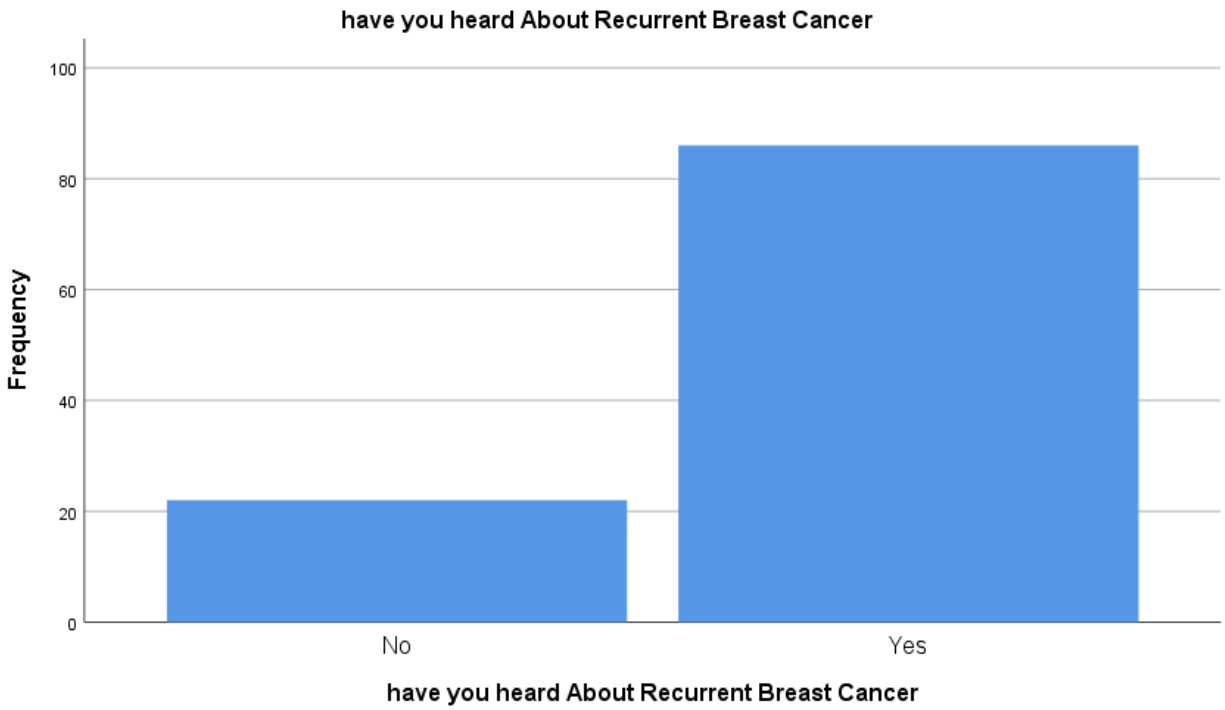


Figure 9 5.1.1.3

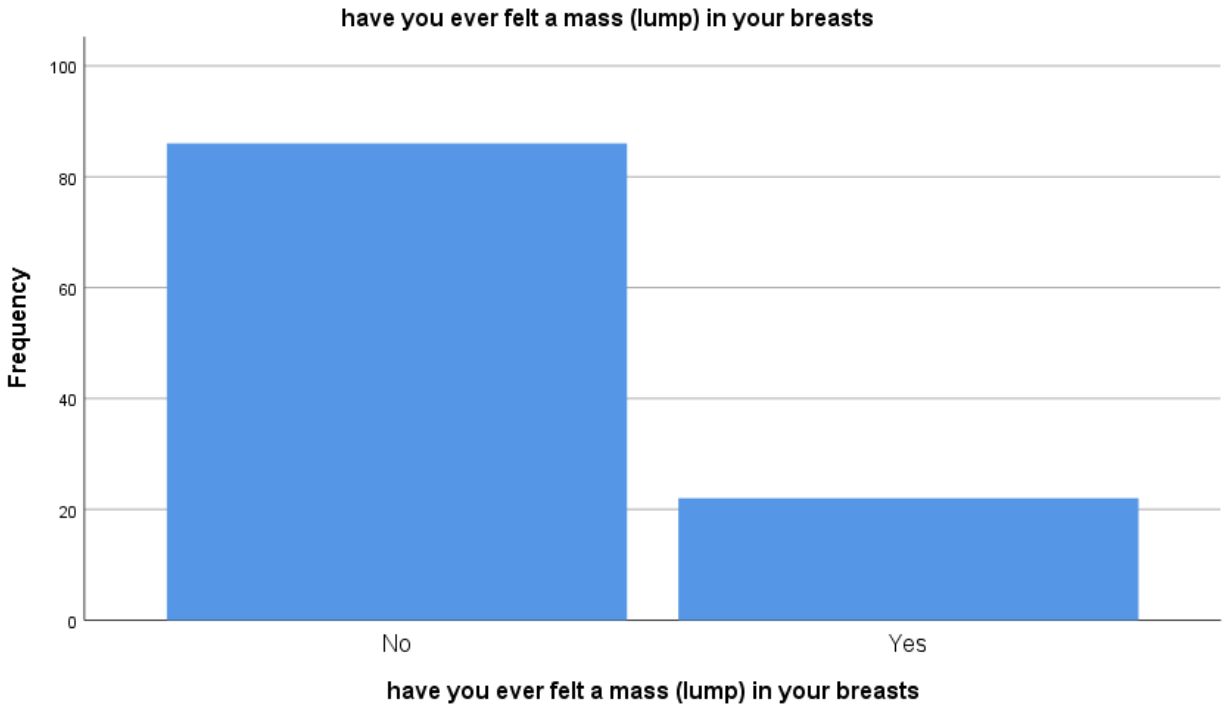


Figure 10 5.1.1.4

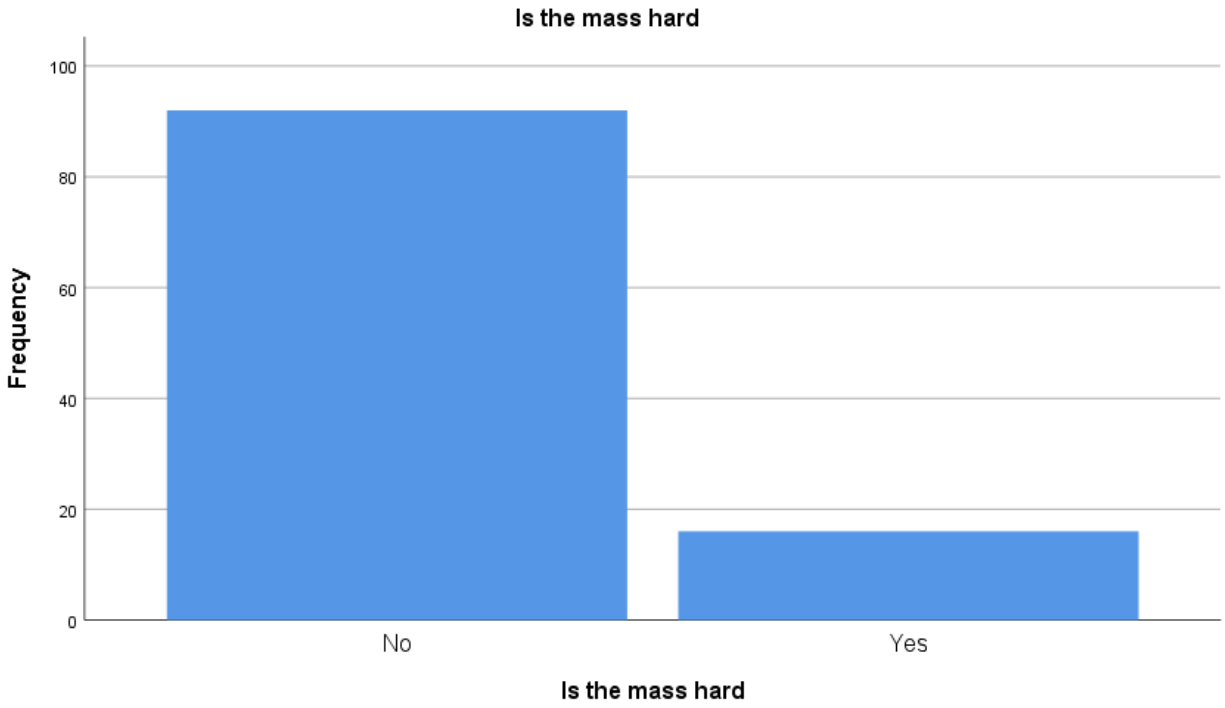


Figure 11 5.1.1.5

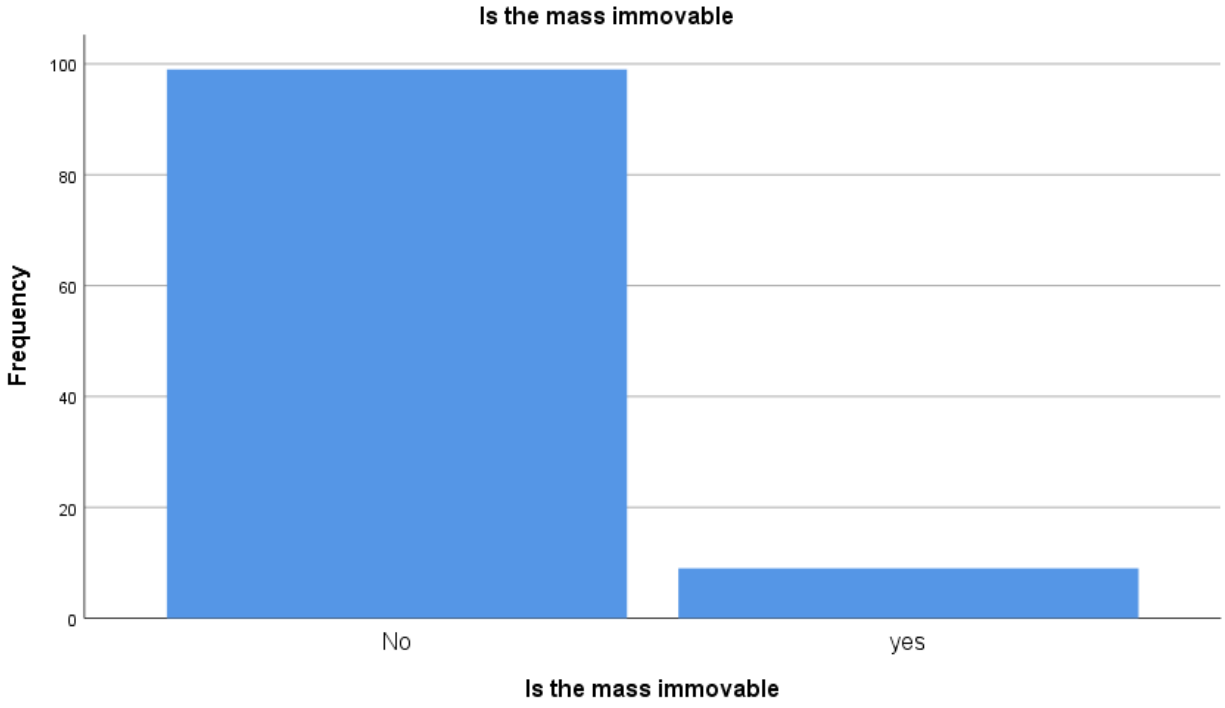


Figure 12 5.1.1.6

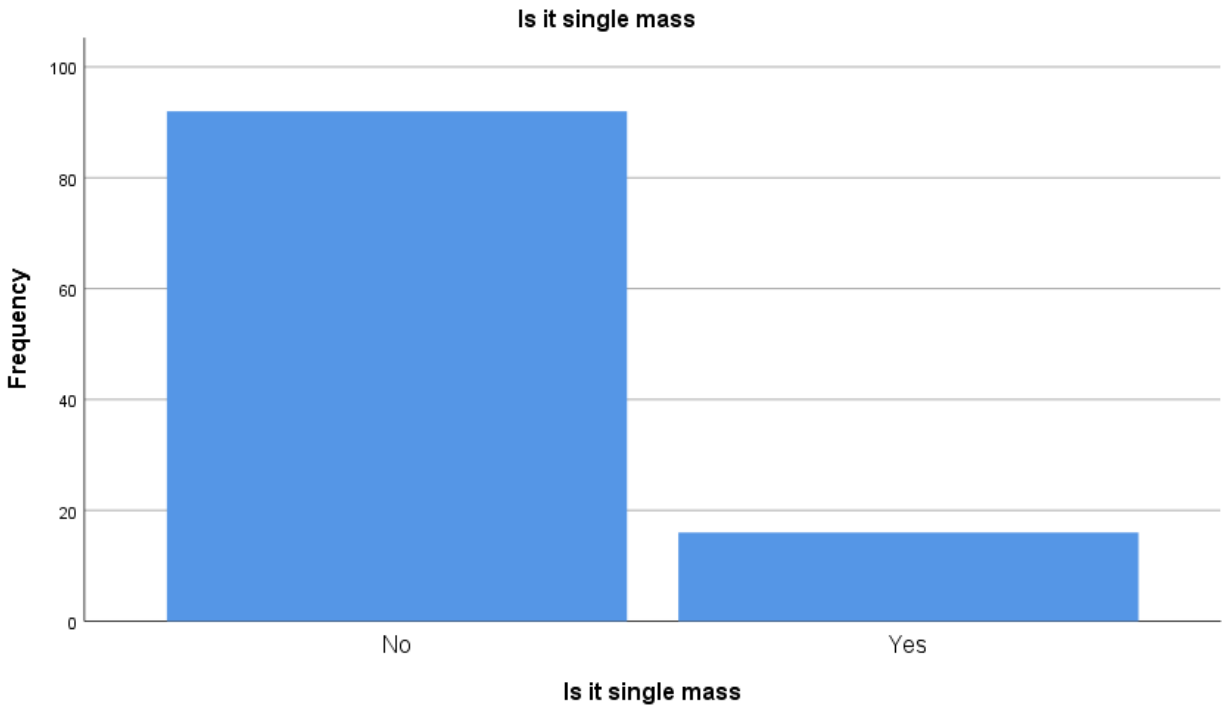


Figure 13 5.1.1.7

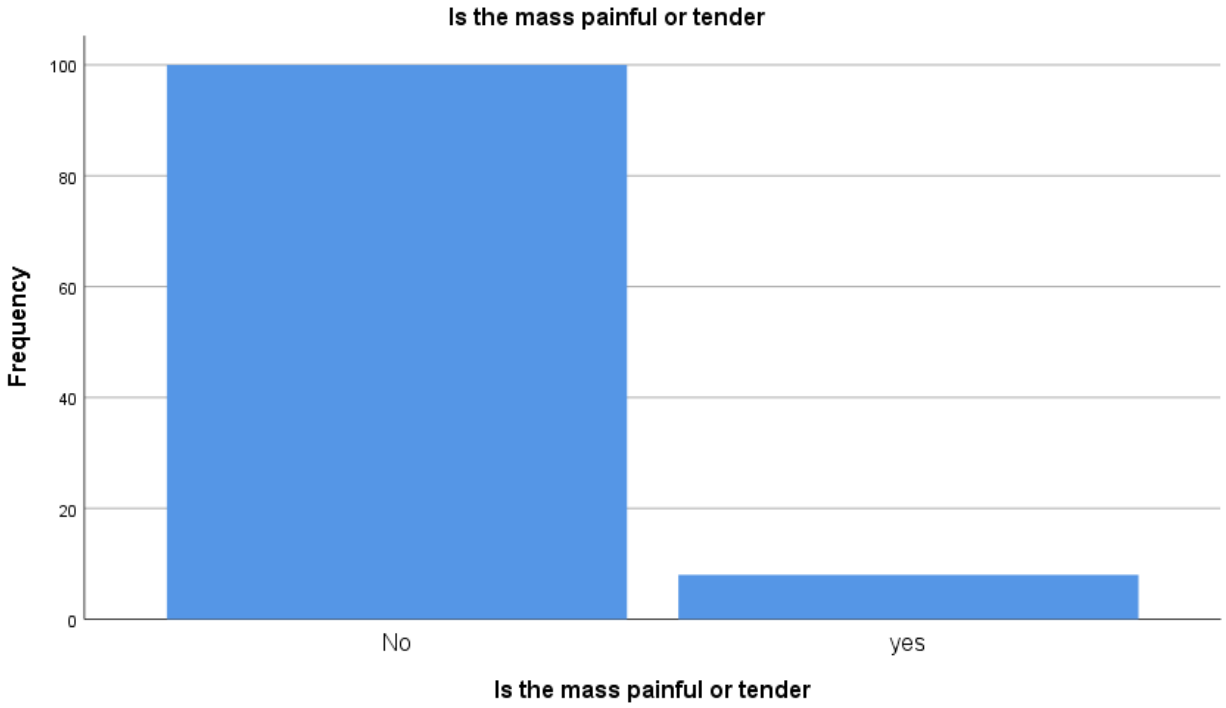


Figure 14 5.1.1.8

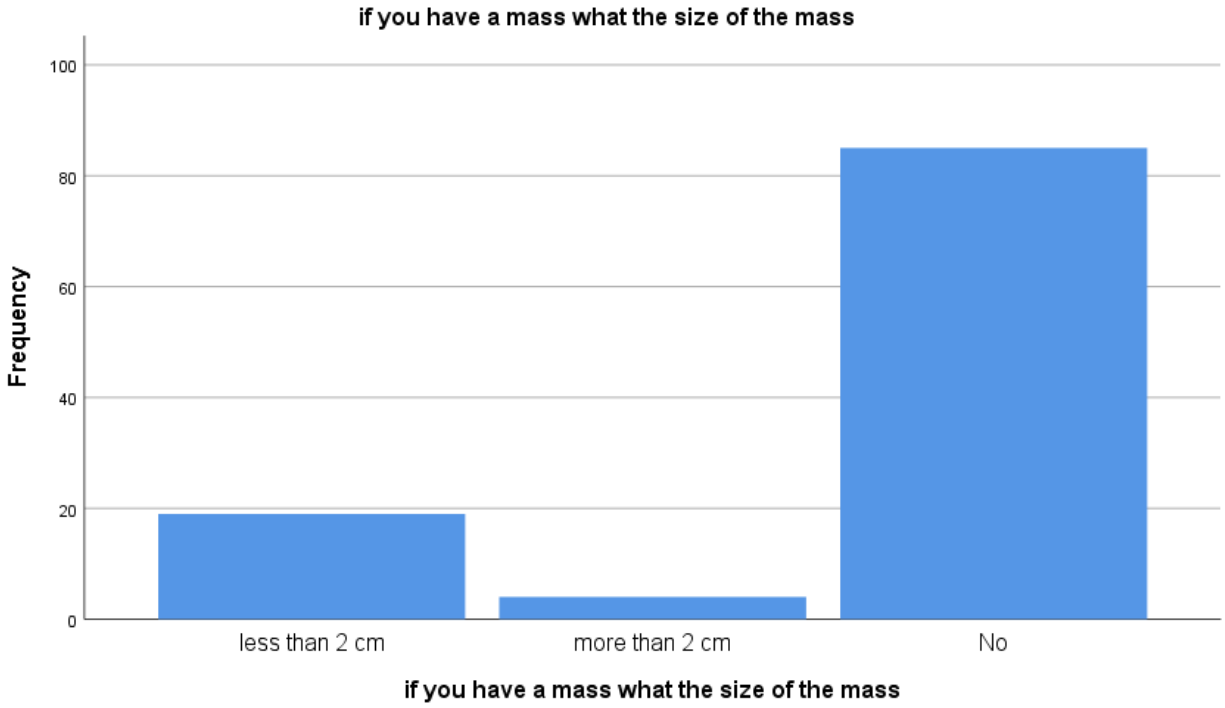


Figure 15 5.1.1.9

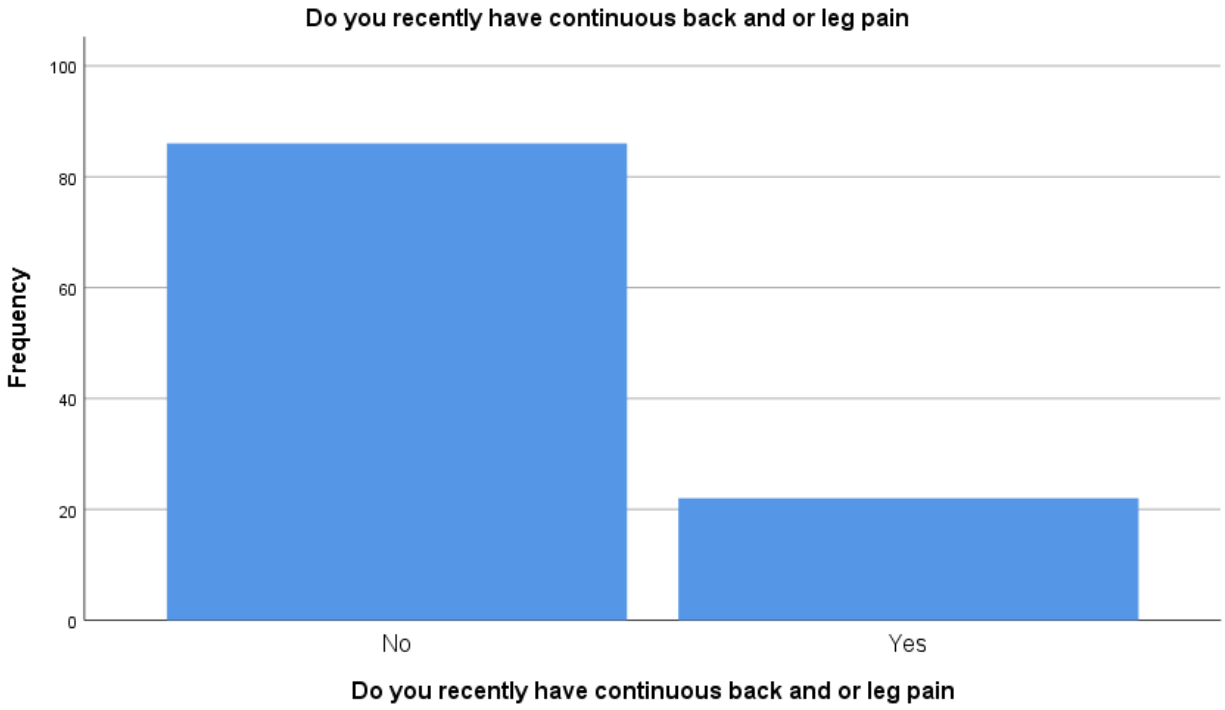


Figure 16 5.1.1.10

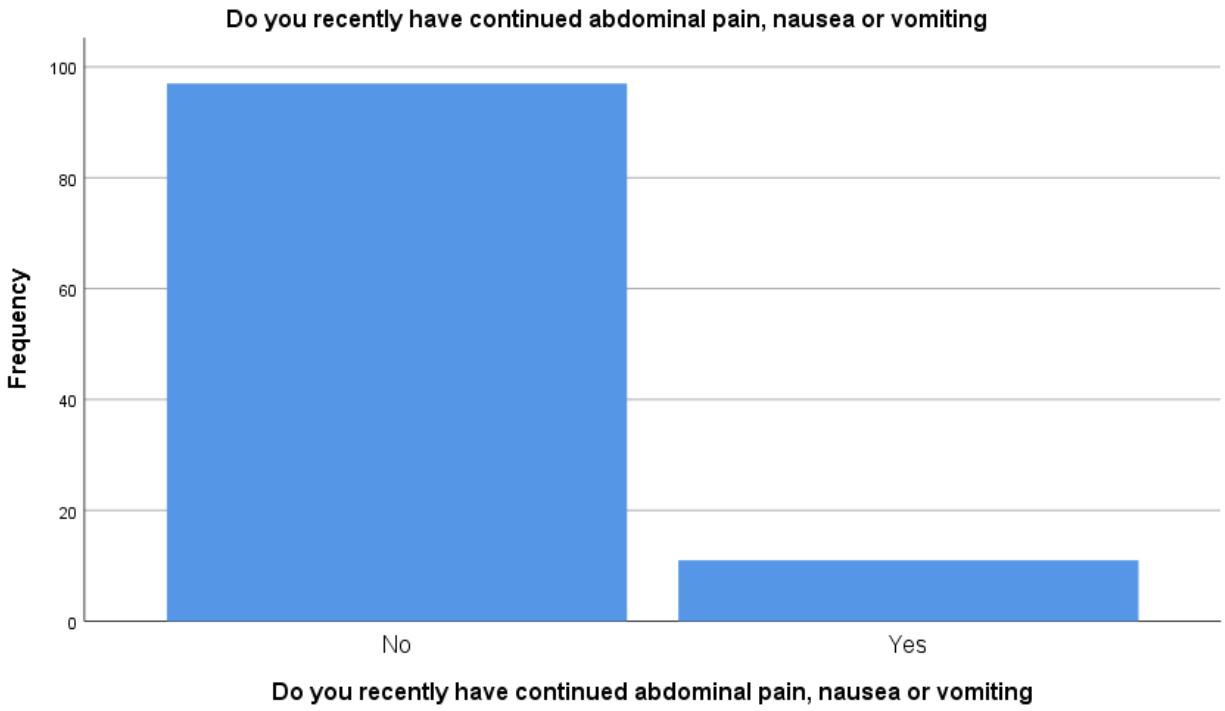


Figure 17 5.1.1.11

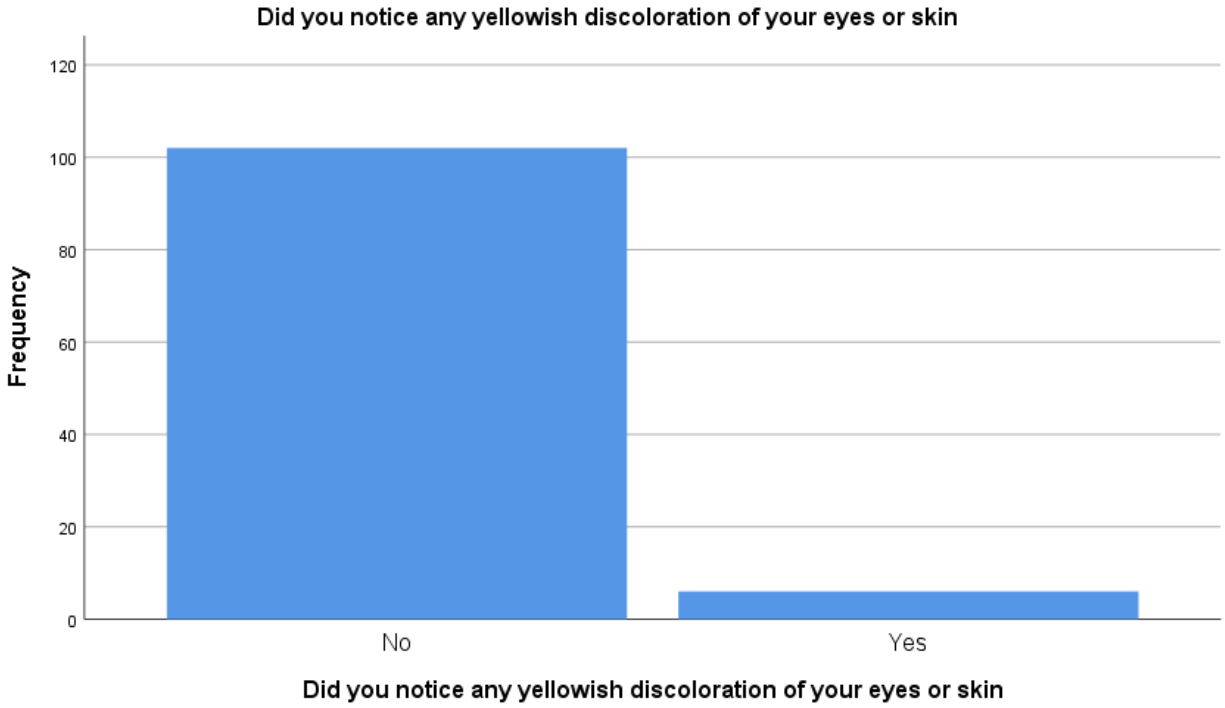


Figure 18 5.1.1.12

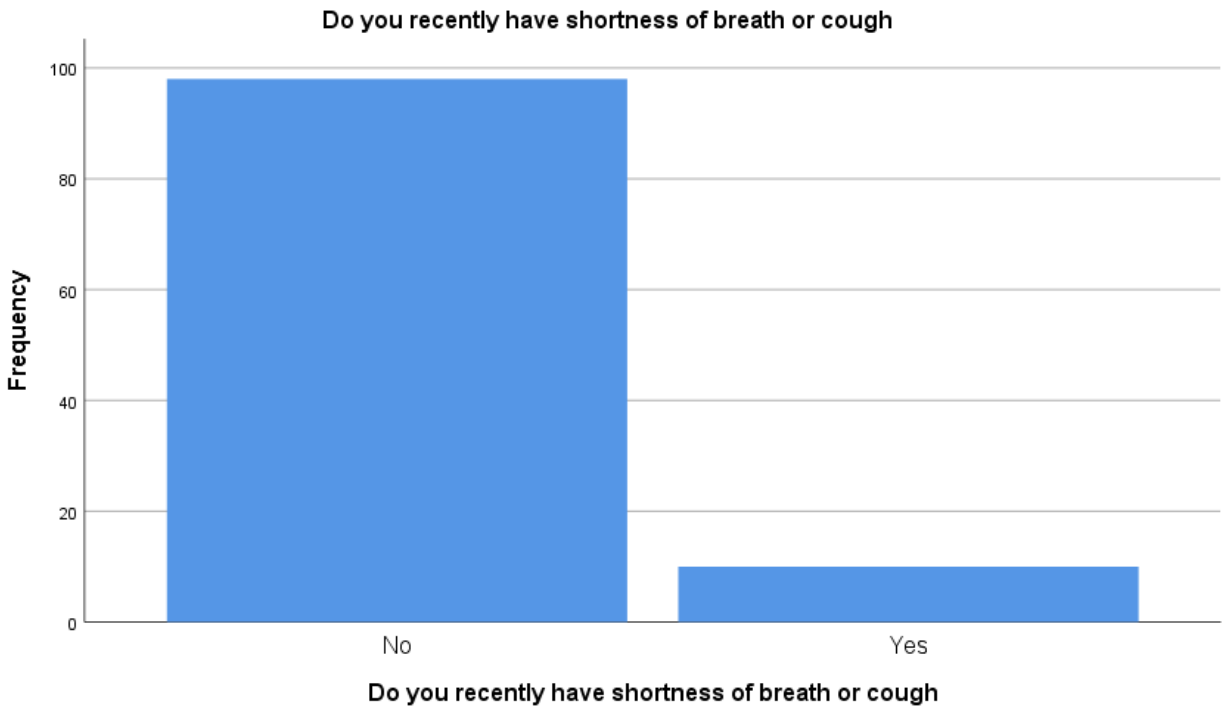


Figure 19 5.1.1.13

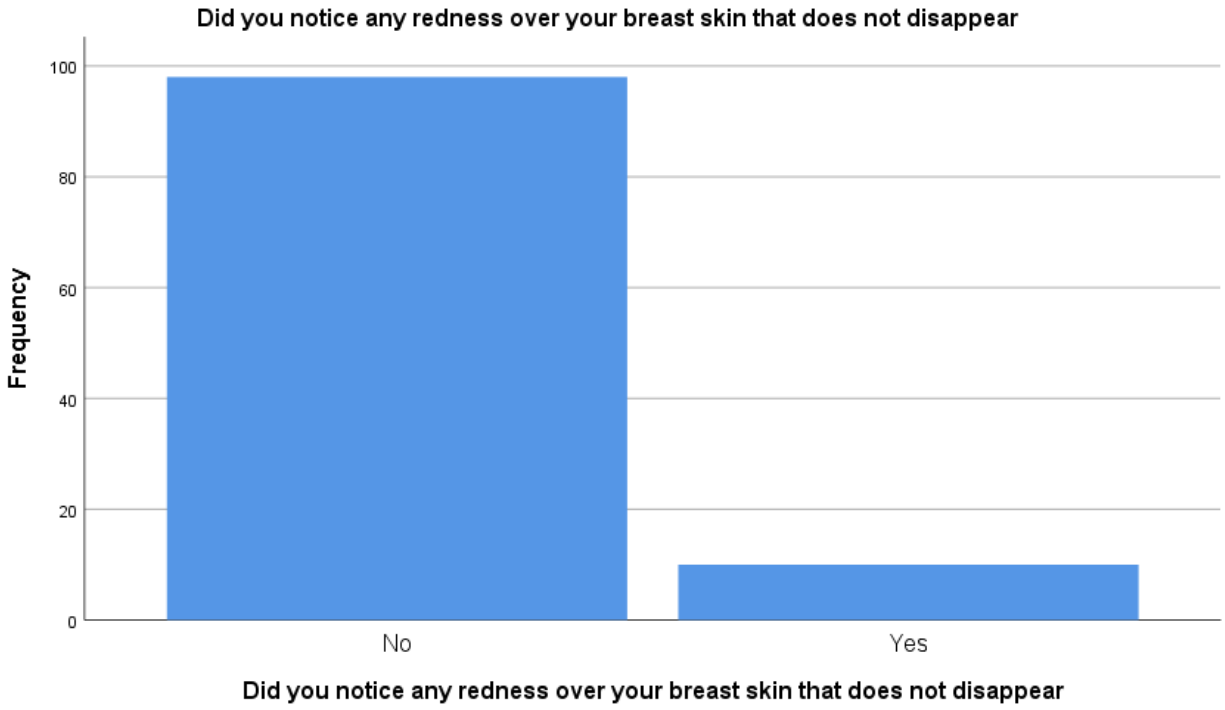


Figure 20 5.1.1.14

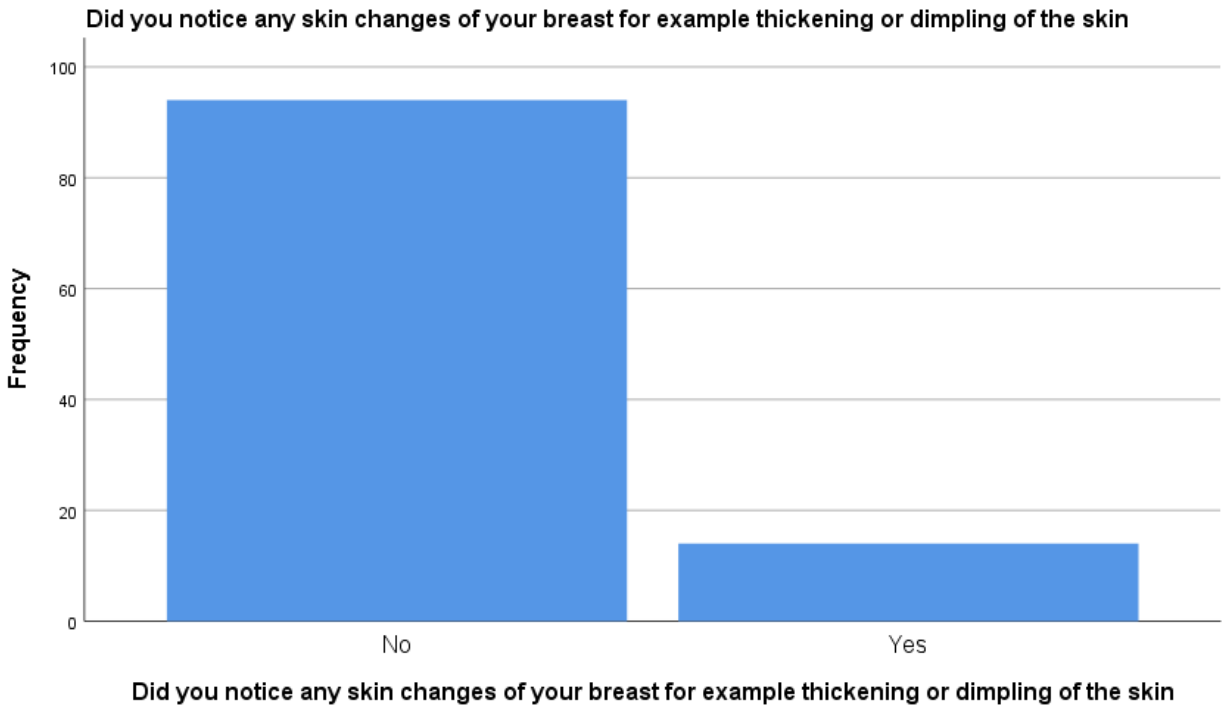


Figure 21 5.1.1.15

Figure 22 5.1.1.16

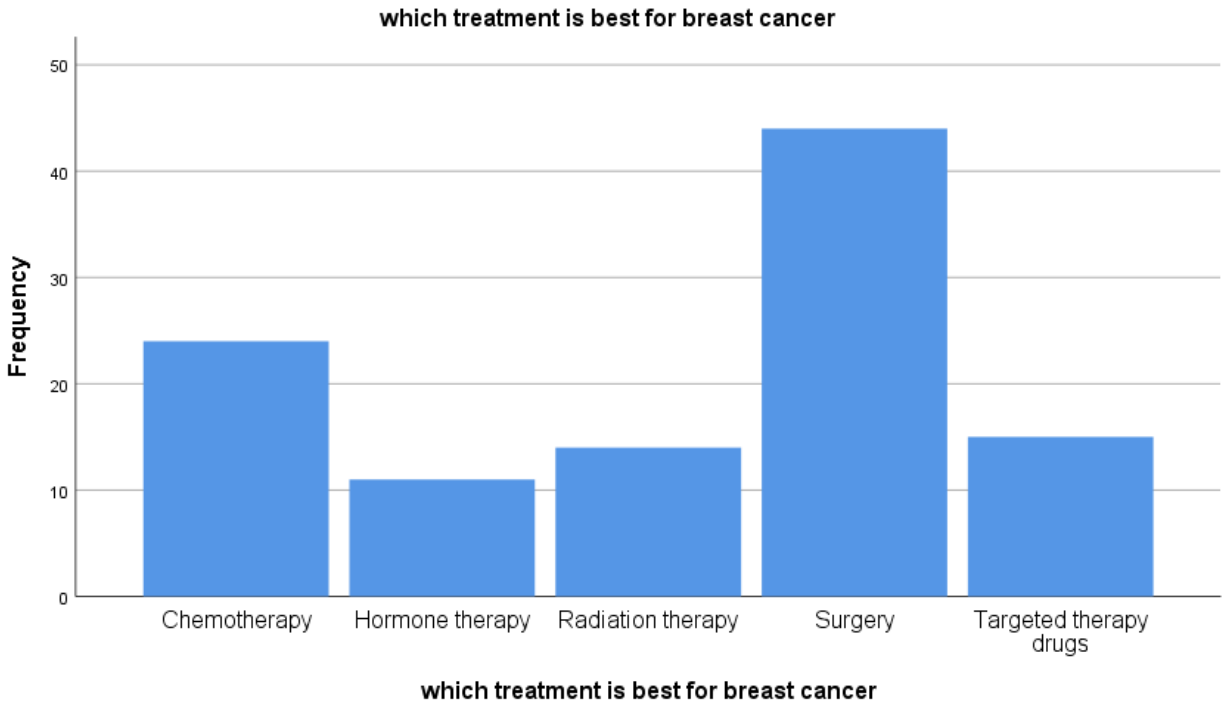


Figure 23 5.1.1.17

We used online survey to get the responds, but after analyzing the dataset in the beginning of graduation project 2 we found some important missing values, also some of the responders were not aware of what symptoms they really have and that made contrary in answers, so that leads to data inaccuracy. And since we didn't do testing before we couldn't find appropriate solutions to these problems in order to fix our dataset (see figure 5.1.1.18). for example, line number 1 answered that they did not felt a mass but also answered that they still feel some pain. Line number 5 answered that they feel all of the other symptoms but not mass questions. Line number 7 answered that also did not felt mass but they answered the size question.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
No	Female	Yes	No	no	no	no	No	No	Yes	Yes	No	No	No	No	No	Surgery
No	Male	Yes	No	no	no	no	No	No	no	No	No	No	No	No	No	Surgery
No	Male	Yes	No	no	no	no	No	No	no	No	No	No	No	No	No	Hormone therapy
No	Female	Yes	No	no	no	no	No	No	no	No	No	No	No	No	No	Chemotherapy
No	Female	Yes	No	no	no	no	No	No	Yes	Yes	Yes	Yes	Yes	Yes	No	Targeted therapy drugs
No	Female	Yes	No	no	no		No	No	no	No	No	No	No	No	No	Chemotherapy
No	Female	Yes	No	No			No	less than 2 cm	Yes	Yes	Yes	Yes	Yes	Yes	No	Radiation therapy
No	Female	Yes	No	No	no	No	No	No	No	No	No	No	No	No	No	Surgery
No	Female	Yes	No	no		No	No	No	No	No	No	No	No		No	Surgery
No	Female	Yes	No	No	no	No	no	No	No	No	No		No	No	No	Chemotherapy
No	Female	Yes	No	no	no	No	no	No	No	No	No		No	No	No	Surgery
No	Female	No		No		No	no	No	No	No	No	No			No	Hormone therapy
No	Female	Yes	No	No	no	No	no	No	No	No	No	No	No	No	No	Surgery
No	Female	Yes		No		No	no	No	No	No	No	No	No		No	Targeted therapy drugs
No	Female	Yes	No	No	No	No	No	No	No	No	No	No	No		No	
No	Male	No	Yes	Yes		Yes	yes	more than 2 cm	No	No	No	Yes	Yes	No	No	
No	Female	Yes	No	No	No		No	No	No	No	No			No	No	Chemotherapy
No	Female	Yes	No	no	no	no	No	No	No	No	No	No	No	No	No	Targeted therapy drugs
Yes	Female	Yes	No	no	no	no	No	No	No	No	No	No	No	No	No	Radiation therapy
No	Female	No	No	no	no	no	No	No	No		No		No		No	Targeted therapy drugs
No	Male	Yes	No	no	no	no	No	No	No	No	No	No	No	No	No	Targeted therapy drugs
No	Female	Yes	Yes	No	no	Yes	No	less than 2 cm	No		No	No	No	No	No	Surgery
No	Female	Yes	No	no	no	no	No	No	no	No	No	No	No	No	No	Radiation therapy

Figure 24 5.1.1.18 smaple of survey dataset

After showing the problem that faced us we decided to use another breast cancer dataset that have been used by most of the previous researchers (see Appendix 2), it is a dataset provided online by UCI Wisconsin machine learning repository. Having 10 real value attribute and class attribute with total 11 attribute and 699 instances. All of information about the dataset are listed down.

5.1.1.1 Wisconsin Dataset

This breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. If you publish results when using this database, then please include this information in your acknowledgements. Also, please cite one or more of:

1. O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

2. William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.

3. O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.

4. K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).

1. Title: Wisconsin Breast Cancer Database (January 8, 1991)

2. Sources:

-- Dr. WIlliam H. Wolberg (physician) University of Wisconsin Hospitals Madison, Wisconsin USA

- Donor: Olvi Mangasarian (mangasarian@cs.wisc.edu)

Received by David W. Aha (aha@cs.jhu.edu) Date: 15 July 1992

3.Past Usage:

Attributes 2 through 10 have been used to represent instances. Each instance has one of 2 possible classes: benign or malignant.

1. Wolberg, W. H., & Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In it Proceedings of the National Academy of Sciences, it 87. 9193--9196.

-Size of data set: only 369 instances (at that point in time)

- Collected classification results: 1 trial only

- Two pairs of parallel hyperplanes were found to be consistent with 50% of the data

- Accuracy on remaining 50% of dataset: 93.5%

- Three pairs of parallel hyperplanes were found to be consistent with 67% of data

- Accuracy on remaining 33% of dataset: 95.9%

2. Zhang, J. (1992). Selecting typical instances in instance-based learning. In it Proceedings of the Ninth International Machine Learning Conference. pp. 470--479. Aberdeen, Scotland: Morgan Kaufmann.

- Size of data set: only 369 instances (at that point in time)

- Applied 4 instance-based learning algorithms

- Collected classification results averaged over 10 trials

- Best accuracy result: 1-nearest neighbor: 93.7%

- trained on 200 instances, tested on the other 169

- Also of interest:

- Using only typical instances: 92.2% (storing only 23.1 instances)

- trained on 200 instances, tested on the other 169

4. Relevant Information:

Samples arrive periodically as Dr. Wolberg reports his clinical cases. The database therefore reflects this chronological grouping of the data. This grouping information appears immediately below, having been removed from the data itself:

Group 1: 367 instances (January 1989)

Group 2: 70 instances (October 1989)

Group 3: 31 instances (February 1990)

Group 4: 17 instances (April 1990)

Group 5: 48 instances (August 1990)

Group 6: 49 instances (Updated January 1991)

Group 7: 31 instances (June 1991)

Group 8: 86 instances (November 1991)

Total: 699 points (as of the donated database on 15 July 1992)

Note that the results summarized above in Past Usage refer to a dataset of size 369, while Group 1 has only 367 instances. This is because it originally contained 369 instances; 2 were removed. The following statements summarizes changes to the original Group 1's set of data:

Group 1: 367 points: 200B 167M (January 1989)

Revised Jan 10, 1991: Replaced zero bare nuclei in 1080185 & 1187805

Revised Nov 22,1991: Removed 765878,4,5,9,7,10,10,10,3,8,1 no record

Removed 484201,2,7,8,8,4,3,10,3,4,1 zero epithelial

Changed 0 to 1 in field 6 of sample 1219406

Changed 0 to 1 in field 8 of following sample: 1182404,2,3,1,1,1,2,0,1,1,1

5. Number of Instances: 699 (as of 15 July 1992)

6. Number of Attributes: 10 plus the class attribute

7. Attribute Information: (class attribute has been moved to last column)

Attribute	**Domain**
1. Sample code number	id number
2. Clump Thickness	1 - 10

3. Uniformity of Cell Size	1 - 10
4. Uniformity of Cell Shape	1 - 10
5. Marginal Adhesion	1 - 10
6. Single Epithelial Cell Size	1 - 10
7. Bare Nuclei	1 - 10
8. Bland Chromatin	1 - 10
9. Normal Nucleoli	1 - 10
10. Mitoses	1 - 10
11. Class:	2 for benign, 4 for malignant

Table 1 Wisconsin Dataset

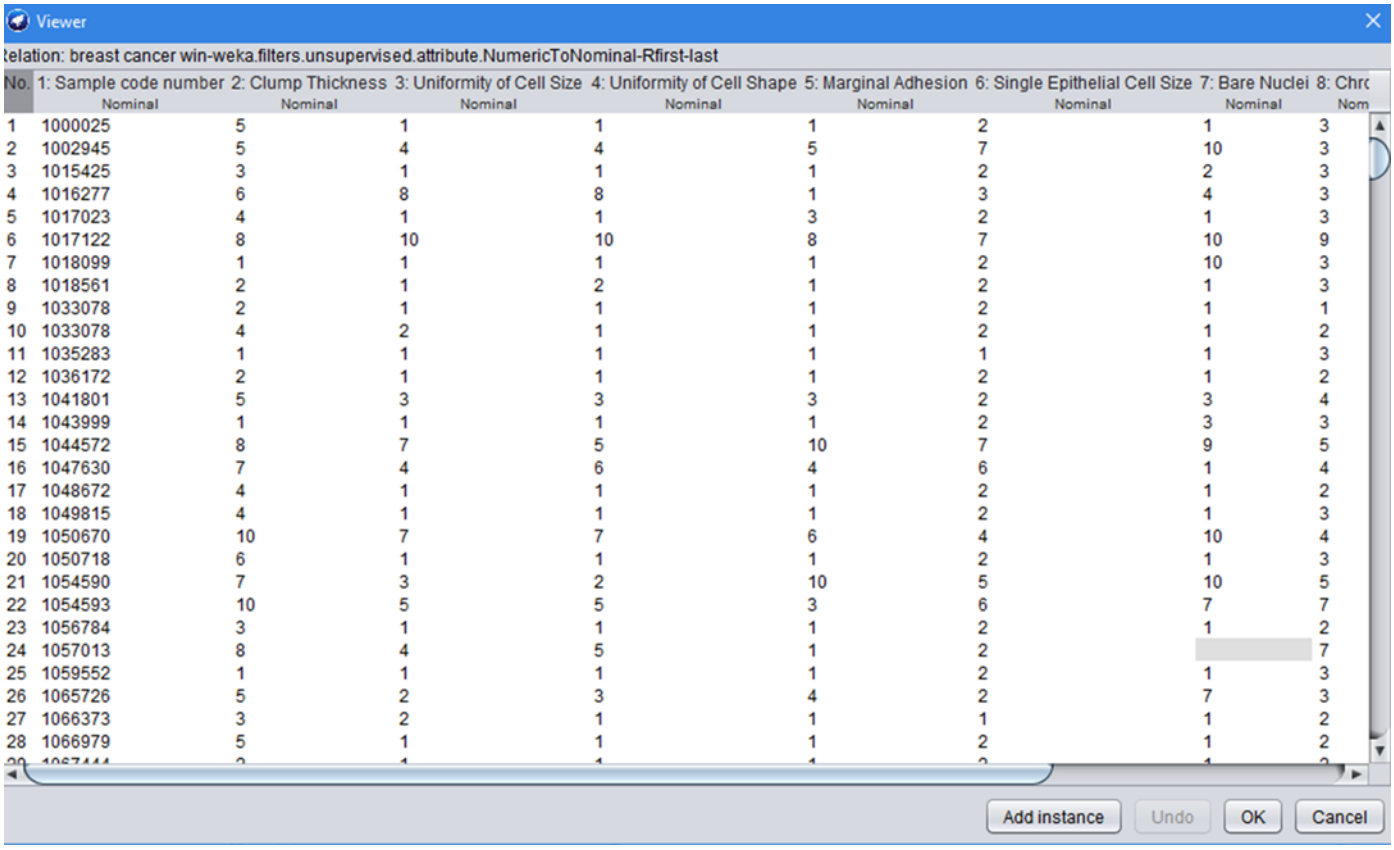
8. Missing attribute values: 16

There are 16 instances in Groups 1 to 6 that contain a single missing (i.e., unavailable) attribute value, now denoted by "?".

9. Class distribution:

Benign: 458 (65.5%) Malignant: 241 (34.5%)

All of information from (The UCI Machine Learning Repository, n.d).



No.	1: Sample code number	2: Clump Thickness	3: Uniformity of Cell Size	4: Uniformity of Cell Shape	5: Marginal Adhesion	6: Single Epithelial Cell Size	7: Bare Nuclei	8: Chro
	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nom
1	1000025	5	1	1	1	2	1	3
2	1002945	5	4	4	5	7	10	3
3	1015425	3	1	1	1	2	2	3
4	1016277	6	8	8	1	3	4	3
5	1017023	4	1	1	3	2	1	3
6	1017122	8	10	10	8	7	10	9
7	1018099	1	1	1	1	2	10	3
8	1018561	2	1	2	1	2	1	3
9	1033078	2	1	1	1	2	1	1
10	1033078	4	2	1	1	2	1	2
11	1035283	1	1	1	1	1	1	3
12	1036172	2	1	1	1	2	1	2
13	1041801	5	3	3	3	2	3	4
14	1043999	1	1	1	1	2	3	3
15	1044572	8	7	5	10	7	9	5
16	1047630	7	4	6	4	6	1	4
17	1048672	4	1	1	1	2	1	2
18	1049815	4	1	1	1	2	1	3
19	1050670	10	7	7	6	4	10	4
20	1050718	6	1	1	1	2	1	3
21	1054590	7	3	2	10	5	10	5
22	1054593	10	5	5	3	6	7	7
23	1056784	3	1	1	1	2	1	2
24	1057013	8	4	5	1	2		7
25	1059552	1	1	1	1	2	1	3
26	1065726	5	2	3	4	2	7	3
27	1066373	3	2	1	1	1	1	2
28	1066979	5	1	1	1	2	1	2

Figure 25 5.1.1.1.1 sample of Wisconsin dataset by WEKA software

5.1.2 Preprocessing

Second step is preprocessing. Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size and their likely origin from multiple, heterogeneous sources. Low-quality data will lead to low-quality mining results. From that preprocessing is concerned with How can the data be preprocessed in order to help improve the quality and efficiency of the data. There are several data preprocessing techniques. Data cleaning can be applied to remove noise and correct inconsistencies in data. Data integration merges data from multiple sources into a coherent data store such as a data warehouse. Data reduction can reduce data size by, for instance, aggregating, eliminating redundant features, or clustering. Data transformations (e.g., normalization) may be applied, where data are scaled to fall within a smaller range like 0.0 to 1.0. This can improve the accuracy and efficiency of mining. (Jiawei, Micheline and Jian, 2012)

Since Wisconsin dataset has already preprocessed multiple of times that makes it an accurate dataset. There weren't a lot to do but, what we did was on the phase of transformation. we convert numeric data to nominal. After searching to find a solution to solve a problem we faced that the classification step will not start for some reason. And converting to nominal

solve this problem. (see figures 5.1.2.1 and 5.1.2.2) in figure 5.1.2.1 the type of data was numeric and figure 5.1.2.2 shows that most of the classification algorithm did not work for it.

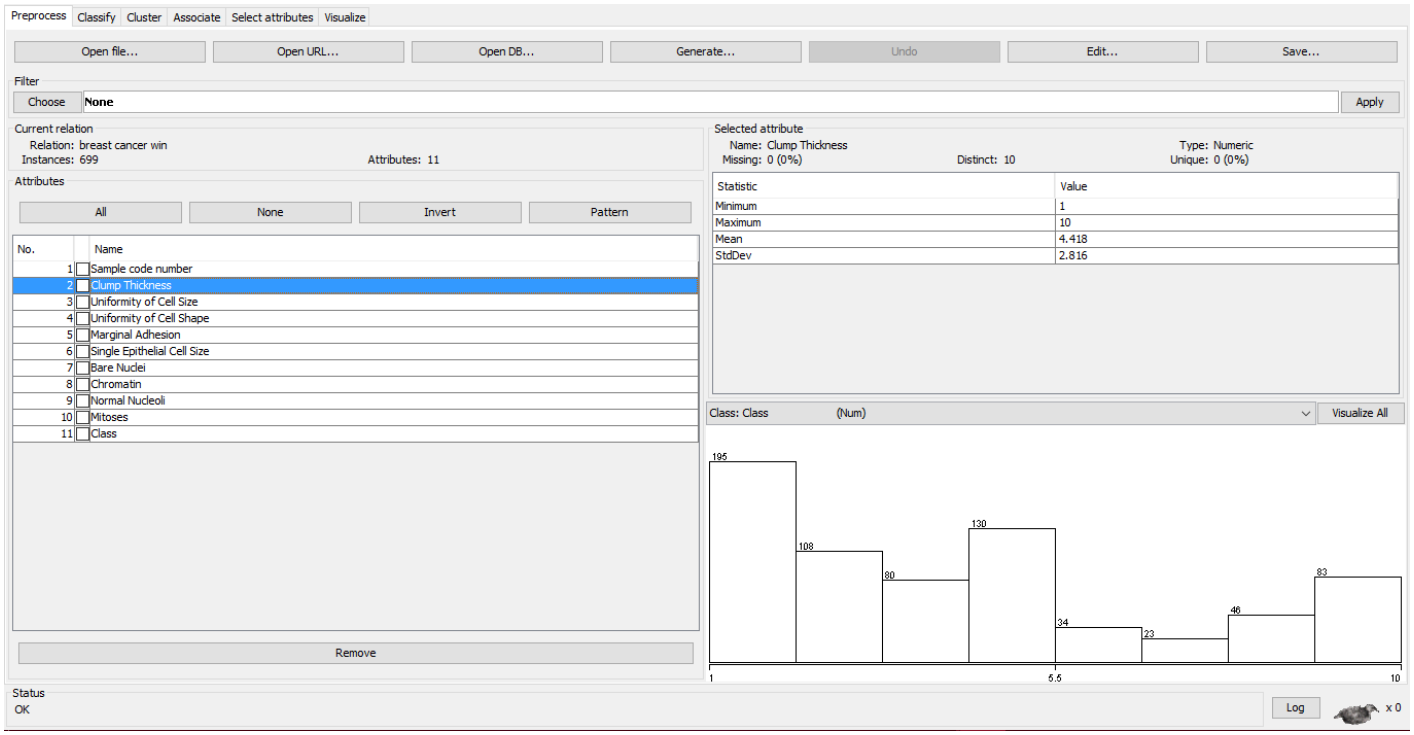


Figure 26 5.1.2.1

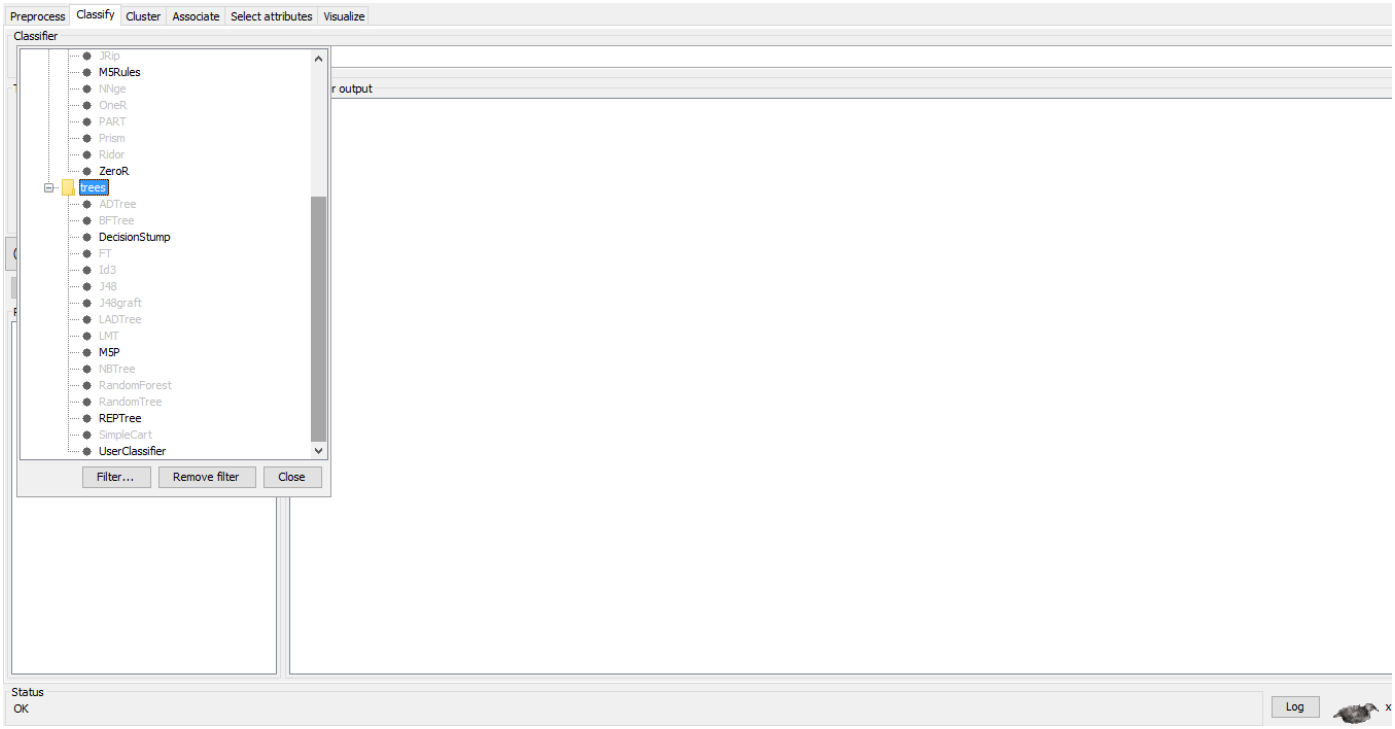


Figure 27 5.1.2.2

5.1.3 Classification

"Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical (discrete, unordered) class labels." (Jiawei, Micheline and Jian ,2012, p 327)

There are many types of classification techniques, let's define the important ones from them in general.

5.1.3.1 Decision Tree Induction

"Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node." (Jiawei, Micheline and Jian ,2012, p 330)

Random Forest, Random Tree, REPTree, J48 and LMT are all decision tree algorithms.

5.1.3.2 Bayes Classification Methods

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities. Studies comparing classification algorithms have found a simple Bayesian classifier known as the naïve Bayesian classifier to be comparable in performance with decision tree and neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases. Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. (Jiawei, Micheline and Jian ,2012, p 350)

5.1.3.3 Rule-Based Classification

A rule-based classifier uses a set of IF-THEN rules for classification. An IF-THEN rule is an expression of the form: IF condition THEN conclusion. (Jiawei, Micheline and Jian ,2012, p 355)

ZeroR algorithm comes under rule based classification.

5.1.3.4 k-Nearest-Neighbor Classifiers

KNN is lazy learner. But first what is lazy learner?

When given a training tuple, a lazy learner simply stores it (or does only a little minor processing) and waits until it is given a test tuple. Only when it sees the test tuple does it perform generalization to classify the tuple based on its similarity to the stored training tuples. lazy learners do less work when a training tuple is presented and more work when making a classification or numeric prediction. (Jiawei, Micheline and Jian ,2012)

k-nearest-neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuples are the k "nearest neighbors" of the unknown tuple. (Jiawei, Micheline and Jian ,2012)

Now we are going to list down the results for each algorithm we tested

5.2 Experimental Results

5.2.1 ZeroR Result Analysis:

Total Instances: 699

Attributes: 11

Test mode: 10-fold cross-validation

ZeroR predicts class value: benign

Correctly Classified Instances	458 (65.5222 %)

Incorrectly Classified Instances	241 (34.4778 %)
Kappa statistic	0
Mean absolute error	0.452
Root mean squared error	0.4753
Relative absolute error	100 %
Root relative squared error	100%
Total Number of Instances	699

Table 2 5.2.1.1 summery for ZeroR Decision tree

TP Rate	FP Rate	Precision	Recall	Class
1	1	0.655	1	2 (benign)
0	0	0	0	4 (malignant)

Table 3 5.2.1.2 Accuracy measures for ZeroR decision tree

Classifier	benign	malignant
A	458	0
B	241	0

Table 4 5.2.1.3 Confusion matrix for ZeroR decision tree

5.2.2 J48 result analyses:

Correctly Classified Instances	660(94.4206%)
Incorrectly Classified Instances	39 (5.5794 %)
Kappa statistic	0.8769
Mean absolute error	0.0796
Root mean squared error	0.218
Relative absolute error	17.6026 %

Root relative squared error	45.8562%
Total Number of Instances	699

Table 5 5.2.2.1 summery for J48

TP Rate	FP Rate	Precision	Recall	Class
0.954	0.075	0.960	0.954	2 (benign)
0.925	0.045	0.914	0.925	4 (malignant)

Table 6 5.2.2.2 Accuracy measures for J48

Classifier	Benign	malignant
A	437	21
B	18	223

Table 7 5.2.2.3 Confusion matrix for J48

5.2.3 SMO result analyses:

Correctly Classified Instances	669(95.7082%)
Incorrectly Classified Instances	30 (4.2918 %)
Kappa statistic	0.9048
Mean absolute error	0.0429
Root mean squared error	0.2072
Relative absolute error	9.4959%
Root relative squared error	43.5866%
Total Number of Instances	699

Table 8 5.2.3.1 summery for SMO

TP Rate	FP Rate	Precision	Recall	Class
0.969	0.066	0.965	0.969	2 (benign)
0.934	0.031	0.941	0.934	4

				(malignant)

Table 9 5.2.3.2 Accuracy measures for SMO

Classifier	Benign	malignant
A	444	14
B	16	225

Table 10 5.2.3.3 Confusion matrix for SMO

5.2.4 BayesNet results analyses:

Correctly Classified Instances	681(97.4249%)
Incorrectly Classified Instances	18(2.5751%)
Kappa statistic	0.9437
Mean absolute error	0.0272
Root mean squared error	0.1586
Relative absolute error	6.023 %
Root relative squared error	33.3594%
Total Number of Instances	699

Table 11 5.2.4.1 summery for BayesNet

TP Rate	FP Rate	Precision	Recall	Class
0.967	0.012	0.993	0.967	2 (benign)
0.988	0.033	0.941	0.988	4 (malignant)

Table 12 5.2.4.2 Accuracy measures for BayesNet

Classifier	Benign	malignant
A	443	15
B	3	238

Table 13 5.2.4.3 Confusion matrix for BayesNet

5.2.5 NaïveBayes results analyses:

Correctly Classified Instances	681(97.4249%)
Incorrectly Classified Instances	18(2.5751%)
Kappa statistic	0.9437

Mean absolute error	0.0276
Root mean squared error	0.1588
Relative absolute error	6.1045 %
Root relative squared error	33.4204%
Total Number of Instances	699

Table 14 5.2.5.1 summery for NaiveBayes

TP Rate	FP Rate	Precision	Recall	Class
0.967	0.012	0.993	0.967	2 (benign)
0.988	0.033	0.941	0.988	4 (malignant)

Table 15 5.2.5.2 Accuracy measures for NaiveBayes

Classifier	Benign	malignant
A	443	15
B	3	238

Table 16 5.2.5.3 Confusion matrix for NaiveBayes

5.2.6 IBK results analyses:

Correctly Classified Instances	668(95.5651%)
Incorrectly Classified Instances	31(4.4349%)
Kappa statistic	0.901
Mean absolute error	0.0467
Root mean squared error	0.1939
Relative absolute error	10.3355 %
Root relative squared error	40.7879%
Total Number of Instances	699

Table 17 5.2.6.1 summery for IBK

TP Rate	FP Rate	Precision	Recall	Class
0.976	0.083	0.957	0.976	2 (benign)
0.917	0.024	0.953	0.917	4

				(malignant)

Table 18 5.2.6.2 Accuracy measures for IBK

Classifier	Benign	malignant
A	447	11
B	20	221

Table 19 5.2.6.3 Confusion matrix for IBK

5.2.7 LBR (Lazy Bayesian Rules) results analyses:

Correctly Classified Instances	681(97.4249%)
Incorrectly Classified Instances	18(2.5751%)
Kappa statistic	0.9437
Mean absolute error	0.0276
Root mean squared error	0.1588
Relative absolute error	6.1016 %
Root relative squared error	33.4137%
Total Number of Instances	699

Table 20 5.2.7.1 summery for LBR

TP Rate	FP Rate	Precision	Recall	Class
0.967	0.012	0.993	0.967	2 (benign)
0.988	0.033	0.941	0.988	4 (malignant)

Table 21 5.2.7.2 Accuracy measures for LBR

Classifier	Benign	malignant
A	443	15
B	3	238

Table 22 5.2.7.3 Confusion matrix for LBR

5.2.8 REPTree results analyses:

Correctly Classified Instances	467(66.8097%)
Incorrectly Classified Instances	232(33.1903%)
Kappa statistic	0.0532

Mean absolute error	0.4125
Root mean squared error	0.4623
Relative absolute error	91.2727%
Root relative squared error	97.2674%
Total Number of Instances	699

Table 23 5.2.8.1 summery for REPTree

TP Rate	FP Rate	Precision	Recall	Class
0.996	0.954	0.665	0.996	2 (benign)
0.046	0.004	0.846	0.046	4 (malignant)

Table 24 5.2.8.2 Accuracy measures for REPTree

Classifier	Benign	malignant
A	456	2
B	230	11

Table 25 5.2.8.3 Confusion matrix for REPTree

5.2.9 RandomForest results analyses:

Correctly Classified Instances	669(95.7082%)
Incorrectly Classified Instances	30(4.2918%)
Kappa statistic	0.9043
Mean absolute error	0.1935
Root mean squared error	0.2478
Relative absolute error	42.8117 %
Root relative squared error	52.1362%
Total Number of Instances	699

Table 26 5.2.9.1 summery for RandomForest

TP Rate	FP Rate	Precision	Recall	Class
0.976	0.079	0.959	0.976	2 (benign)
0.921	0.024	0.953	0.921	4

				(malignant)

Table 27 5.2.9.2 Accuracy measures for RandomForest

Classifier	Benign	malignant
A	447	11
B	19	222

Table 28 5.2.9.3 Confusion matrix for RandomForest

5.2.10 Final results:

We analyzed UCI data set using WEKA tool and compared several algorithms to find the best algorithm among them and the result we got is that LMT algorithm gives a good accuracy with 99.85%

But first let's discuss what is LMT?

Logistic Model Tree algorithm, or LMT for short. It combines the logistic regression models with tree induction. A logistic model tree basically consists of a standard decision tree structure with logistic regression functions at the leaves, much like a model tree is a regression tree with regression functions at the leaves. As in ordinary decision trees, a test on one of the attributes is associated with every inner node. For a nominal (enumerated) attribute with *k* values, the node has *k* child nodes, and instances are sorted down one of the *k* branches depending on their value of the attribute. (Niels, Mark and Eibe, 2005)

the following algorithm for building logistic model trees:

− Tree growing starts by building a logistic model at the root using the LogitBoost algorithm to iteratively fit simple linear regression functions, using fivefold cross validation.
− A split for the data at the root is constructed. Splits are either binary (for numeric attributes) or multiway (for nominal ones). Tree growing continues by sorting the appropriate subsets of data to the child nodes and building the logistic models at the child nodes in the following way: The LogitBoost algorithm is run on the subset associated with the child node, but starting with the committee, weights and probability estimates of the last iteration performed at the parent node.
− Splitting of the child nodes continues in this fashion until some stopping criterion is met.
− Once the tree has been built it is pruned using CART-based pruning. (Niels, Mark and Eibe, 2005)

From WEKA software: "numBoostingIterations - Set a fixed number of iterations for LogitBoost. If >= 0, this sets a fixed number of LogitBoost iterations that is used everywhere in the tree. If < 0, the number is cross-validated."
And for our experiment the value of numBoostingIteration was originally set to -1, so in order to increase the accuracy we fixed it to 5. And that given us an increased accuracy from 98% to 99.85% with 698 correctly classified instances out of 699.

Correctly Classified Instances	698 (99.8569 %)
Incorrectly Classified Instances	1 (0.1431 %)
Kappa statistic	0.9968
Mean absolute error	0.016
Root mean squared error	0.0585
Relative absolute error	3.5374%
Root relative squared error	12.318%
Total Number of Instances	699

Table 29 5.2.10.1 Summery for LMT

TP Rate	FP Rate	Precision	Recall	Class
1	0.004	0.998	1	2 (benign)
0.996	0	1	0.996	4 (malignant)

Table 30 5.2.10.2 Accuracy measures for LMT

Classifier	benign	malignant
A	458	0
B	1	240

Table 31 5.2.10.3 Confusion matrix for LMT

5.2.11 Comparison of all algorithm:

Classifier	**Accuracy**
ZeroR	65.5222 %

J48	94.4206 %
SMO	95.7082 %
NaiveBayes	97.4249 %
BayesNet	97.4249 %
IBK	95.5651 %
LBR (Lazy Bayesian Rules)	97.4249 %
REPtree	66.8097 %
Random Forest	95.7082 %
LMT	99.8569 %

Table 32 5.2.11.1 Comparison of Algorithms

Chapter 6: Conclusion and Future Work

6.1 Conclusion:

Web data is huge and can be used in many ways to help the world to extract useful information using data mining techniques and methods. In this research we studied and compared different classification algorithm to determine the best classification algorithm to detect breast cancer. LMT logistic model tree achieved higher accuracy comparing to another classification algorithm. Using Wisconsin breast cancer dataset, LMT detected the presence of breast cancer determining cancer cells to be either benign or malignant. this algorithm gives 99.85 % accuracy with 698 correct classified instances out of 699. This classification model will be useful in future to determine the existence of tumors.

6.2 Future Work:

LMT can be used to detect different diseases since it gives high accuracy it has high expectation to success on different fields. The research on this algorithm is in it is beginning, so there is a lot of development to be achieved as time complexity can be a problem for LMT but I believe it can be fixed in future work.

References:

Ahmed, M. and Ayman, M. (2016). Breast cancer classification using deep belief networks. Expert Systems with Applications,46:139–144

Deeba, K., and Amutha, b. (2016). Classification Algorithms of Data Mining. *Indian Journal of Science and Technology*, *9*(39), 0974-5645

Fincher, Sally, Marian Petre, and Martyn Clark, eds. *Computer science project work: principles and pragmatics*. Springer Science & Business Media, 2001.

Jahanvi, J., Rinal D., & Jigar P., Ph. (2014). Diagnosis of Breast Cancer using Clustering Data Mining Approach. *International Journal of Computer Applications,101*(10), 0975 – 8887

Jiawei, H., Micheline, K., & Jian, P. (2012). *Data Mining Concepts and Techniques* (3rd ed). Waltham, USA: Morgan Kaufmann

John, W., Robert, B., & Stephen, D. (2012). *Systems Analysis and Design in a Changing World* (6th ed). Boston, USA: Joe Sabatino.

Lucid chart Inc. (n.d.).retrieved from www.lucidchart.com

Mayo clinic. (June 05, 2014). *Recurrent breast cancer*. Retrieved from https://www.mayoclinic.org/diseases-conditions/recurrent-breast-cancer/symptoms-causes/syc-20377135

Ministry of health. (19/Jumada al-Thani/1436). *National Breast Cancer Awareness Campaign*. Retrieved from https://www.moh.gov.sa/HealthAwareness/Campaigns/Breastcancer/Page/stat.aspx

Niels, L., Mark, H., and Eibe, F. (2005). Logistic model trees. Machine Learning,59(1/2):161–205.

Sommerville, I. (2011). *Software engineering* (9th ed). Boston, USA: Pearson Education.

Souad Demigha, (2016). Mining Knowledge of the Patient Record: "The Bayesian Classification to Predict and Detect Anomalies in Breast Cancer". *The Electronic Journal of Knowledge Management, 14*(3) (pp128-139), 1479-4411.

The UCI Machine Learning Repository. (n.d.). *UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set*. Retrieved from https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)

The University of Waikato. (n.d.). Weka 3 - Data Mining with Open Source Machine Learning software in java. Retrieved from https://www.cs.waikato.ac.nz/ml/weka/

Vikas, C., and Saurabh, P. (2014). A Novel Approach for Breast Cancer Detection using Data Mining Techniques. *International Journal of Innovative Research in Computer and Communication Engineering*, *2(*1), 2320-9801.