Faculty of Graduate Studies

Master of applied statistics & Data Science

Comparing several machine learning algorithms in predicting breast cancer

المقارنة بين الادوات المختلفة لتعلم الألة للتنبؤ بسرطان الثدي

A Master Thesis

By

Ashraf Fashafsheh

1185207

Supervisor

Dr. Hassan Abu Hassan

2021

Faculty of Graduate Studies

Master of applied statistics & Data Science

Comparing several machine learning algorithms in predicting breast cancer

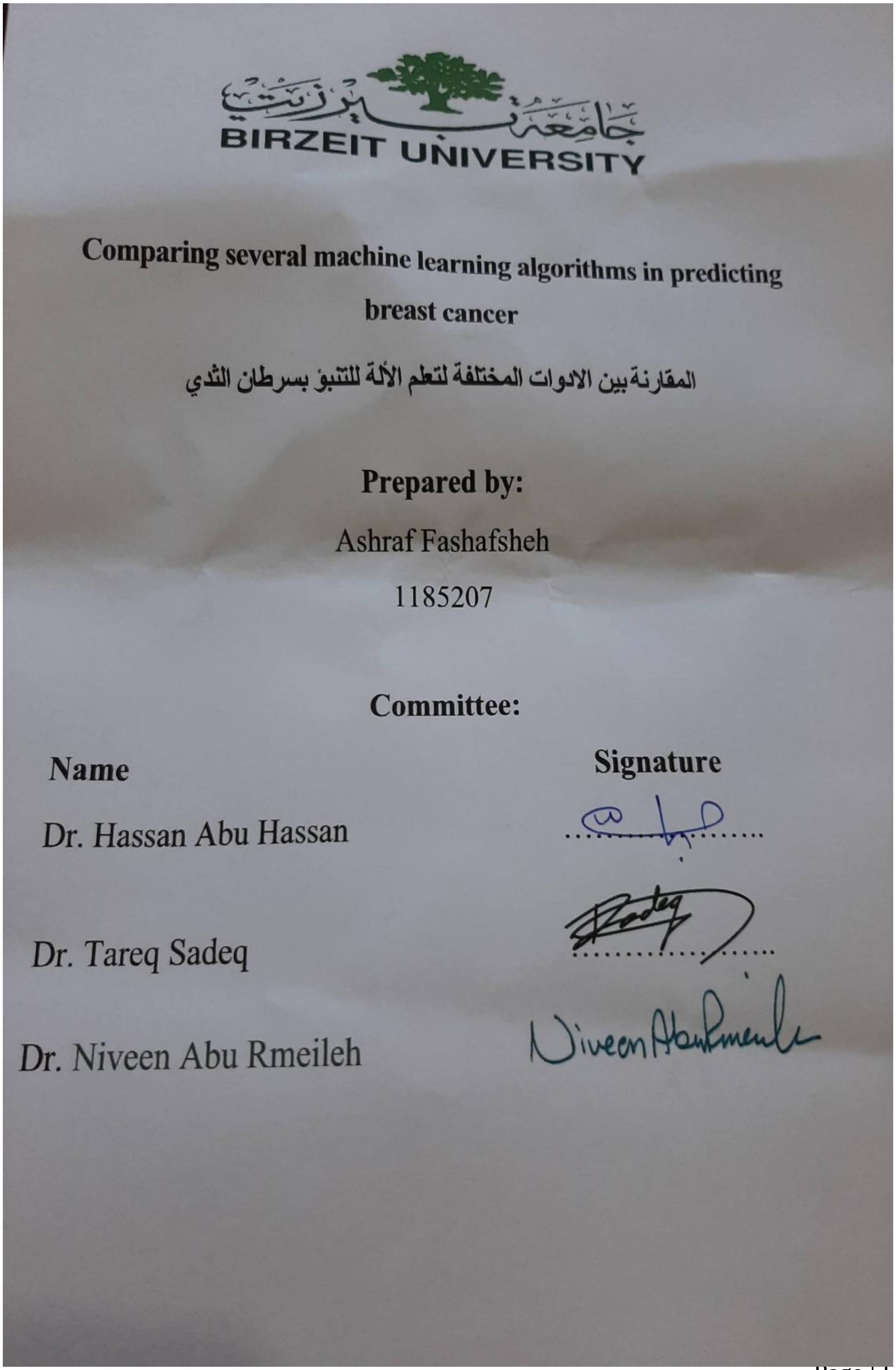
المقارنة بين الادوات المختلفة لتعلم الألة للتنبؤ بسرطان الثدي

A Master Thesis

By : Ashraf Fashafsheh

Supervisor : Dr. Hassan Abu Hassan

Submitted in partial fulfillment of the requirements of the "Master Degree in Applied Statistic and Data Science" from the faculty of Graduate Studies at Birzeit University - Palestine



BIRZEIT UNIVERSITY

جامعة بيرزيت

Comparing several machine learning algorithms in predicting breast cancer

المقارنة بين الادوات المختلفة لتعلم الألة للتنبؤ بسرطان الثدي

Prepared by:

Ashraf Fashafsheh

1185207

Committee:

Name	**Signature**
Dr. Hassan Abu Hassan	…………………
Dr. Tareq Sadeq	Sadeq …………………
Dr. Niveen Abu Rmeileh	Niveen AbuRmeile

Dedication:

To those who have supported me in all stages of my studies.

To my parents, brothers, and my friends

To my wife and my children without whom this thesis would not have been completed.

I dedicate this thesis.

With all my love and respect

Acknowledgment:

I would like to express many thanks to my supervisor, Dr. Hassan Abu Hassan, for his effort, valuable comments, guidance, and assistance during my research; it would have been next to impossible to write this thesis without his help. It is a pleasure to thank all the administrative staff, especially Lina Al-Jundi Administrative Assistant - Faculty Of Graduate Studies.

Contents

List of Figures:

List of Tables

List of Abbreviations:

ANN	Artificial Neural Network
SVM	Support Vector Machine
KNN	K-Nearest Neighbor Algorithm
DT	Decision Tree
MOH	Ministry Of Health
BRCA1	Breast cancer gene 1
BRCA2	Breast cancer gene 2
WHO	World Health Organization
DM	Data Mining
ML	Machine Learning
WEKA	Waikato Environment for Knowledge Analysis
AUC	Area Under the Curve
ROC curve	Receiver Operating Characteristic curve
MLP	Multilayer Perceptron
RMSE	Root Mean Square Error

Abstract:

Breast cancer is more common among females in Palestine compared to other types of cancer, as the number of women diagnosed by this disease increases significantly from 2016 to 2019) according to the reports of the Palestinian Ministry of Health, the incidence of breast cancer increased from 13.2 cases per 100,000 females in 2016 to 40 cases per 100,000 females people in 2019, they attribute this increase to the patients not being diagnosed in the early stages.

This research aims to determine the most accurate algorithm among the algorithms used in the search (SVM, DT, NB, and ANN), as well as to determine the factors most associated with the disease according to the data of the Palestinian Ministry of Health.

The study is a collection of secondary data from the directorates of the Palestinian Ministry of Health in the West Bank, where 1140 cases were diagnosed in the period from April 2019 to October 2020.

The data were analyzed using the WEKA software package, as well as the R program. And tables and drawings were used for analysis, where the SVM algorithm was identified as the best algorithm of the algorithms used in the research. Also, the factors most associated with the disease were: the shape and size of the mass, Alcohol consumption, hormonal therapy, the family record, and place of residence (Region).

ملخص:

يعتبر سرطان الثدي اكثر شيوعا بين الاناث في فلسطين مقارنة مع أنواع السرطان الأخرى ,حيت تتزايد اعداد الاصابات في هذا المرض بشكل ملحوظ, حسب تقارير وزارة الصحة الفلسطينية فان نسبة الإصابة بمرض سرطان الثدي ارتفعت من 13.2 حالة لكل 100000أنثى في عام 2016 الى 40 حالة لكل 100000أنثى في عام 2019 , ويعزو هذا الارتفاع الى عدم تشخيص المرضى في المراحل المبكرة.

وهدف هذا البحث لتحديد الخوارزمية الأكثر دقة من بين الخوارزميات المستخدمة في البحث (SVM, DT, NB, and ANN), كذلك الامر تحديد العوامل الأكثر ارتباطا بالمرض حسب بيانات وزارة الصحة الفلسطينية. وذلك من اجل الحد من الارتفاع نسب الإصابة بالمرض.

والدراسة هي عبارة عن جمع بيانات ثانوية من مديريات وزارة الصحة الفلسطينية في الضفة الغربية, حيث تم جمع 1140 حالة تم تشخيصه في الفترة الواقعة من شهر نيسان 2019 الى شهر تشرين اول من عام 2020.

تم تحليل البيانات بواسطة الحزمة البرمجية ويكا وكذلك الامر تم استخدام برنامج R. وقد تم عمل الجداول والرسومات الضروية للتحليل ,حيث تم تحديد الخوارزمية SVM كافضل خوارزمية من الخوارزميات المستخدمة في البحث .وكانت العوامل الأكثر ارتباطا بالمرض هي :شكل وحجم الكتلة ,وتناول الكحول ,واستخدام الهرمونات و السجل العائلي ومكان السكن.

Chapter One

Introduction

1.1 Background

In 2018, approximately 9.6 million deaths occurred as a result of cancer diseases. This is the second leading cause of death globally. The most common types of cancers in men are lung, stomach, colorectal, and liver cancer, while the most common types among women are breast, colorectal, lung, cervical, and thyroid cancer. (World Health Organization, 2018)

Cancer is a rapid reproduction and unusual growth in the cells of one of the bodys organs and this is not subject to it tumor of factors that regulate and control the growth and division of organ cells under normal conditions. The most frequently diagnosed cancer among females is breast cancer.It arises from the breast tissues either from the inner lining of milk or from lobules that supply the ducts with milk. (Sumbaly, Vishnusri & Jeyalatha, 2014)

There are two types of tumors:

- Benign: this type of tumor is not dangerous to the human and cannot invade neighboring tissues.

- Malignant: this type of tumor is dangerous for the human and can invade nearby tissues and lead to death**.**

Breast cancer is the most common deadly cancer among women in Palestine. The breast cancer cases increased very rapidly between 2016 and 2019. In 2016, breast cancer ranked first in the West Bank among all types of cancer, with 388 cases, or 15.3% of all reported cancer cases. Breast cancer was the first type of cancer reported among females, as it constituted 28.9% of the total cancer cases reported among females in Palestine, with an incidence rate of 13.2 per 100,000 females. In 2019, there were 529 cases, representing 31.8% of all cancer cases, with a rate of 40 cases per 100,000 females. (Annual Health Report, Palestine 2019 - Ministry of Health)

Symptoms of Brest cancer:

Breast cancer often has no symptoms, however, sometimes symptoms may occur at an advanced stage:

- ✓ A solid painless lump in the breast or under the arm.
- ✓ Swelling in the breast.
- ✓ Unusual discharges from the breast.
- ✓ Change in the size, shape of the breast, or wrinkly skin.
- ✓ Inverted nipple.
- ✓ Itching, ulcers, or rash around the breast.
- ✓ There is rarely a feeling of pain.

The appearance of lumps does not necessarily mean that it is cancer; it may be due to the presence of cysts or infection. (Ministry of Health - Kingdom of Saudi Arabia)
The applications of Machine Learning (ML) in medical science may play an essential role in cancer care because of its high performance in predicting outcomes (*malignant* tumor or benign tumor), reducing costs of medicine, promoting patients' health. Moreover, they improve healthcare value and quality and help in taking correct decisions to save people's lives. (Asri, Mousannif, Al Moatassime & Noel, 2016)

Recently, the Machine Learning role has spread, and it's used in classifying and predicting cancer disease, also, ML is used to classify and predict breast cancer. We studied four classifiers: SVM, Navies' Base, DT, and ANN which are among the most accurate machine learning algorithms.

We aim to evaluate the efficiency and effectiveness of those algorithms in terms of accuracy, sensitivity, specificity, and precision, and to decide which is the best prediction algorithm for breast cancer.

1.2 Research Problem:

Given the lack of local studies that talk about predicting breast cancer through Machine Learning tools and the Comparison of machine learning tools in terms of prediction accuracy, the objectives of the study can be summarized as follows.

- How to predict breast cancer using Machine Learning tools.
- Comparison of several machine learning tools to predict breast cancer in terms of predictive accuracy and to determine the best among them.

1.3 **Research Objectives**:

The main objective of the research is to use a machine learning algorithm on real data of breast cancer patients from the ministry of health in the West Bank to predict breast cancer based on some attributes of the patients.

Additionally, this research aims to identify the most related attributes that affect breast cancer. However the specific objectives of the thesis are:

1. Apply several ML algorithms (SVM, DT, NB, and ANN) to predict breast cancer, and determine the best one, among them, in terms of the prediction accuracy.
2. Calculate classification performance using metrics such as accuracy, sensitivity, and specificity.
3. Identify the most related attributes that affect breast cancer.

1.4 Importance of the study:

Providing the best algorithm performance in machine learning (SVM, NB, ANN,DT)in order to use breast cancer prediction.

What distinguishes our study is the determination of the most accurate algorithm in the process of predicting breast cancer likewise, knowing the most important variables that causing the disease including those controllable by humans such as smoking, to help prevent disease.

1.5 Research Ethics:

Prior to the commencement of the study, ethical compliance was obtained to protect participants. The data that was used during this research was collected from 15 cancer centers in Palestine Heath Directorates. Before the collection of data, official approval was sent to the Palestinian Ministry of Health, after getting a written permission for the collection and use of data shown in the Appendix.

The researcher undertakes to not misuse the data, during the research and reference to previous studies, copyrights were preserved through quotation and was carried out honestly.

1.6 Thesis Outline

The study is divided into five chapters:

1. Introduction.
2. Literature Review.
3. The Methodology
4. Results and Discussion
5. Conclusions and Recommendations.

Chapter Two

Literature Review

2.1 Literature Review

In recent years, there has been more literature on breast cancer using data mining and Machine Learning such as:

There are many algorithms of Machine learning that are used by researchers in the problem of classification, such as decision trees, Bayesian classifiers, SVM, and neural networks, to determine which one of all these has the best performance.

Anusha et al (2018) compared four algorithms k-Nearest Neighbours (KNN), Decision Tree, Naive Bayes (NB), and Support Vector Machine (SVM) Classifiers of Breast Cancer Detection, the paper is compared between prediction accuracy for all algorithms used in this study, they concluded that SVM using the Gaussian kernel is the fittest technique for recurrence/non-recurrence prediction of breast cancer.

(Mengjie,2017) tested four different machine learning algorithms (such as SVM, KNN, Logistic Regression, and Naive Bayes in this model) for breast cancer prediction. The main component analysis was used to reduce the dimension for the original correlated dataset; he concluded that the highest Area Under the Curve (AUC) value of 0.9944 was achieved by SVM with a linear kernel.

(Akinsola et al.,2017) evaluated and investigated three selected classification algorithms using the Waikato Environment for Knowledge Analysis (WEKA for short). WEKA is an open-source data mining software mainly used for research purposes and academic. Experimental results show that the Decision Tree (C4.5) proves to be the best algorithm with the highest accuracy from other algorithms tested Decision Tree (C4.5), Multilayer Perceptron, and Naive Bayes.

(Hiba et al., 2016) compared four machine learning algorithms: Vector Machine Support (SVM), Decision Tree (C4.5), Naive Bayes (NB), and k-Nearest Neighbors (k-NN). The researchers found that the best classification accuracy is the support vector machines (SVM), where the accuracy reached 97.13%. Likewise, the SVMs have high efficiency in predicting and diagnosing breast cancer and achieve the best performance in terms of accuracy and low error.

(Deepika et al.2016) compared between Artificial Neural Network (ANN), Genetic Algorithms (GA), Support Vector Machines (SVM), Decision Trees (DT), and Genetic Algorithms (GA), Classifier of Breast Cancer Detection, the paper aims to review various data mining techniques that are specifically considered on breast cancer prediction. He infers that there is still lacking early diagnosis, accuracy, sensitivity, and specificity of the breast cancer data.

(Tolga Ensari, 2019) compared the most commonly used machine learning techniques to classify the performance of these technologies, using the values of accuracy, ROC area, retrieval, and accuracy. The best performance is also obtained through the Support Vector Machine technology with the highest accuracy.

(Ali Al Bataineh, 2019) evaluated the performance in classifying data concerning the efficiency and effectiveness of five machine learning algorithms via Multilayer Perceptron (MLP), K-Nearest Neighbors (KNN), Classification and Regression Trees (CART), Gaussian Nave Bayes (NB), and Support Vector Machines (SVM).

(Mohammed,2017) built a model to solve the difficulty of determining the level of risk to disease and the best practices, and applied classification technique such as SVM, ANN, and KNN in this model, the model achieved an accuracy of 78%, ,a list of the most related attributes that affect breast cancer risk is produced.

There is a need to fine-tuning parameters for algorithms of machine learning predictive model to get a result in better accuracy, recall, and precision.

(Abdull,2011) used three Machine-Learning Techniques: Decision Trees (DTs), Neural Networks (NNs), and Logistic Regression (LR). The performance of logistic and neural network classifiers can be enhanced by utilizing significant inputs (independent variables) using forward selection algorithms instead of selection algorithms. These significant variables, are: family history, height, length of breastfeeding, work related to radiation, breastfeeding, kinds of meat consumed, sporting activity and weight, are utilized in the final model.

(The American Cancer Society, 2019-2020) in a report concerning breast cancer found that about a 33% of postmenopausal breast cancer cases were associated with modifiable factors like postmenopausal obesity, lack of physical activity, alcohol consumption, and lack of breastfeeding.

(Deepika et al.,2016) compared between Decision Trees (DT), Artificial Neural Network (ANN), Genetic Algorithms (GA), and Support Vector Machines (SVM) Classifier of Breast Cancer Detection, he inferred that there is a still lack of early diagnosis, accuracy, sensitivity, and specificity of the breast cancer data. In their research, they reviewed a group of variables (risk factors) that led to the infection of breast cancer. (American Cancer Society,2020).

(Mümine KAYA KELEŞ,2019) revealed that many researchers have concluded that data mining methods play an important role in diagnosing breast cancer. And to diagnose it, they used the Weka data mining tool. Likewise, they tested all classification algorithms and determined the most successful algorithms based on accuracy rates.

2.2 Machine Learning (ML) Algorithms

ML, a branch of artificial intelligence, is concerned with learning from data samples to the concept of general inference. Likewise, ML is important in biomedical sciences, where acceptable generalization is achieved by searching in the n dimensional space of a specific group of biological samples using different algorithms. (Kourou et al., 2015)

Machine learning algorithms are classified into two types as shown in Fig.3.3.

- Supervised learning: a set of labeled training data is used to estimate the input data or map the input data to the desired output.
- Unsupervised learning: no labeled examples are provided, and nothing is known about learning outcomes.

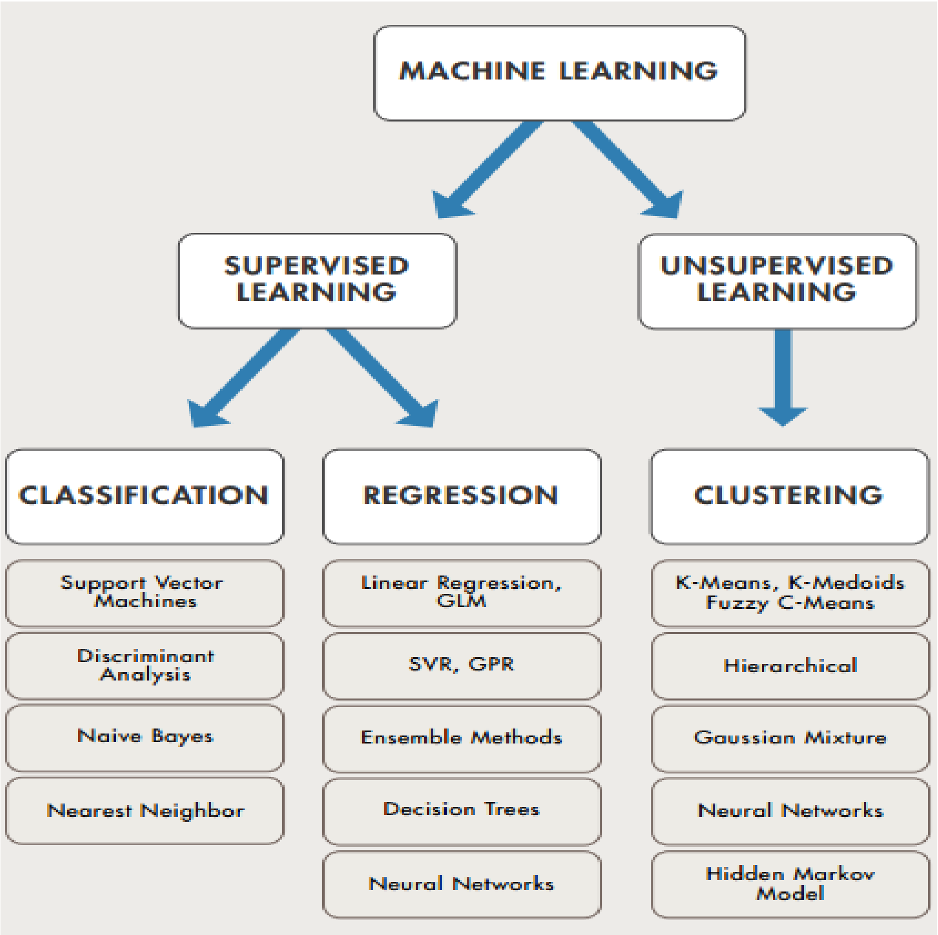


Figure (2.1): Supervised Learning versus Unsupervised Learning, by (Vincent Granville, 2017).

2.2.1 Support Vector Machines:

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification and regression problems. SVMs map the input vector into a higher dimensional feature space and define the hyperplane that divides the data points into two groups. The marginal distance between the decision hyperplane and the cases nearest to the boundary is maximized as shown in Figure 3.4. The resulting classifier achieves substantial generalizability and can therefore be used for the accurate classification of new samples. (Bharat et al., 2018) Vector Machine (SVM) framework has the same benefits as parametric techniques with respect to reduced computing time for testing and storage requirements. (Awad, Khanna., 2015)

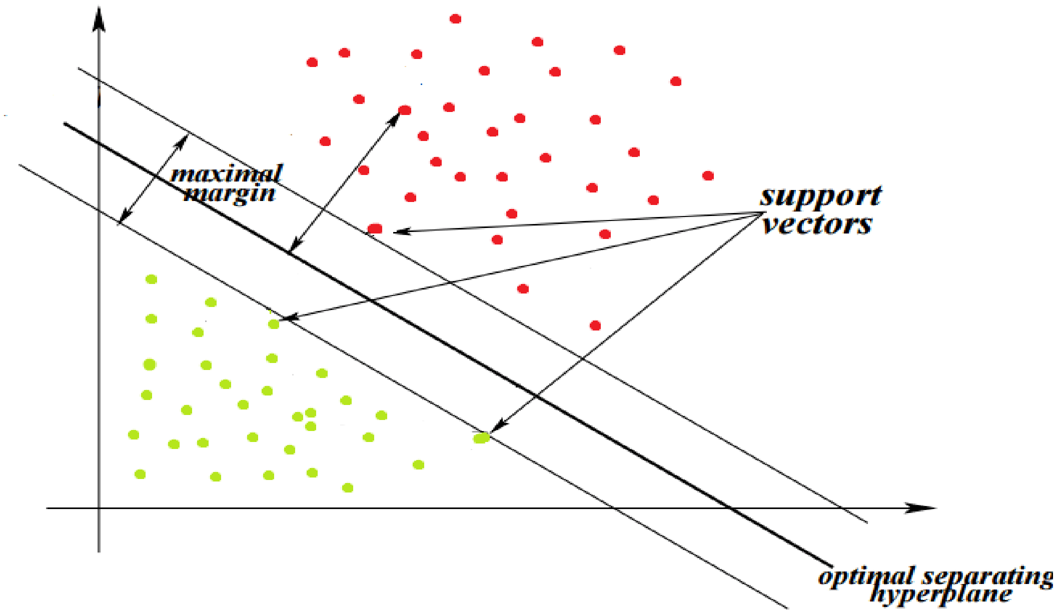


Figure 2.2 How SVM works (Rashmi Jain,2017)

2.2.2 Naive Bayes:

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is especially suited when the dimensions of the inputs are large (Rathi,2012), and it's one of the most popular machine learning methods, as it is characterized by the speed in processing and efficiency in prediction operations it relies on the statistical concept Bayes ’theorem, which calculates the probability of a specific result by verifying what is available and known and is called Naive because it depends on the principle of Independence Assumptions.

p$(k|y)$is estimated by following Bayes theorem, which assumes that the attributes are independent given the target class

p$(k_{1,}k_{2,}k_{3,}$…..,k_n|y)=p$(k_1|y)$ p$(k_2|y)$p$(k_3|y)$….. p$(k_n|y)$

Where k_i is the i^{th} the dimension of the feature vector, y is the target variable (dependent variable).

And the Naïve Bayes classifier can be then written as

p$(k|y)$=p(y)$\prod_{i=1}^{N}$ p$(k_i$ | y).

Find the probability of a given set of inputs for all possible values of the class variable y to construct a classifier model and pick up the output with the highest probability. The following formula can be used to express this.

$y = arg_{y\in Y} maxp(y)$ $\prod_{i=1}^{N}$ p$(k_i{}^{new}$| y).

Where $arg_{y\in Y} maxp(y)$: - maximum value of the class.

2.2.3 Artificial Neural Network:

An artificial neural network (ANN) is a computing system inspired by a biological nervous system. An ANN consists on a collection of connected nodes called artificial neurons. The neurons are connected by weighted links passing signals. (El-Shahat, 2014)

Neural networks are the same as SVMs in which a supervised machine learning algorithm can handle several classification or pattern recognition problems.

They are trained to produce the output as a set of input variables. As shown in Fig (2.3), multiple hidden layers representing the neural connections are used for regression and classification. (Kourou et al., 2015)

Neural Network Features are:

- ✓ Serves as a gold standard method in several classification tasks.
- ✓ Is characterized as a "black-box” technology.

Among their disadvantages that they're generic layered structure proves to be time-consuming while it can lead to very poor performance. (Kourou et al., 2015)

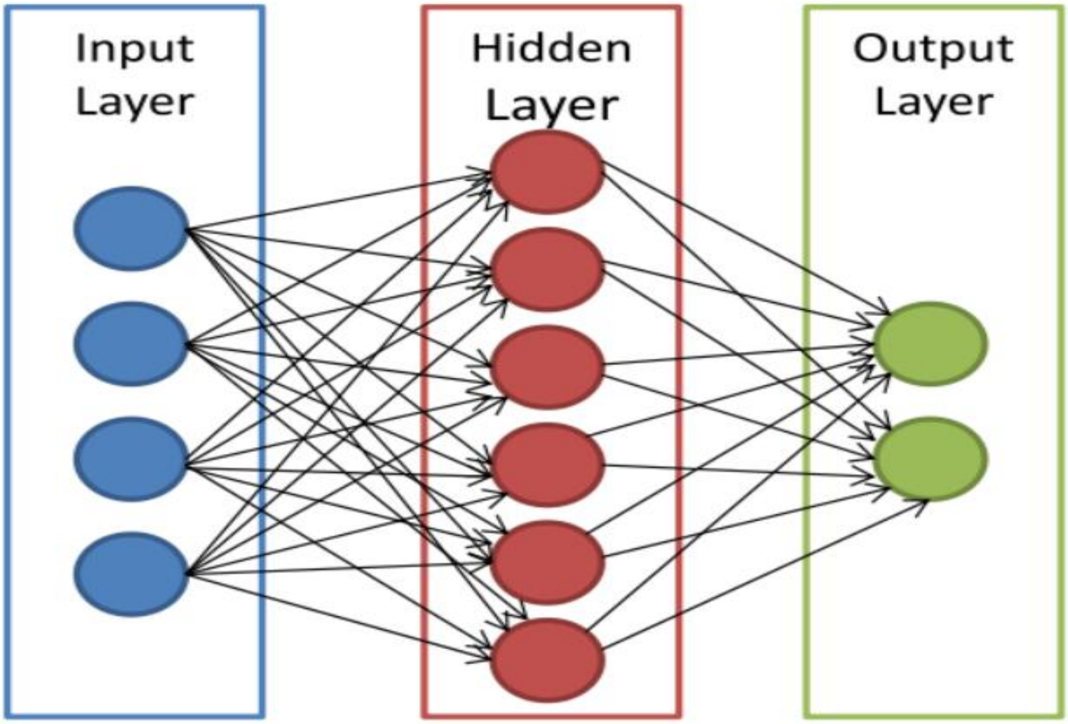


Figure (2.3)An illustration of the ANN structure. The arrows connect the output of one node to the input of another. (Konstantina et al, 2015)

2.2.4 Decision Tree:

A decision tree (DT) is a supervised learning method that is used for classification and regression. A decision tree helps decision-makers in knowing all of the possible alternatives and the possibilities of obtaining and using the best option among future cases, for example, Predict whether a bank client will be a good debtor or not.

Decision Tree Induction Approach is one of the key approach to the process of knowledge discovery. The structure of the tree consists of root nodes, intermediate nodes and leaves. This is the intermediate nodes are the symbol of the decision test and the results are indicated by the edges. The node of the leaf is associated with a label of class. This method works on the basis of the rules it has been established. This technique may be a very efficient method for the prediction of breast cancers. The gain from this the model is that it's very fast. (Preetha and Vanilla. , 2019)

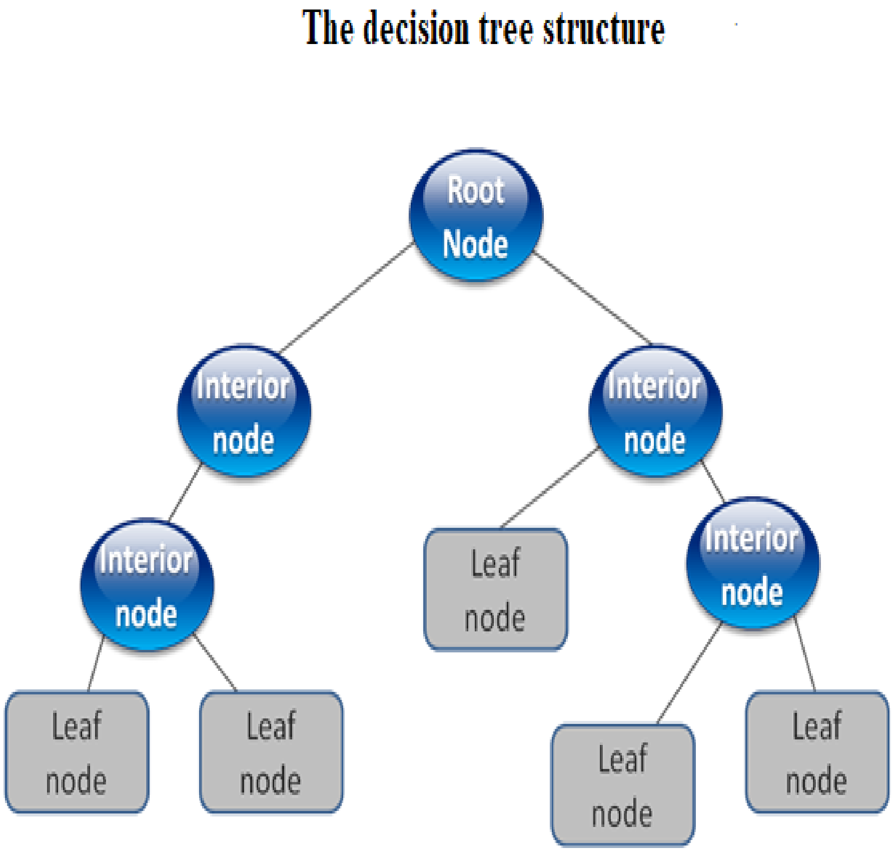

The decision tree structure

Figure (2.4), the decision tree structure, by (Researcher)

2.3 Summary

Medical data of the cancer disease and especially breast cancer has been used to build intelligent models using classification or prediction algorithms. Most of the reviewed literature has used data mining and machine learning techniques supervised or unsupervised.

Most of the literature review in our thesis concluded that using machine learning tools to predict breast cancer has a high accuracy of over 90%.

Most previous studies used not real data, but in our study, we will use real data of breast cancer patients from the Palestinian Ministry of Health, and we will use Machine learning techniques to predict breast cancer and discover the most related attributes to high prediction.

Chapter Three

Methodology

3.1 Introduction

In this study, we compared the performance of the machine learning algorithms (Support Vector Machines (SVM), Decision Tree, Artificial Neural Network (ANN), and Bayesian Based Classifiers) to predict breast cancer type (bening or malignant). Also, we determined the most important related attributes.

3.2 Study population:

The study population consists of patients who visit breast cancer clinics in the Central Health Directorates of the West Bank, who may have a benign tumor or malignant tumor.

3.3 Dataset:

The dataset is the breast cancer dataset that we obtained from the Central Health Directorates of the West Bank. The data set consists of the following independent variables (features) as shown in table 4-1:

Table (3.1): Attributes descriptions

Independent Variables (features)	**Description**
Marital Status	(0:Married , 1:Divorced, 2: 'Widowed', 3:Single , 4: Separated)
Hormonal therapy	0: No; 1: Yes
Age Group	(0: 20-30, 1: 31-40, 2: 41-50, 3:51-60, 4: more than 60)
Family History	Does family History have family members who had breast cancer? (1:Yes; 0:No)
Region	West Bank (North = 0, middle=1, south=2)
BMI Group	Body Mass Index (0:Under weight = <18.5, 1: Normal weight = 18.5–24.9,2: Overweight = 25–29.9 ,3:Obesity = BMI of 30 or greater)
Age Pregnancy Group	Age at first child birth(0: <30 ; 1: >=30)
Breastfeeding	Did you breastfeed your children? (0: No, 1: Yes)
Age at Menarche	The first menstruation (menarche) is the most definitive sign of puberty in females(0 : <12 ; 1:==13; 2:>=14)

Smoking Status)	0: Non-smoker; 1: Smoker currently; 2: Passive smoker
Drinking Alcohol	0: No; 1:Yes
Expose Radiation	expose of radiation-Chest area before the age of thirty (0: No, 1: Yes)
Physical Activity	do you do physical activity? (0: No, 1: Yes)
Education level	(0:Tawjihi or less , 1-Diploma,2- Bachelor(BA) , 3: Graduate Studies) 0: <= 12 , 1: >12 and <16 , 2:=16 , 3: >= 17
Type of locality	(Village= 0 ; City = 1 ; Refugee Camp = 2)
Mass shape	(0: Oval;1: Round; 2: Lobulated ; 3:Irregular
Mass Margin	(0: Indistinct ; 1:circumscribed ; 2: Micro lobulated; 3:Spiculated, 4: Obscured)
Career	Housewife = 0 ; Worker = 1 ;Employee = 2;Not Working=3 ;Farmer=4

<u>The Dependent Variable(Target Variable)</u>

- ✓ Class: 0: Benign tumor; 1: Malignant tumor

3.4 Procedure of Work-Study:

The strategy used to analyze the breast cancer data consist of the following steps.

- **Collecting data:** Data was collected from MOH and selecting the attributes variables (independent variables) and the target variable (dependent variable), then classifying the attributes into controlled and uncontrolled variables. The controlled attributes refer to those variables that can be controlled by humans, such as physical activity, smoking, weight, breastfed, number of children, while the uncontrolled attributes are those that cannot be controlled. They include exposure to radiation-Chest area, age at menarche, and age at menopause, a family history of breast cancer, or a family history of other genetic conditions.
- **Data preprocessing**: We used data mining to transform patient’s data into a suitable format, to be used by machine learning algorithms for predicting breast cancer, there are many important steps in data preprocessing. Data cleaning contains missing data analysis, outlier detection, range constraints, data-type constraints, Flirting and selection removed patients' records that all cells have null, and transformations, we have two variables each with zero (Age at first pregnancy and BMI), and these values indicate that these missing values because of the age of the lady at the birth of the first child is not equal to zero, and also for the other feature where it cannot be zero since it depends on two variables, namely height and weight. Based on the above we processed the missing values by using the following steps:

1-Replaced the zeros to Nan.

2-Replaced all missing values for nominal and numeric attributes in a dataset with the modes and means from the data, by using weka.filters replaced missing values.

One of the best practices for training a Neural Network is to normalize your data to achieve a mean close to 0. Normalizing the data usually accelerates learning and leads to faster convergence. (Timo Stöttner., 2019)

- **Splitng the Data**: the data set was divided into two subsets.
  - ✓ Training set (80%): a subset to train the models and estimate their parameters or thresholds.
  - ✓ Testing set (20%): a subset to test and evaluate models produced in the above step.

We have created four machine learning models. After that, we produced accuracy measures like sensitivity, specificity, accuracy, precision, and receiver operating characteristic (ROC) curves.

Accuracy is estimated as the fraction of the number of correct predictions to a total number of productions, as shown in the Equation.

$$\text{Accuracy} = \frac{\text{number of correct predictions}}{total\ number\ of\ productions}$$

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

Accuracy = $\frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$

The sensitivity, precision, specificity, F-Measure and RMS and RMSE can also be calculated in terms of positives and negatives as follows:

Sensitivity= Recall= $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$

Precision = $\frac{\mathrm{TP}}{\mathrm{TP+FP}}$.

Specificity = $\frac{\mathrm{TN}}{\mathrm{TN+FP}}$

F-Measure = $\frac{\mathrm{TP}}{\mathrm{TP}+\frac{1}{2}(\mathrm{FP+FN})}$

Where:

TP = True Positives, *TN* = True Negatives, *FP* = False Positives, and *FN* = False Negatives.

RMSE = $\sqrt{\sum_{i=1}^{n}\frac{(\hat{y_\imath}-\mathrm{y}_i)^2}{\mathrm{n}}}$

Where

$\widehat{y_1}, \widehat{y_2}, \widehat{y_3}, \ldots\ldots, \widehat{y_n}$ are predicted values

$\mathrm{y}_1, \mathrm{y}_2, \mathrm{y}_3, \ldots\ldots, \mathrm{y}_n$ are observed values

n is the number of observations

However the sensitivity and specificity are the most accepted ways to quantify the prediction accuracy.

The sensitivity, precision and specificity can also be calculated in terms of positives and negatives as follows:

Sensitivity measure indicates how is the ability of a test to correctly identify those with the disease.

Precision is ratio of total number of correctly classified positive cases and the total number of predicted positive cases.

Specificity measure indicates how is the ability of a test to correctly identify those who don't have the   disease.

F-measure is a measure of a test's accuracy. It is calculated from the precision and recall of the test.

Root Mean Squared Error (RMSE) is the square root of the mean of the squared differences between the actual value and predicted value.

The ROC curve (receiver of the operating characteristic curve) is a graph showing the efficiency of the classification model at all classification thresholds. This curve plots the following two parameters:

- True positive rate.
- False-positive rate.

An excellent model has AUC **(Area under the** ROC **Curve)** near to the 1, which means it has a perfect measure of separability. A poor model has AUC near to the 0 which means it has the worst measure of separability. It means it is reciprocating the result. It is predicting zero's as one's and one's as zeros. And when AUC is 0.5, it means the model has no class separation capacity whatsoever.

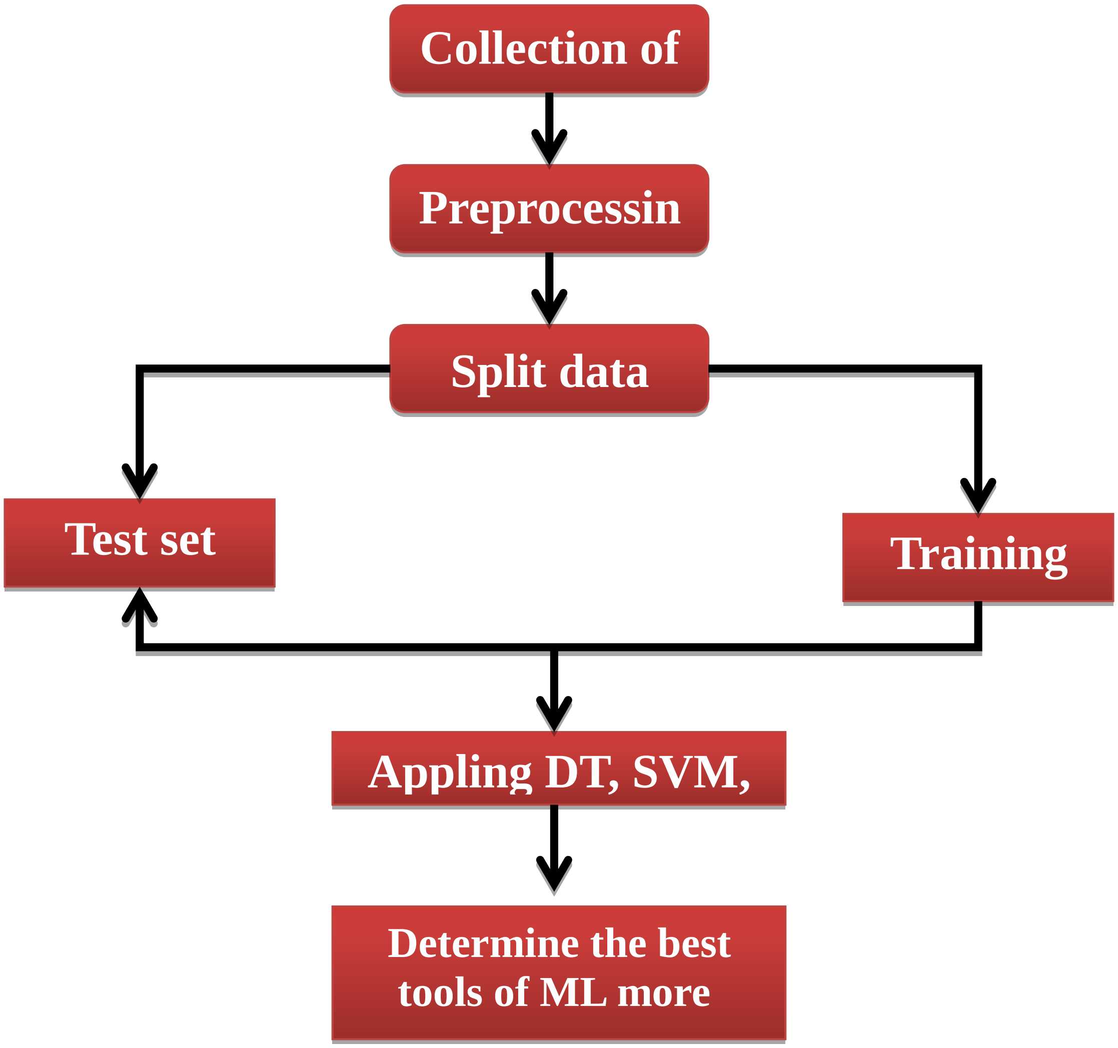


Figure (3.1): Procedure of Work-Study (by Researcher)

3.5 Data Analysis:

Data analysis was conducted using Weka software and R (programming language) to double-checks.

Weka (Waikato Environment for Knowledge Analysis) is a data mining platform that uses a series of machine learning algorithms written in Java, Weka is a free software application available under the GNU General Public License. The Weka workbench includes a set of visualization tools and algorithms for data analysis and predictive modeling, along with graphical user interfaces for quick access to this functionality. (Dr. Sudhir &Kodge,2013)

Weka is a collection of tools for:

- Data pre-processing
- Clustering
- Regression
- Classification

The researcher has discussed the prediction of breast cancer in Palestine, covering the period from April 2019 to October 2020.

Chapter four

Results and Discussion

4.1 Introduction

Our dataset is consisted of 1140 patients divided into two subsamples, one of which acted as a training set containing data from 912 patients, while the other subset acted as a validation set containing data from 228 patients.

4.2 Imbalanced Dataset:

The proportion of the malignant tumor in the data set is 29% while the proportion of benign tumor is 71%, as shown in figure 4.1. So our dataset is imbalanced, and it may give misleading accuracy and biased prediction since the algorithm in the imbalanced dataset doesn’t get the necessary information about the minority class to make an accurate prediction.

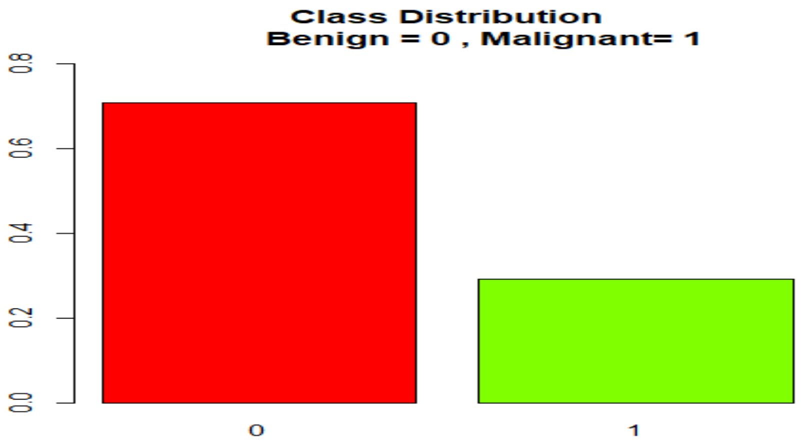


Figure 4.1, the proportions of class distribution of a dependent variable

To deal with the imbalanced dataset, we converted the dataset to nominal class and selected filters resample, we have used this filter as shown in figure 4.2, to get better performance results.

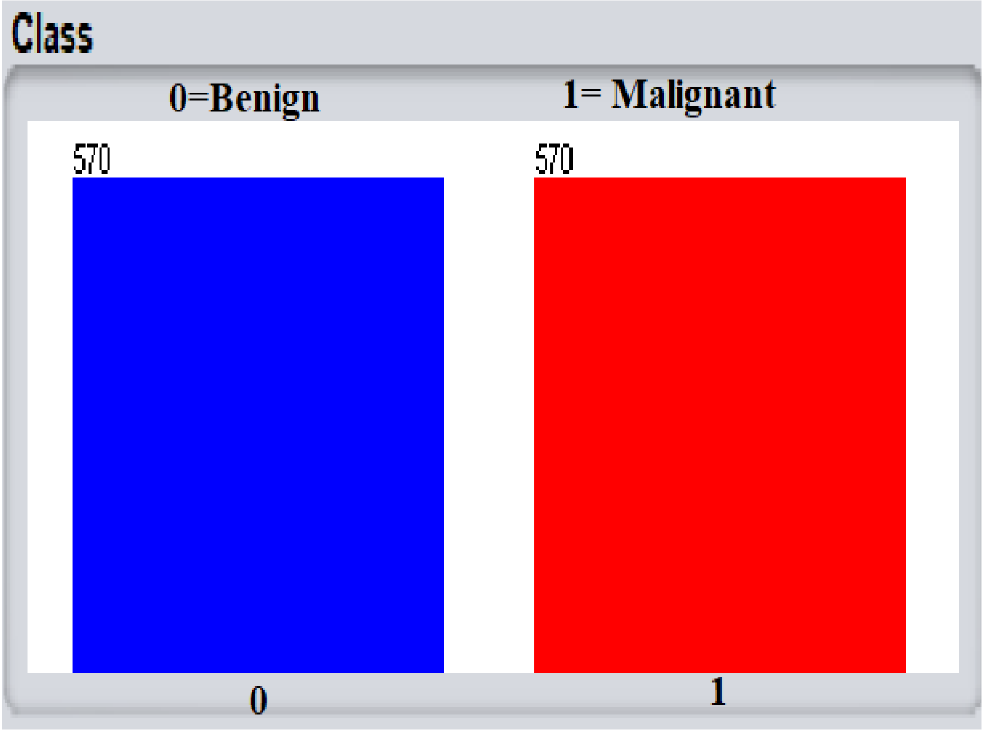


Figure 4.2 the proportions of class distribution of a balanced dependent variable**.**

4.3 **SVM Method**:

In the first algorithm SVM method, we have gotten 91.22% accuracy, 90.7% sensitivity (Recall), and 91.7% specificity, in medical researches, sensitivity and specificity are the statistical measures of performance of a binary classification test. In general, the sensitivity of the test indicates the ability of the model to correctly identify patients who have the disease, where the test specificity indicates the ability of the model to correctly identify patients who do not have the disease. The incorrect identity of patients with or without a disorder is related to the definition of type I and type II test hypothesis errors. The sensitivity of the test is, therefore, proportional to the strength of the test in the hypotheses test. (Sharma et al., 2009)

From the applied SVM algorithm, also we got a confusion matrix as shown in table 4.1

Table 4.1 the confusion matrix for the SVM Algorithm

n=228	Predicted (benign)	Predicted(malignancy)	Totals
Actual benign	110	10	120
Actual malignancy	10	98	108
Totals	120	108	228

Table 4.2 detailed accuracy for the SVM Algorithm

Sensitivity	Specificity	Precision	Accuracy	F-Measure	ROC Area	RMSE
0.907	0.917	0.907	0.912	0.907	0.912	0.2962

From table 4.1 we concluded, the SVM algorithm correctly classified 208 instances from 228 instances, and the model has correctly predicted 98 females as having the malignancy from 108 females who actually do have the malignancy tumor, while the model correctly predicted 110 females as having the benign tumor from 120 females who actually do have the benign tumor.

In other words, the model correctly identified 91.7% of the females who have benign tumor, but it wrongly predicted 8.3% of them as having malignancy tumor, on the other hand, the model correctly identified 90.7% of those who do have malignancy tumor, but it wrongly predicted 9.3% of them as having benign tumor.

4.3.1 The most related attributes that affect breast cancer by using SVM Algorithm

The results of the SVM algorithm showed that the most related attributes that affect breast cancer are:

- M-shape
- Hormonal therapy
- Family History
- M-margins
- Age pregnancy
- Education level

4.4 Decision Tree Method:

In the second algorithm, the decision tree method we have gotten 76.75% accuracy, 75.9 % sensitivity, and 77.5% specificity; also, we have a confusion matrix as shown in table 4.3

Table 4.3 the confusion matrix for Decision Tree Algorithm

n=228	Predicted (benign)	Predicted(malignancy)	Totals
Actual benign	93	27	120
Actual malignancy	26	82	108
Totals	119	109	228

Table 4.4 detailed accuracy for Decision tree

Sensitivity	Specificity	Precision	Accuracy	F-Measure	ROC Area	RMSE
0.759	0.775	0.752	.7675	0.756	0.821	0.407

From table 4.3 we concluded, the Decision tree algorithm correctly classified 175 instances from 228 instances, and the model has correctly predicted 82 females as having the malignancy from 108 females who actually do have the malignancy tumor, while the model correctly predicted 93 females as having the benign tumor from 120 females who actually do have the benign tumor.

In other words, the model correctly identified 77.5% the females who have benign tumor, but it wrongly predicted 22.5% of them as having malignancy tumor, on the other hand, the model correctly identified 75.9% of those who do have malignancy tumor, but it wrongly predicted 24.1% of them as having benign tumor.

As shown in figure 4.2 we see that the decision tree is drawn upside down with its root, at the top (mass shape) being the most significant element. The end of the branch that no longer separates is the decision/leaf, in this case, tumor malignant or tumor benign (0 or 1).

The decision tree is also a map of the possible outcomes of a number of similar choices. It allows a person or organization to weigh potential acts against each other on the basis of their costs, probabilities and benefits.

We will mention all possible outcomes resulting from the decision tree.

- If the patient has a mass shape irregular or lobulated (mshape=3 or 2), the model predicts that the patient has a malignant tumor.
- If the patient has a mass shape round (mshape=1), and she doesn't have family members who had breast cancer (family history=0), and she has mass margin indistinct or spiculated (mmargin=0 or 3) the model predicts that the patient has a malignant tumor.
- If the patient has a mass shape round (mshape=1), and she doesn't have family members who had breast cancer (family history=0), and she has mass margin circumscribed, Micro lobulated, Obscured) (mmargin=1 or 2 or 4), the model predicts that the patient has a benign tumor.
- If the patient has a mass shape round (mshape=1), and she has family members who had breast cancer (family history=1), and she drinks alcohol (alcohol consumption=1), the model predicts that the patient has a malignant tumor.

- If the patient has mass shape round (mshape=1), and she has family members who had breast cancer (family history=1), and she doesn't drink alcohol (alcohol consumption=0), and she is living in the north region (Jenin, Nablus, Qalqilia, Tubas, and Tulkarm) or she lives in the south of the West Bank(Bethlehem, or Hebron),(Region=0,2) the model predicts that the patient has the benign tumor.
- If the patient has mass shape round (mshape=1), and she has family members who had breast cancer (family history=1), and she doesn't drink alcohol (alcohol consumption=0), and she is living in the middle regions (Sulfate, Ramallah, Jericho, and East Jerusalem) (Region=1) the model predicts that the patient has a malignant tumor.
- If the patient has a mass shape oval (mshape=0), and she has hormonal therapy (hormonal therapy =1), the model predicts that the patient has a malignant tumor.
- If the patient has mass shape oval (mshape=0), and she wasn't treated for hormonal therapy (hormonal therapy =0), and the patient has mass margin Indistinct, circumscribed, Micro lobulated or Obscured),(mmargin=0,1,2,4) the model predicts that the patient has a benign tumor.
- If the patient has mass shape oval (mshape=0), and she wasn't treated for hormonal therapy (hormonal therapy =0), and the patient has mass margin Spiculated, (mmargin=3) the model predicts that the patient has a malignant tumor.

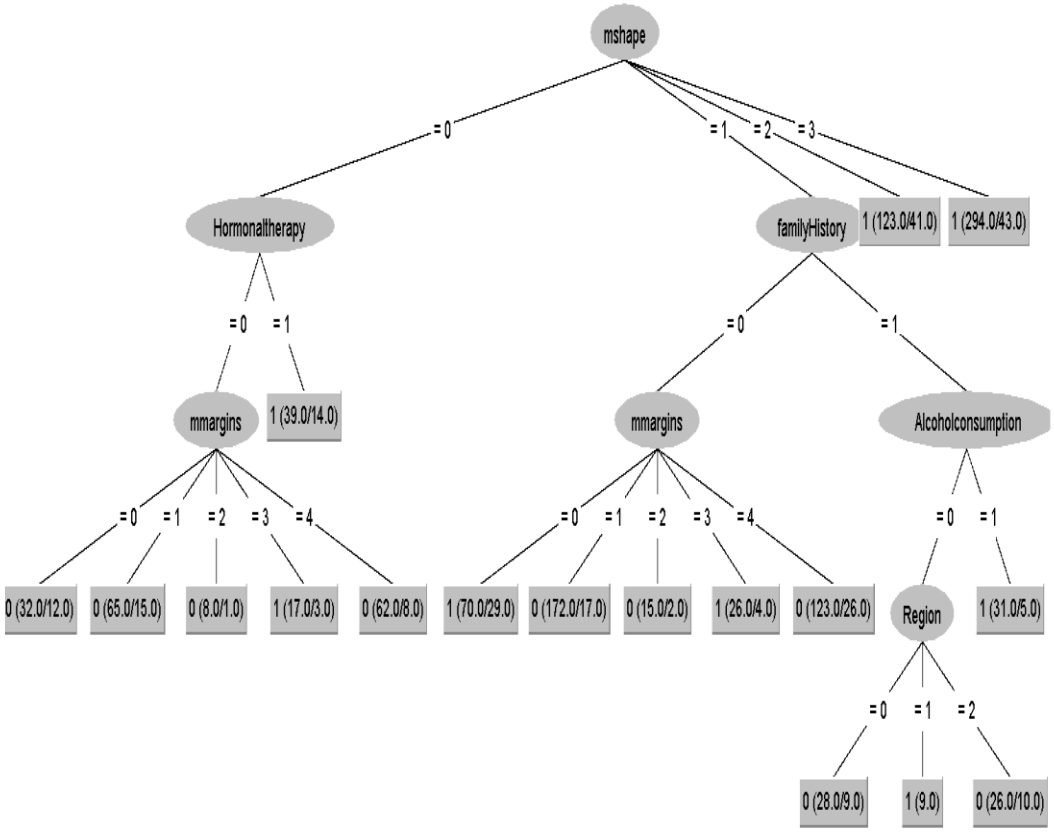


Figure 4.3, shows the Decision tree Classifier of our dataset.

4.4.1 The most related attributes that affect breast cancer

The results of the decision tree algorithm showed that the most related attributes that affect breast cancer are:

- M-shape
- M-margins
- Family's history

- Region
- Hormonal therapy
- Age group

This result is consistent with the report (Breast Cancer Facts&Figures 2019-2020) by the American Cancer Society) which concluded that the most related risk factors affecting breast cancer were hormones, alcohol consumption and not breastfeeding.

Also, study (Abdull, 2011) that concluded to the most related factor risk that affects breast cancer is family history.

However, our result dissimilar with the study (Abdul, 2011) that concluded to the most related factors risk that affects breast cancer is work-related to radiation, breastfeeding.

4.5 Naïve Base algorithm:

In the third algorithm Naïve Base method, we have gotten 77.2% accuracy, 74.1% sensitivity, and 80% specificity, a confusion matrix is shown in table 4.5 below.

Table 4.5 the confusion matrix for the Naïve Base Algorithm

n=228	Predicted (benign)	Predicted(malignancy)	Totals
Actual benign	96	24	120
Actual malignancy	28	80	108
Totals	114	104	228

Table 4.6 detailed accuracy for Naïve Base

Sensitivity	Specificity	Precision	Accuracy	F-Measure	ROC Area	RMSE
0.741	0.80	0.769	0.772	0.755	0.856	0.394

From table 4.5 we concluded, the Naïve Base algorithm correctly classified 176 instances from 228 instances, and the model has correctly predicted 80 females as having the malignancy from 108 females who actually do have the malignancy tumor, while the model correctly predicted 96 females as having the benign tumor from 120 females who actually do have the benign tumor.

In other words, the model correctly identified 80% of the females who have benign tumor, but it wrongly predicted 20%, of them as having malignancy tumor, on the other hand, the model correctly identified 74.1% of those who do have malignancy tumor, but it wrongly predicted 25.9% of them as having benign tumor.

4.5.1 The most related attributes that affect breast cancer

The results of the Naïve Base algorithm showed that the most related attributes that affect breast cancer are:

- Mass shape
- Hormonaltherapy
- Family's history
- Smoking

- Exposeradiation
- Age group

4.6 Neural Network (Multilayer Perceptron):

The last algorithm used a neural network, we executed the Multilayer Perceptron and applied normalized by using WEKA, and we got 84.6% accuracy, 88% sensitivity, and 81.7% specificity, a confusion matrix is shown in table 4.7 below.

Table 4.7 the confusion matrix for the Neural Network Algorithm

n=228	Predicted (benign)	Predicted(malignancy)	Totals
Actual benign	98	22	120
Actual malignancy	13	95	108
Totals	111	117	228

Table 4.8 detailed accuracy by class for Neural Network

Sensitivity	Specificity	Precision	Accuracy	F-Measure	ROC Area	RMSE
0.88	0.817	0.812	0.846	0.844	0.903	0.351

From table 4.7 we concluded, the Multilayer Perceptron algorithm correctly classified 193 instances from 228 instances, and the model has correctly predicted 95 females as having the malignancy from 108 females who actually do have the malignancy tumor, while the model correctly predicted 98 females as having the benign tumor from 120 females who actually do have the benign tumor.

In other words, the model correctly identified 81.7% of the females who have benign tumor, but it wrongly predicted 18.3% of them as having malignancy tumor, on the other hand, the model correctly identified 88% of those who do have malignancy tumor, but it wrongly predicted 12% of them as having benign tumor.

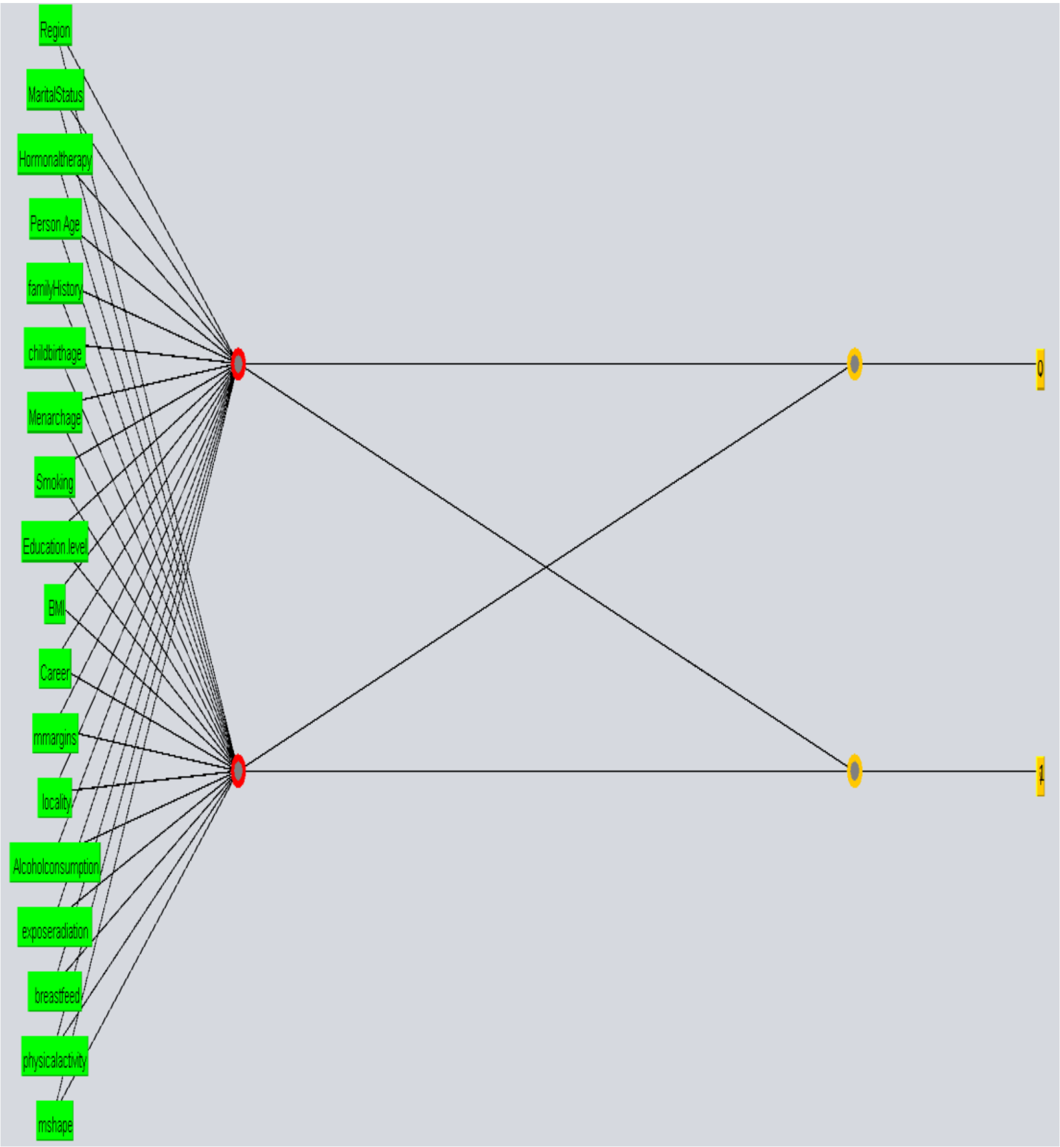


Figure 4.4, shows the ANN architecture has two layers.

From figure 4.3 we see that's ANN consists of three parts, an input layer, several hidden layers, and an output layer. The layers are usually arranged from the left (input) to the right (output).

4.6.1 The most related attributes that affect breast cancer

The results of the ANN algorithm showed that the most related attributes that affect breast cancer are:

- M-shape
- Region
- breastfeed
- Hormonaltherapy
- familyHistory
- Smoking

4.7 Comparing (SVM, DT, NB, and ANN)

4.7.1 Experiment Results:

The following table compares accuracy, sensitivity, and specificity between (SVM, DT, NB, and ANN).

Table 4.9 Comparing accuracy, sensitivity, and specificity between (SVM, DT, NB, and ANN)

Algorithms	SVM	Decision tree	Naive base	Neural network
Accuracy	91.23%	76.75%	77.2%	84.64%
sensitivity	90.7%	75.9%	74.1%	88%
specificity	91.7%	77.5%	80%	81.7%
Precision	90.7%	75.2%	76.9%	81.2%

From table 4.9 we have found that the SVM algorithm has the highest value of 91.23.6% accuracy, and 90.7% sensitivity. While the NB algorithm got the lowest value of sensitivity (sensitivity = 0.741).

As mentioned earlier, our sample are 1140 records, so SVM has more performance than other algorithms that used in our research because of the Support Vector Machines (SVM) works fine when we have a small data set, it does not require much time to train model.

Table 4.10 comparing the most related attributes that affect breast cancer between (SVM, DT, NB, and ANN)

Ranking of important attributes	SVM	Decision tree	Naive base	Neural network
First	M-shape	M-shape	M-shape	M-shape
Second	Hormonal therapy	M-margins	Hormonal therapy	Region
Third	Family History	Family History	Family History	Breastfeed
Fourth	M-margins	Region	Smoking	Hormonal therapy
Fifth	Age pregnancy	Hormonal therapy	Expose radiation	Family History
Sixth	Education level	Age-group	Age-group	Smoking

From table 4.10 we used weight vector which is of the size of the number of features. We used this weight vector to select the 6 most important features by just selecting the 6 features which the highest weights as shown in the appendix page 63.

Chapter Five

Conclusions and Recommendations.

5.1 Introduction

This chapter focuses mainly on the conclusions, recommendations, and future work.

5.2 Conclusions of this research.

The study has concluded a set of results, here are the most important of them.

1. We found that the Support Vector Machines (SVM) was the best predictor of breast cancer in our model compared with the other three mentioned classifiers (SVM, DT, and NB).

- We found the most related attributes that affect breast M-shape, Hormonal therapy, Family's history, M-margins, Age pregnancy and Education level.
- Support Vector Machines (SVM) achieved the highest average accuracy of 91.23, and 90.7% sensitivity.
- Support Vector Machines (SVM) works fine when we have a small data set because it does not require much time to train the model.
- Our study is consistent with the study of (Hiba et al., 2016) where the study showed that Support Vector Machines (SVM) have proven their efficiency in Breast Cancer prediction and diagnosis and achieves the best performance. Also, the study of (Tolga Ensari, 2019) concluded that the best performance

has been obtained by the Support Vector Machine technique with the highest accuracy.

5.2 Recommendations:

In light of the results of this study, a set of recommendations are required for this study to achieve its objectives in a correct manner, which are as follows: -

- ❖ In order to increase the accuracy of the results in our research, the author recommends studying other variables such as: number of children, type of meat consumed, age at last pregnancy, type of vegetables consumed.
- ❖ Save data set in a secured way on the Palestinian Ministry of Health website after applying privacy measures.
- ❖ Adding other variables related to society, health, geography and nutritional aspects in order to have a better understanding of the disease.
- ❖ Utilize machine learning algorithms to predict the disease.
- ❖ Spread awareness among females on what to avoid such as: smoking, alcohol consumption and contraception use.

5.3 Limitations

This study has several limitations, for example, three factors were excluded from the study:

- BRACA1 and BRACA2, since it's a very expensive test and is not found in the records of the Palestinian Ministry of Health.
- Electronic data included are breast cancer patients’ information from the period 2019 to September of 2020.
- The Source of data is the directorates of health in the West Bank only.

References

- Akinsola, J., Awodele, O., Hinmikaiye, J., Olakanmi, O & Akinjobi, J. (2017). *Supervised Machine Learning Algorithms: Classification and Comparison*. Paper presented at the International Journal of Computer Trends and Technology (IJCTT) – Volume 48 Number 3 June 2017 ISSN: 2231-2803 http://www.ijcttjournal.org Page 128.
- Ali Bataineh, A.(2019) *A Comparative Analysis of Nonlinear Machine LearningAlgorithms for Breast Cancer Detection.* Paper presented at the International Journal of Machine Learning and Computing, Vol. 9, No. 3, June 2019.
- Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). *Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis*. Procedia Computer Science 83 (2016) 1064 – 1069**.**
- Awad, M., & Khanna, R. (2018). *Using Machine Learning algorithms for breast cancer risk prediction and diagnosis.* Papers publication at https://www.researchgate.net/publication/300723807
- Bharat, A., Pooja, N., & Reddy, A. (2018).*Using Machine Learning algorithms for breast cancer risk prediction and diagnosis.* Paper presented at the IEEE Third International Conference on Circuits, Control, Communication, and Computing.

- Borkar, M., Deulkar., K., & Garg, A. (2015). *Prediction of Breast Cancer using Artificial Neural Networks.* Paper presented at the International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181.
- Deepika, M., Gladence, L., & Keerthana, R. (2016). *A review on prediction of breast cancer using various data mining techniques.* ISSN: 0975-8585.
- Dhahri, H., Maghayreh., E., Mahmood., A., & Elkilani, W. (2019). *Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms.* Paper presented at the Hindawi Journal of Healthcare Engineering Volume 2019, Article ID 4253641, 11 pages https://doi.org/10.1155/2019/4253641
- El-Shahat, A. (2018). *Introductory Chapter: Artificial Neural Networks.* Paper presented at the

  http://dx.doi.org/10.5772/intechopen.73530
- Jagtap, S.,& Kodge, B. (2013). *Census Data Mining and Data Analysis using WEKA.* . Paper presented at the International Conference in “Emerging Trends in Science, Technology and Management-2013, Singapore.
- KELEŞ, M.(2019). *Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study*. Paper presented at the ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
- Kourou, K., Themis., E., Konstantinos, E., Michalis, E., & Dimitrios, F. (2015*).* *Machine learning applications in cancer prognosis and prediction*.

Papepresented at Computational and Structural Biotechnology Journal Volume 13, 2015, Pages 8-17

- Preetha, R., and Jinny, S. (2019). *AResearch on Breast Cancer Prediction using Data Mining Techniques*. Paper presented at the International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-11S2, September 2019.
- Rathi, M. (2012). A *Breast Cancer Prediction using Naïve Bayes Classifier*. Paper presented at the, I International Journal of Information Technology & Systems, Vol. 1; No. 2: ISSN: 2277-9825 (July-Dec. 2012).
- Salem, A. M. The *Data mining techniques and breast cancer prediction: A case study of Libya*. Doctor thesis. College of social studies, The Sheffield Hallam University- Sheffield-U.K,2011.
- Saritas, I. (2011). *Prediction of Breast Cancer Using Artificial Neural Networks.* Published online: 12 August 2011, Springer Science+Business Media, LLC 2011.
- Sharma, D., Yadav., U., & Sharma, P. (2009). *The concept of sensitivity and specificity in relation to two types of errors and its application medical research*. Paper presented at the Journal of Reliability and Statistical Studies (ISSN: 0974-8024) Vol. 2, Issue 2(2009): 53-58.

- Stark, G., Gregory, H., Nartowt, B. & Deng, J. (2018). *Predicting breast cancer risk using personal health data and machine learning models.* Paper presented at the PLoS ONE 14(12): e0226765. https://doi.org/10.1371/journal.pone.0226765
- Stat, Mengjie, *Breast Cancer Prediction Using Machine Learning Algorithm*, the University of Texas at Austin, 2017.16.
- Sumbaly, R., Vishnusri., N., & Jeyalatha, S. (2014*). Diagnosis of Breast Cancer using Decision Tree Data Mining Technique*. Paper presented at the International Journal of Computer Applications (0975 – 8887)Volume 98– No.10, July 2014.
- Sudhir, J., & Kodge, G. (2013*). Census Data Mining and Data Analysis using WEKA*. Paper presented at the (ICETSTM – 2013) International Conference in "Emerging Trends in Science, Technology and Management-2013, SingaporeTolga, E.(2019). Comparison of Machine Learning Methods for. Breast Cancer Diagnosis. Paper presented at the DOI: 10.1109/EBBT.2019.8741990.
- Tafish, M.H, Breast Cancer Severity Degree Prediction Using Data Mining Techniques in the Gaza Strip, the Islamic University–Gaza /2017.
- American Cancer Society(2020). *Breast Cancer Facts & Figures 2019-2020*. (https://www.cancer.org). Retrieved-from https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2019-2020.pdf

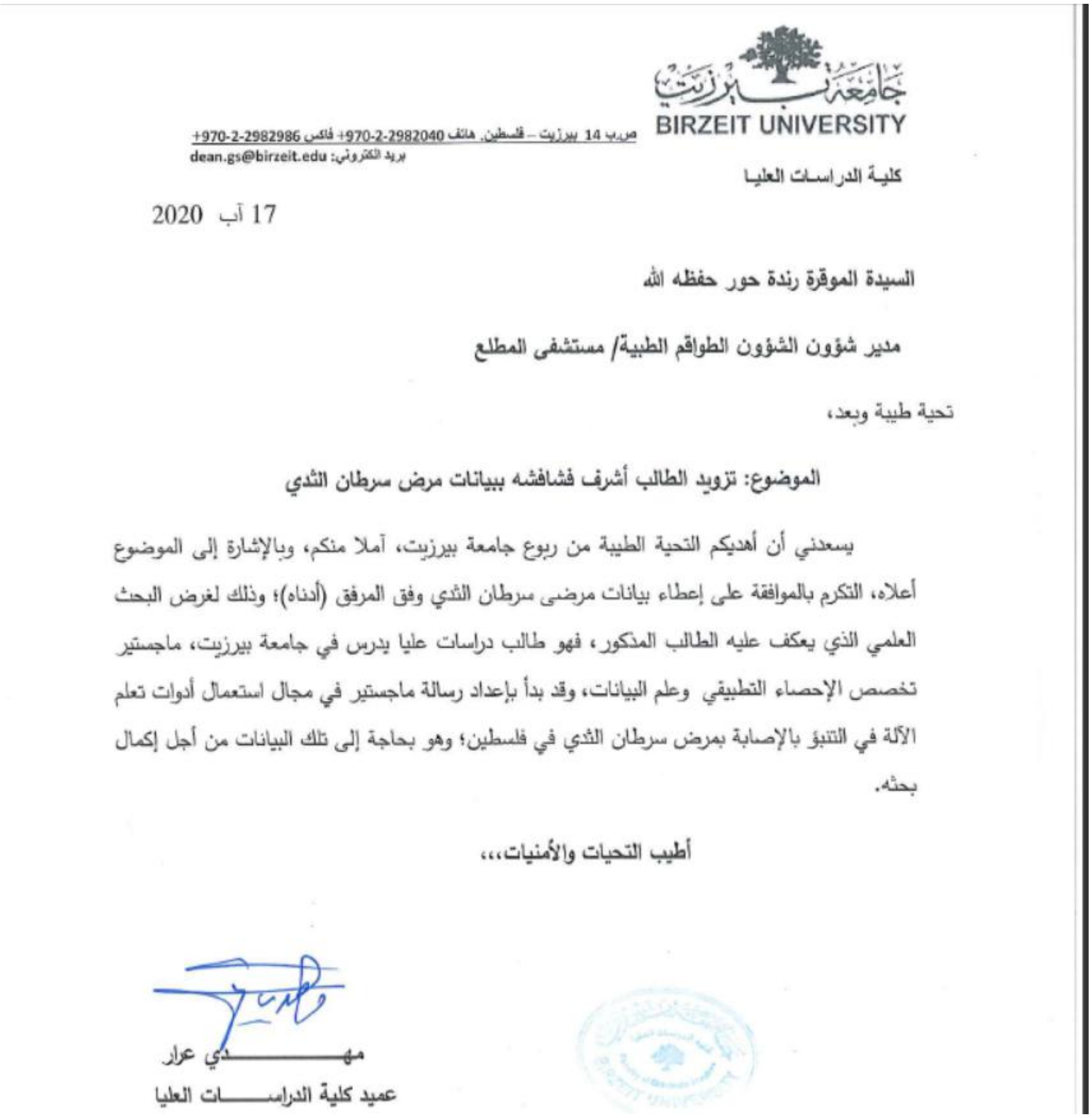
- Stöttner,Timo (May 16,2019). *Why Data should be Normalized before Training a Neural Network*.(www. https://towardsdatascience.com). Retrieved-from https://towardsdatascience.com/why-data-should-be-normalized-before-training-a-neural-network-c626b7f66c7d
- The World Health Organization. (2018) .*Cancer*.Retrieved: June from https://www.who.int/health-topics/cancer#tab=tab_1
- Palestinian Ministry of Health. (2020). Annual Health Report Palestine 2019, Retrieved-from http://site.moh.ps/Content/Books/HYM2UGrm8hFDOPe1AW6z2W6ZDvbJbuYGykdfV6B1lEulthrx5QMAyC_5WFKDTWWGKW3O7rk4vgIUzRlhJdSYyQXxFKscP6Uqz3UhrxoWLcHlT.pdf
- Palestinian Ministry of Health. (2020). Annual Health Report Palestine 2016, Retrieved from

  http://site.moh.ps/Content/Books/ZxRcynmiUofNqt66u4CrHRgmJR6Uv7z77srjjIEAho6xnz5V3rgLTu_RhO7xf2j2VusNiIvWkjwp84yXHLdGleB97gKrHHI5iZ9oPJ25owGEN.pdf.
- Saudi Ministry of Health. (2020) .Breast Cancer,

  Retrieved from https://www.moh.gov.sa/en/awarenessplateform/ChronicDisease/Pages/BreastCancer.aspx

Appendix (A):- The Books to provide Data

The Books to provide Data



جامعة بيرزيت
BIRZEIT UNIVERSITY

ص.ب 14 بيرزيت – فلسطين. هاتف 2982040-2-970+ فاكس 2982986-2-970+
بريد الكتروني: dean.gs@birzeit.edu

كليـة الدراسـات العليـا

17 آب 2020

السيدة الموقرة رندة حور حفظه الله

مدير شؤون الشؤون الطواقم الطبية/ مستشفى المطلع

تحية طيبة وبعد،

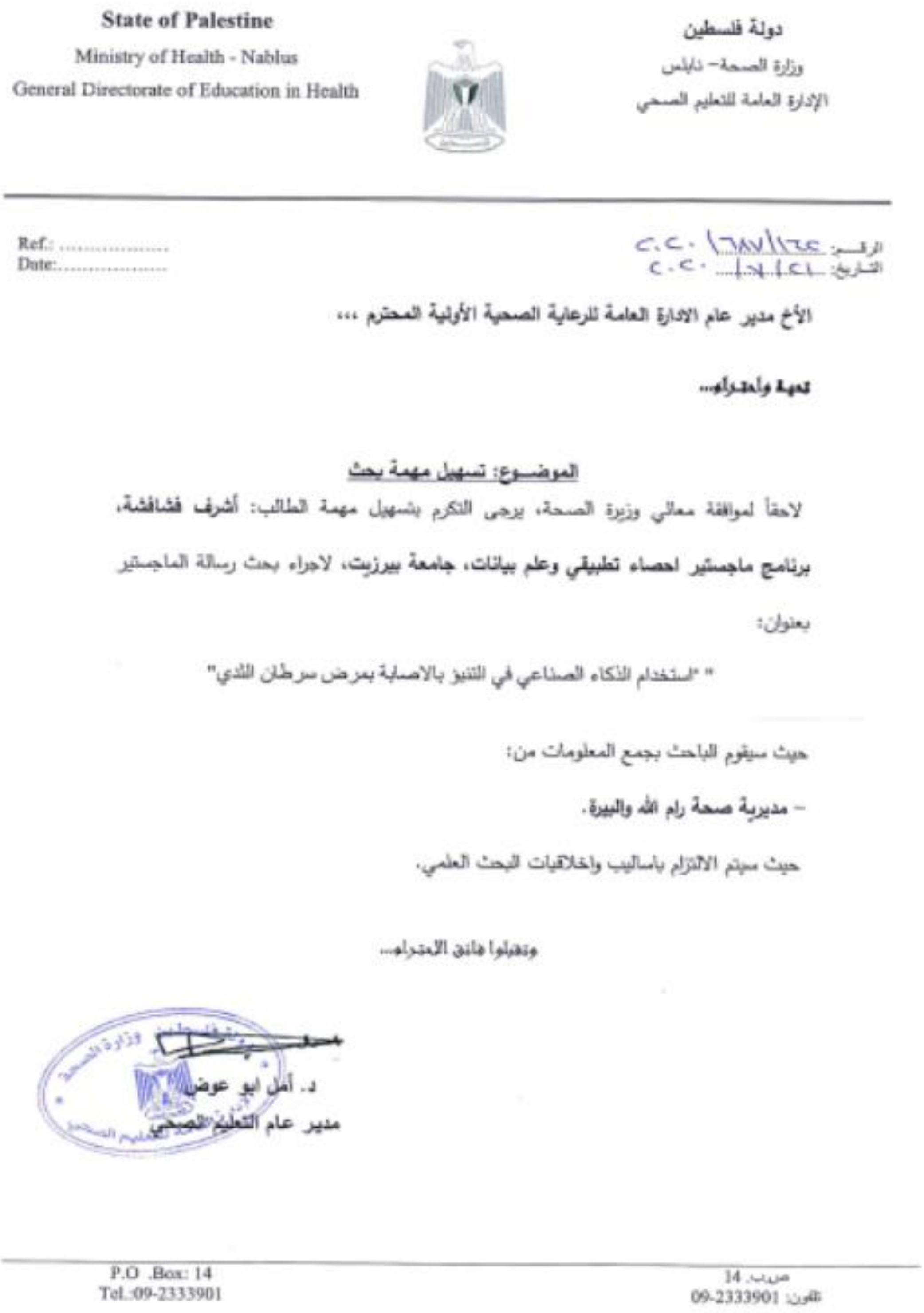
الموضوع: تزويد الطالب أشرف فشافشه ببيانات مرض سرطان الثدي

يسعدني أن أهديكم التحية الطيبة من ربوع جامعة بيرزيت، آملا منكم، وبالإشارة إلى الموضوع أعلاه، التكرم بالموافقة على إعطاء بيانات مرضى سرطان الثدي وفق المرفق (أدناه)؛ وذلك لغرض البحث العلمي الذي يعكف عليه الطالب المذكور، فهو طالب دراسات عليا يدرس في جامعة بيرزيت، ماجستير تخصص الإحصاء التطبيقي وعلم البيانات، وقد بدأ بإعداد رسالة ماجستير في مجال استعمال أدوات تعلم الآلة في التنبؤ بالإصابة بمرض سرطان الثدي في فلسطين؛ وهو بحاجة إلى تلك البيانات من أجل إكمال بحثه.

أطيب التحيات والأمنيات،،،

مهـــــــــــــدي عرار
عميد كلية الدراســـــات العليا

BIRZEIT UNIVERSITY



State of Palestine
Ministry of Health - Nablus
General Directorate of Education in Health

دولة فلسطين
وزارة الصحة– نابلس
الإدارة العامة للتعليم الصحي

Ref.: ....
Date:..................

الرقم: ٢٠٢٠/١٦٨٧/١٢٤
التاريخ: ٢٠٢٠/٧/٢١

الأخ مدير عام الادارة العامة للرعاية الصحية الأولية المحترم ،،،

تحية واحترام...

<u>الموضــوع: تسهيل مهمة بحث</u>

لاحقاً لموافقة معالي وزيرة الصحة، يرجى التكرم بتسهيل مهمة الطالب: **أشرف فشافشة، برنامج ماجستير احصاء تطبيقي وعلم بيانات، جامعة بيرزيت،** لاجراء بحث رسالة الماجستير بعنوان:

" "استخدام الذكاء الصناعي في التنبؤ بالاصابة بمرض سرطان الثدي"

حيث سيقوم الباحث بجمع المعلومات من:

– مديرية صحة رام الله والبيرة.

حيث سيتم الالتزام باساليب واخلاقيات البحث العلمي.

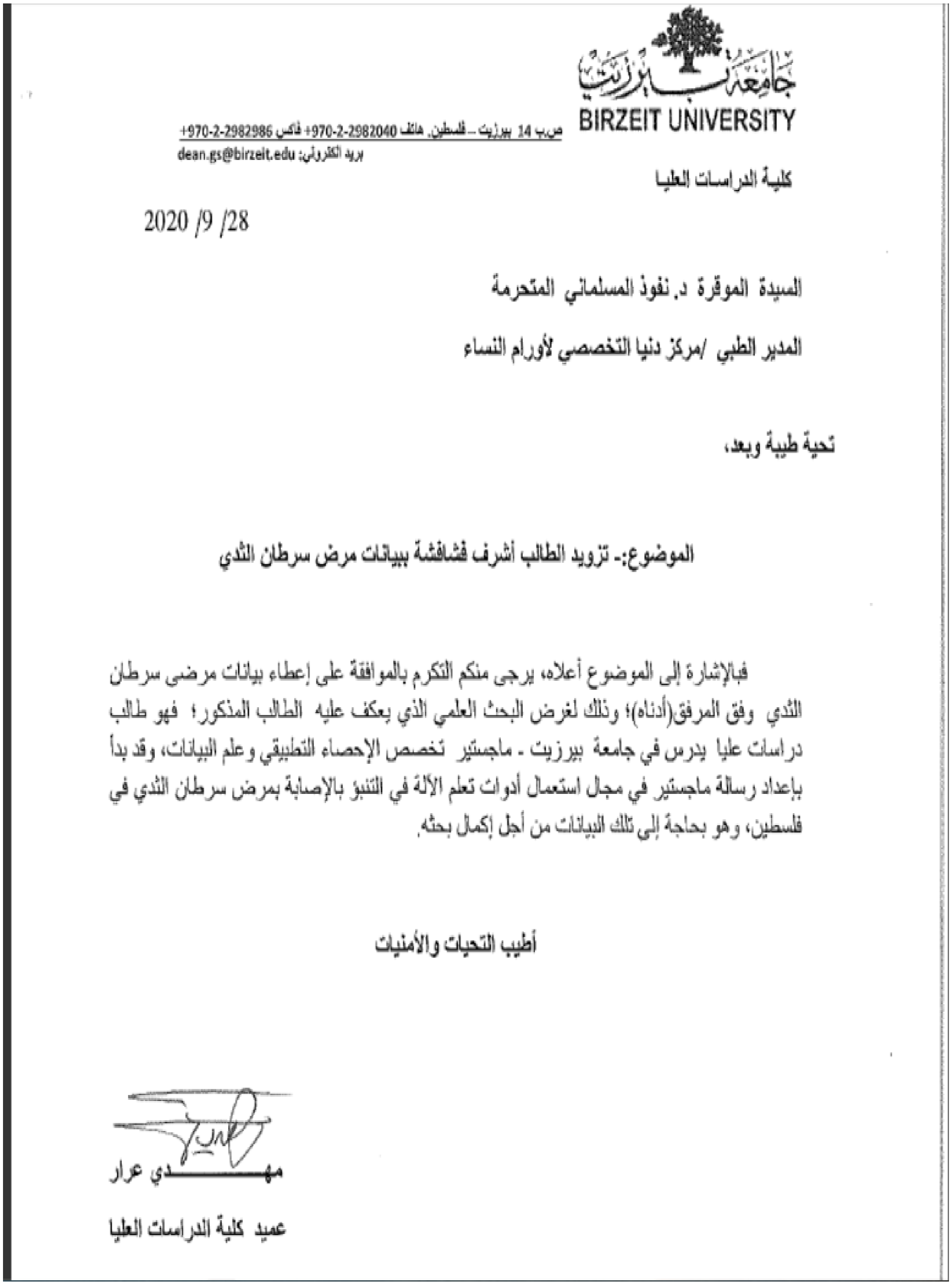
وتقبلوا فائق الاحترام...

د. أمل ابو عوض
مدير عام التعليم الصحي

وزارة الصحة

P.O .Box: 14
Tel.:09-2333901

ص.ب. 14
هاتف: 09-2333901



جامعة بيرزيت
BIRZEIT UNIVERSITY

ص.ب 14 بيرزيت – فلسطين، هاتف 2982040-2-970+ فاكس 2982986-2-970+
بريد الكتروني: dean.gs@birzeit.edu

كليـة الدراسـات العليـا

2020 /9 /28

السيدة الموقرة د. نفوذ المسلماني المتحرمة

المدير الطبي /مركز دنيا التخصصي لأورام النساء

تحية طيبة وبعد،

الموضوع:- تزويد الطالب أشرف قشاقشة ببيانات مرض سرطان الثدي

فبالإشارة إلى الموضوع أعلاه، يرجى منكم التكرم بالموافقة على إعطاء بيانات مرضى سرطان الثدي وفق المرفق(أدناه)؛ وذلك لغرض البحث العلمي الذي يعكف عليه الطالب المذكور؛ فهو طالب دراسات عليا يدرس في جامعة بيرزيت - ماجستير تخصص الإحصاء التطبيقي وعلم البيانات، وقد بدأ بإعداد رسالة ماجستير في مجال استعمال أدوات تعلم الآلة في التنبؤ بالإصابة بمرض سرطان الثدي في فلسطين، وهو بحاجة إلى تلك البيانات من أجل إكمال بحثه.

أطيب التحيات والأمنيات

مهـــــــــــدي عرار

عميد كلية الدراسات العليا

Appendix (B): Result of WEKA and R

Result of WEKA and R

((Split Data 80% for training model and 20% for testing or evaluation model))

SVM Algorithm

SVM type:-SVM classification

Epsilon =0.001

Gamma =0.125

Where

Epsilon: The tolerance of the termination criterion.

Gamma: - The gamma to use, if 0 then 1/max_index is used.

Instances: 1140

Attributes: 19

Region

Marital Status

Hormonal therapy

Person Age

Family History

Child birth age

Menarche age

Smoking

```
Education level

BMI

Career

M margins

Locality

Alcohol consumption

Expose radiation

Breastfeed

Physical activity

M shape

Class

Test mode:    split 80.0% train, remainder test

Time taken to build model: 0.11 seconds

Evaluation on test split

Time taken to test model on test split: 0.02 seconds
```

❖ **Summary**

Correctly Classified Instances	208	19.22%
Incorrectly Classified Instances	20	8.77%
Kappa statistic		0.8241
Mean absolute error		0.0877
Root mean squared error		0.2962
Relative absolute error		17.53%
Root relative squared error		59.18%
Total Number of Instances		228

Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.917	0.093	0.917	0.917	0.917	0.824	0.912	0.884	0
0.907	0.083	0.907	0.907	0.907	0.824	0.912	0.867	1

Confusion Matrix for the SVM Algorithm

n=228	Predicted (No)	Predicted (Yes)	Totals
Actual No	110	10	120
Actual Yes	10	98	108
Totals	120	108	228

❖ **Decision Tree (trees.J48) Algorithm**

Run information

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 25

Relation: Book11111-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last-weka.filters.supervised.instance.Resample-B1.0-S1-Z100.0

Instances: 1140

Attributes: 19

Test mode: split 80.0% train, remainder test

Time is taken to build the model: 0 seconds

Evaluation on test split

Time is taken to test the model on test split: 0 seconds

Summary

Correctly Classified Instances	175	76.75%
Incorrectly Classified Instances	53	23.24%
Kappa statistic		0.534
Mean absolute error		0.3168
Root mean squared error		0.4069
Relative absolute error		63.32%
Root relative squared error		81.31%
Total Number of Instances		228

Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.775	0.241	0.782	0.775	0.782	0.53	0.821	0.799	0
0.759	0.225	0.752	0.759	0.756	0.534	0.821	0.781	1

Confusion Matrix

n=228	Predicted (No)	Predicted (Yes)	Totals
Actual No	93	27	120
Actual Yes	26	82	108
Totals	119	109	228

❖ Naive Bayes Classifier

Run information

Scheme: weka.classifiers.bayes.NaiveBayes

Relation: Book11111-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last-weka.filters.supervised.instance.Resample-B1.0-S1-Z100.0

Instances: 1140

Attributes: 19

Test mode: split 80.0% train, remainder test

Time taken to build model: 0.01 seconds

Evaluation on test split

Time taken to test model on test split: 0 seconds

Summary

Correctly Classified Instances	176	77.19%
Incorrectly Classified Instances	52	22.80%
Kappa statistic		0.5417
Mean absolute error		0.2932
Root mean squared error		0.3938
Relative absolute error		58.59%
Root relative squared error		78.70%
Total Number of Instances		228

Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.80	0.259	0.774	0.80	0.787	0.542	0.856	0.867	0
0.741	0.200	0.769	0.741	0.755	0.542	0.856	0.848	1

Matrix Confusion Matrix

n=228	Predicted (No)	Predicted (Yes)	Totals
Actual No	96	24	120
Actual Yes	28	80	108
Totals	124	104	228

- **Multilayer Perceptron (Neutral Network)**

Scheme: weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H 2

Relation: Book11111-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last-weka.filters.supervised.instance.Resample-B1.0-S1-Z100.0

Instances: 1140

Attributes: 19

Test mode: split 80.0% train, remainder test

Time taken to build model: 1.94 seconds

Evaluation on test split

Time taken to test model on test split: 0 seconds

Summary

Correctly Classified Instances	193	84.6491%
Incorrectly Classified Instances	35	15.3509%
Kappa statistic	0.6934	
Mean absolute error	0.1979	
Root mean squared error	0.3508	
Relative absolute error	39.56%	
Root relative squared error	70.11%	
Total Number of Instances	228	

Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.817	0.120	0.883	0.817	0.848	0.696	0.903	0.914	0
0.88	0.183	0. 812	0.88	0.844	0.696	0.903	0.899	1

Confusion Matrix

n=228	Predicted (No)	Predicted (Yes)	Totals
Actual No	98	22	120
Actual Yes	13	95	108
Totals	111	117	228

Output of R(determine The most related attributes that affect breast cancer by using four algorithms)

```
print(w,cat(' The most related attributes that affect breast cancer by using SVM
algorithm:\n'))
 The most related attributes that affect breast cancer by using SVM algorithm:
            mshape     Hormonaltherapy      familyHistory          mmargins
age.pregnancy.group     Education.level            Career
         24.141318          19.633282          15.571395          11.122489
9.289583           8.488007           7.621107
            Region           agegroup     exposeradiation        breastfeed    physicalactivity
locality           BMIgroup
          7.592410           7.297156           6.556821           6.368542           5.623189
5.440871           5.310359
 Alcoholconsumption       MaritalStatus            Smoking   age.Menarch.group
          4.598907           2.366250           1.700077           1.120479
> print(vi_tree,cat('The most related attributes that affect breast cancer by using
Decision tree algorithm:\n'))
The most related attributes that affect breast cancer by using Decision tree algorithm:
            mshape           mmargins       familyHistory             Region
Hormonaltherapy            agegroup            BMIgroup
       78.18021136        37.44229544        22.38341872        14.18814557
11.28966828        10.34118126        9.96822780
 Alcoholconsumption              Career            Smoking       MaritalStatus
exposeradiation age.pregnancy.group     Education.level
        5.03209549         4.71163027         2.60966173         2.27894753
2.15556009          1.89664175          1.25834900
          locality
        0.05547895
> #####How to determine the most related attributes that affect breast cancer
> varImp(naive_bayes,cat('The most related attributes that affect breast cancer by
using Naive Bayes algorithm:\n'))
loess r-squared variable importance

                    Overall
mshape            100.00000
Hormonaltherapy     28.77455
familyHistory       27.18641
Smoking             17.17703
exposeradiation     13.96465
agegroup            12.75499
```

```
mmargins            7.51288
MaritalStatus       6.85815
breastfeed          4.74228
Education.level     3.99890
Alcoholconsumption  1.41417
BMIgroup            0.96806
age.pregnancy.group 0.75848
age.Menarch.group   0.45052
locality            0.20575
physicalactivity    0.04735
Region              0.02975
Career              0.00000
```

```
> imp<-varImp(model)###How to determine the most related attributes that affect breast cancer
```

The most related attributes that affect breast cancer by using Neural network algorithm:

```
                    Overall
mshape              100.000
Region               98.509
breastfeed           71.708
Hormonaltherapy      68.579
familyHistory        58.645
Smoking              50.816
exposeradiation      39.972
age.pregnancy.group  38.114
physicalactivity     30.399
BMIgroup             30.102
Career               27.486
locality             26.304
MaritalStatus        22.612
Education.level      21.609
mmargins             17.482
age.Menarch.group    17.446
Alcoholconsumption    8.644
agegroup              0.000
```

Appendix(C) R-codes

R-codes

```
#################Breast Cancer Classification Problem#############
# Loaded needed library
library(ggplot2)
library(caret)
library(dplyr)
library(purrr)
library(tidyverse)
library(car)
library(ggplot2)
library(readr)
library(caret)
library(e1071)
library(rpart)
library(rpart.plot)
library(ROCR)
library(GGally)
library(datasets)
library(haven)
library(olsrr)
library(dplyr)
library(caTools)
library(mice)
library(mctest)
library(corrplot)
library(car)
library(tree)
library(textir)
library(naivebayes)
library(ROCR)
library(VIM)
library(class)
library(neuralnet)
library(reprex)
#Clean cache
rm(list=ls())
library(readxl)
#Load datasets

data <-read_excel("C:/Users/eng_a/OneDrive/Desktop/dataset_brestcancer.xls")
view(data)
attach(data)
summary(data)
#Creating subset to treat missing values
dataset=select(data,childbirthage,BMI)

dataset[dataset=='0']=NA#Replacing 0 with NA
summary(dataset)#This will show the NA present in every individual variable
sum(is.na(dataset))#To check total number is NA present
#Replacing the treated variables with the untreated variables present in main data

data$childbirthage=dataset$childbirthage
data$BMI=dataset$BMI
```

```
View(data)


####################convert  variables to Numeric variable##############

########### Transform  independent variable to numeric
###########???   class (0: benign, 1: Malignant )####
data$Class=recode(data$Class ,"'Malignant' = 1;'Benign' = 0  ")

table(Class)
data$Class<- as.numeric(data$Class)
prop.table(table(Class))

###########???   Marital status (0: married, 1: divorced, 2: widower, 3: Single 4:-
Separated)############################
data$MaritalStatus=recode(data$MaritalStatus, "'Married' = 0 ;'Divorced' = 1
;'Widowed' = 2 ;'Single'=3 ;'Separated'=4")
table(data$MaritalStatus)
data$MaritalStatus<- as.numeric(data$MaritalStatus)

##########???     Hormonal therapy (0:- not given, 1:-
given)####################################
data$Hormonaltherapy<- recode(data$Hormonaltherapy, "'Yes'= 1 ;'No'= 0 ")
table(data$Hormonaltherapy)
data$Hormonaltherapy<- as.numeric(data$Hormonaltherapy)

############################???              Age group. (0:  20-30,    1: 31-40,   2:
41-50,   3:51-60, 4: more than 60)#######################
agegroup = case_when ( data$`Person Age`>= 20   & data$`Person Age` <= 30 ~ '0',
                    data$`Person Age`>= 31   & data$`Person Age` <= 40 ~ '1',
                    data$`Person Age` >= 41  & data$`Person Age` <=50 ~ '2',
                    data$`Person Age` >= 51  & data$`Person Age` <=  60 ~ '3',
                    data$`Person Age` >= 61   ~ '4' )
table(agegroup)
agegroup<-  as.numeric(agegroup)

########Family History of breast cancer (1: Yes,   0: No)########
data$familyHistory<- recode(data$familyHistory," 'Yes' = 1 ;'No' = 0
")######MM_FHR have family members had breast cancer?
table(data$familyHistory)
data$familyHistory<- as.numeric(data$familyHistory)
##########################          Region( North= 0, middle=1,
south=2)###################.
data$Region <- recode(data$Region,"'North' = 0 ;'middle' = 1;'South' = 2 ")
table(data$Region)
data$Region<-  as.numeric(data$Region)

##########################          BMI (Underweight = <18.5, Normal weight =
18.5???24.9, Overweight = 25???29.9 Obesity = BMI of 30 or
greater)###################.
BMIgroup = case_when(data$BMI <= 18.5   ~ '0',
                  data$BMI >18.5  & data$BMI <=24.9 ~ '1',
                  data$BMI >= 25  & data$BMI<=29.9 ~ '2',
                  data$BMI>= 30  ~ '3')
table(BMIgroup)
BMIgroup<-  as.numeric(BMIgroup)
#####     Age at first pregnancy (1: less than 30 years, 2: 30 years or
higher)############
```

```
age.pregnancy.group = case_when(data$childbirthage <30   ~ '0' , data$childbirthage
>= 30  ~ '1')
table(age.pregnancy.group)
age.pregnancy.group<-  as.numeric(age.pregnancy.group)

###################################??? Breastfeeding (0: No,     1: Yes).###
data$breastfeed <- recode(data$breastfeed,"'Yes' = 1 ;'No' = 0 ")##### Did you
breastfeed your children?
table(data$breastfeed)

data$breastfeed<- as.numeric(data$breastfeed)

####################??? Age at Menarche (1: less than 12 years,   2: 13years,
3:14years,    4: more than 14).##############################
age.Menarch.group = case_when(data$Menarchage <= 12   ~ '0',
                              data$Menarchage == 13   ~ '1',
                              data$Menarchage >= 14   ~ '2')
table(age.Menarch.group)
age.Menarch.group<-  as.numeric(age.Menarch.group)

##################################Smoking (0: No, 1: Yes)#####
data$Smoking=recode(data$Smoking,"'Non smoker' = 0 ;'Smoker currently' = 1;
'Passive smoker'=2 ")
table(data$Smoking)
data$Smoking<- as.numeric(data$Smoking)

##################Alcohol consumption (0: No, 1: Yes)#########
data$Alcoholconsumption=recode(data$Alcoholconsumption,"'No' = 0 ;'Yes' = 1
")###MM_PRFR Alcohol consumption?
```

```
table(data$Alcoholconsumption)
data$Alcoholconsumption<- as.numeric(data$Alcoholconsumption)

##################################expose of radiation-Chest area (0: No, 1:
Yes)########
data$exposeradiation=recode(data$exposeradiation,"'No' = 0 ;'Yes' = 1 ")###expose
of radiation-Chest area.

table(data$exposeradiation)
data$exposeradiation<- as.numeric(data$exposeradiation)

############################## Physical activity (0: No,   1: Yes)########
data$physicalactivity<-recode(data$physicalactivity,"'Yes' = 1 ;'No' = 0 ")
table(data$physicalactivity)
data$physicalactivity<- as.numeric(data$physicalactivity)



##############################???          Education level (0- Diploma or less ,
1- Bachelor(BA)  ,2: Graduate Studies)#####################
data$Education.level = case_when(data$Education.level<= 12   ~ '0',
                         data$Education.level>= 13 & data$Education.level<= 15   ~ '1',
                         data$Education.level== 16   ~ '2',


                         data$Education.level >= 17~ '3' )######## (0- Diploma or less ,
1- Bachelor(BA)  ,2: Graduate Studies)###################
table(data$Education.level)
data$Education.level<-  as.numeric(data$Education.level)
```

```
##############################???          Type of locality (Village= 0 ; City = 1
; Camp = 2).####################

data$locality=recode(data$locality ,"'Village' = 0 ;'City' = 1 ;'Camp' = 2 ")###Type of
locality
table(data$locality)

data$locality=as.numeric(data$locality)

#########################???      Mass shape ('Oval' = 0 ;'Round' = 1 ;'Lobulated'
= 2 ; 'Irregular' = 3 ;'Indistinct'=4;'Spiculated'=5"))#######################3
data$mshape=recode(data$mshape,"'Oval'=0 ;'Round'=1 ;'Lobulated'=2 ; 'Irregular'=3
")

table(data$mshape)
data$mshape<- as.numeric(data$mshape)

######################???            Mass Margin (Indistinct = 0 ;circumscribed = 1
; Micro lobulated = 2 ; Spiculated = 3 ; Obscured= 4 )###################
data$mmargins=recode(data$mmargins,"'Indistinct' = 0 ;'circumscribed' = 1 ;'Micro
lobulated' = 2 ;'Spiculated' = 3 ; 'Obscured' = 4  ")
table(data$mmargins)
data$mmargins<- as.numeric(data$mmargins)

#################Occupation: Housewife = 0 ; Worker = 1 ;Employee = 2;Not
Working=3 ;Farmer=4#########################
data$Career=recode(data$Career,"'Housewife' = 0 ;'Worker' = 1 ;'Employee' = 2;'Not
Working'=3 ;'Farmer'=4")
table(data$Career)
```

```
Career<- as.numeric(data$Career)
######remove binary variables that converted to numeric##############
#Removing the Class  column
data$`Person Age`<- NULL
data$BMI<-NULL
data$childbirthage<-NULL

data$Menarchage<- NULL

###################added new numeric variables (binary to numeric) by using cbind###################
newdata<- cbind(data,agegroup,BMIgroup,age.Menarch.group,age.pregnancy.group)
View(newdata)
summary(newdata)
newdata<-data.frame(newdata)
attach(newdata)

########we have missing data,so we used multiple imputation
data1 <- mice(newdata , m=1) ######imputation by using multiple imputation
data<-complete(data1)
summary(data)
View(data)
summary(data)

test_index <- createDataPartition(data$Class, times = 1, p = 0.8, list = FALSE)
testing_set <- data[-test_index, ]
training_set <-data[test_index, ]
table(training_set$Class)
(testing_set$Class)
```

```
hist(testing_set$Class) ####to see how much malignant and benign  in Test set.

barplot(prop.table(table(testing_set$Class)),
        col = rainbow(4),
        ylim = c(0, 0.7),
        main = "Class Distribution")

library(DMwR)



library(doParallel)
library(ROSE)
set.seed(123)
over_sample_train_data <- ovun.sample(Class ~ ., data = training_set,
method="over", N=1272)$data
print('Number of transactions in train dataset after applying Over sampling method')
print(table(over_sample_train_data$Class))

# Undersampling,as Fraud transactions(1) are having less occurrence, so this Under
sampling method will descrease the Good records untill matches Fraud records, But,
you see that we???ve lost significant information from the sample.
under_sample_train_data <- ovun.sample(Class ~ ., data = training_set,
method="under", N=554)$data
print('Number of transactions in train dataset after applying Under sampling method')
print(table(under_sample_train_data$Class))

# Mixed Sampling, apply both under sampling and over sampling on this imbalanced
data
```

```
both_sample_train_data <- ovun.sample(Class ~ ., data =data, method="both",
p=0.5,N=913, seed = 1)$data
print('Number of transactions in train dataset after applying Mixed sampling method')
print(table(both_sample_train_data$Class))

rose_sample_train_data <- ROSE(Class ~ ., data = training_set,  seed=111)$data
print('Number of transactions in train dataset after applying ROSE sampling method')
print(table(rose_sample_train_data$Class))

##########################################  SVM Algorithm #####

mod1<- svm(Class~. ,data=both_sample_train_data  , type = "C",
gamma=0.125,cost=1,epslon=0.001)

################How to determine the most related attributes that affect breast
cancer.###########

cat('SVM model case:\n')
fit1 <- svm(Class ~ ., data = both_sample_train_data)
w <- t(fit1$coefs) %*% fit1$SV                # weight vectors
w <- apply(w, 2, function(v){sqrt(sum(v^2))})  # weight
w <- sort(w, decreasing = T)
print(w,cat(' The most related attributes that affect breast cancer by using SVM
algorithm:\n'))


#prediction  by using SVM

pred_svm<- predict(mod1, testing_set)
```

```
confusionMatrix(pred_svm,factor(testing_set$Class),positive ="1" )######confusion
matrix
confusionMatrix

# summarize results

table<- table(predict(mod1, testing_set),factor(testing_set$Class))
ctable <- as.table(table, nrow = 2, byrow = TRUE)
fourfoldplot(ctable, color = c("#CC6666", "#99CC99"),
             conf.level = 0, margin =1, main = "confusion_matrix for SVM Algorithm")

###############################  Decision tree algorithm  ########
#load libraries
library(caret)
library(rpart)
tree.model<-rpart(Class~.,data =both_sample_train_data,minbucket = 20)
tree.model
tree.model <- rpart(Class ~ .,data =both_sample_train_data, method = "class")
prp(tree.model,main = "Decision tree Classifier")

################How to determine the most related attributes that affect breast
cancer.###########
library(vip)
library(xgboost)
vi_tree <- tree.model$variable.importance

print(vi_tree,cat('The most related attributes that affect breast cancer by using
Decision tree algorithm:\n'))
```

```
vi_svm

par(mar = c(2, 7, 2,2))

barplot( vi_tree , horiz  = TRUE  , las = 1 ,col = rainbow(4),space =2,
cex.names=0.7,cex.axis=0.7,width = c(765, 840, 150),main = "the most related
attributes that affect breast cancer", font = 2)

tree.predict <- predict(tree.model, testing_set, type = "class")
table<-table(testing_set$Class, tree.predict)
factor(testing_set$Class)
tree.predict
tree<-confusionMatrix((tree.predict),factor(testing_set$Class),positive = "1")
tree

#######################    Naive base   #####################
naive_bayes<-
train(factor(Class)~.,data=both_sample_train_data,method="naive_bayes")
SVMMM<-train(factor(Class)~.,data=both_sample_train_data,method="SVM")
naive_bayes

#####How to determine the most related attributes that affect breast cancer
varImp(naive_bayes,cat('The most related attributes that affect breast cancer by using
Naive Bayes algorithm:\n'))
p<-predict(naive_bayes,testing_set)
p
c<- confusionMatrix(factor(p),factor(testing_set$Class),positive = "1")
c
ctable <- as.table(c, nrow = 2, byrow = TRUE)
```

```
fourfoldplot(ctable, color = c("#CC6666", "#99CC99"),
          conf.level = 0, margin =1, main = "confusion_matrix for Naive Base ")
##############Neural network############
nn <- neuralnet(both_sample_train_data$Class~.,
data=data.frame(both_sample_train_data), hidden=2, linear.output=FALSE,
threshold=0.1)
nn$result.matrix
plot(nn)
model <- train(Class~., data=both_sample_train_data, method="nnet")
imp<-varImp(model)###How to determine the most related attributes that affect
breast cancer
plot(imp)

#Test the resultning output
temp_test <- subset(testing_set, select = c("Region"
,"MaritalStatus","Hormonaltherapy","familyHistory", "Smoking","Education.level",
                                  "Career"
,"mmargins","locality","Alcoholconsumption","exposeradiation","breastfeed",

"physicalactivity","mshape","agegroup","BMIgroup","age.Menarch.group",
                                  "age.pregnancy.group"))
head(temp_test)

nn.results<-neuralnet::compute(nn, temp_test)

results <- data.frame(actual =testing_set$Class, prediction =
nn.results$net.result)#Test the resulting output
roundedresults<-sapply(results,round,digits=0)
roundedresultsdf=data.frame(roundedresults)
```

```
attach(roundedresultsdf)
table(prediction,actual)
con<-confusionMatrix(table(actual,prediction),positive = "1")
con

con<-confusionMatrix(table(actual,prediction))
con
ctable <- as.table(con, nrow = 2, byrow = TRUE)
fourfoldplot(ctable, color = c("#CC6666", "#99CC99"),
             conf.level = 0, margin =1, main = "confusion_matrix for Neural Network")
```

The End