Diagnosing Wisconsin Breast Cancer : A Machine Learning Approach

Alberta Araba Johnson
(MS Applied Statistics and Data Science)

Instructor: Dr. Hansapani Rodrigo

School of Mathematical and Statistical Sciences
The University of Texas Rio Grande Valley

December 10, 2022

ABSTRACT

Breast cancer (BC) is one of the most common cancers among women worldwide, and accounts for the majority of new cancer cases and cancer-related deaths, making it a serious public health issue in today's society. Because it can encourage prompt clinical care for patients, an early diagnosis of BC can considerably enhance the prognosis and likelihood of survival. A more precise classification of benign tumors could spare patients from receiving unneeded care. As a result, there is a lot of research into the proper diagnosis of BC and the classification of patients into groups that are malignant or benign. Machine learning (ML) is widely acknowledged as the preferred approach for classifying BC patterns and forecasting future cases due to its distinct benefits in the discovery of important features from complex BC datasets. In this project, nine traditional ML algorithms were considered in BC diagnosis: logistic regression, K-Nearest neighbors, decision tree, naive bayes, neural network, gradient boosting, support vector machines, discriminant analysis and random forest. Then, we investigate their applications in BC. Our primary data is drawn from the Wisconsin breast cancer database (WBCD) which is the benchmark database for comparing the results through different algorithms. The purpose of this was to investigate which of these supervised learning techniques had the highest predictive power for the data under consideration. For precision analysis, sensitivity, specificity, F1-score and the balanced accuracy were used as the comparison criteria. The resulting values obtained suggested that the neural network and the KNN model outperformed the other method with a balanced accuracy and F1-score of 97.73% and 96.30% respectively. Hence by the metrics considered, ANN and KNN were ranked more accurate in prediction than the other methods.

ACKNOWLEDGEMENTS

My gratitude goes first and foremost to the Almighty God whose everlasting protection and mercies strengthens me throughout my studies. To Him be the Glory.

I wish to express my heartfelt gratitude to all those who have provided me with the opportunity to complete this study.

I give a special gratitude to my course instructor, Dr. Hansapani Rodrigo, whose helpful recommendations and support made it possible to finish this project. As my instructor, she has taught me more than I could ever give her credit for here. She has shown me, by her example, what a good Statistician (and person) should be.

Finally, I must express my gratitude for the help received from my other colleagues, especially when it came to acquiring information that enhanced the content of my project work.

Thank you for your assistance.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Problem Statement

Cancer begins with uncontrolled division of cells, resulting in a visible mass called a tumor. Tumors can be either benign or malignant. Malignant tumors grow rapidly, invading and causing damage to surrounding tissues. Breast cancer is a malignant tissue that begins to grow in the breast. It is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today's society [1]. Symptoms of breast cancer include breast volume, changes in breast shape and size, breast skin colour variations, breast pain, and genetic changes.

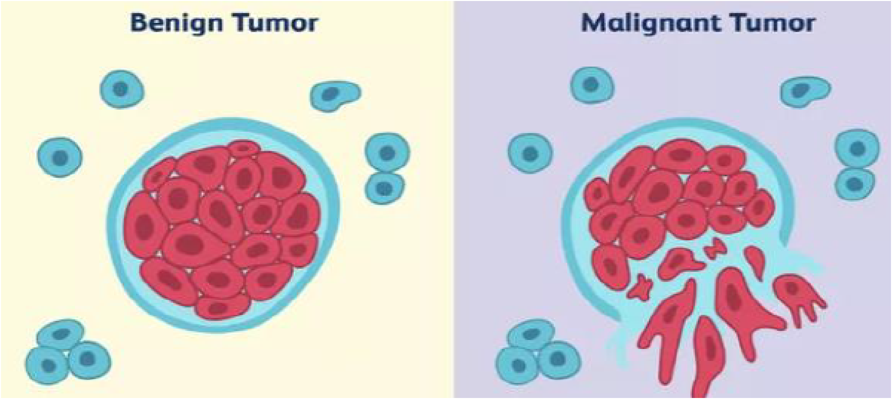


Figure 1.1: The Benign and Malignant Tumor

Breast cancer is the second leading cause of death among women worldwide, and more than 8% of women will develop the disease during their lifetime. The World Health Organization (WHO) reports that approximately 1,000,000 women are newly diagnosed with breast cancer each year, and more than 500,000 of women die from breast cancer [2]. It is estimated that this disease's incidence will increase in the future as we deal with environmental degradation.

In 2008, approximately 182,460 newly diagnosed cases and 40,480 deaths were reported in the United States [3]. Because the cause of breast cancer is still unknown, early detection is key to reducing mortality (over 40%). The earlier the cancer is detected, the more effective it will be. A prerequisite for early detection is an accurate and reliable diagnosis that can distinguish between benign and malignant tumors. A suitable detection approach should yield both low false positive and false negative rates.

Previously, mammography was the most effective method of detecting and diagnosing this. Although the incidence of breast cancers increased over the past decade, breast cancer mortality decreased in women of all ages. This favourable trend in mortality reduction may be related to improved breast cancer treatment and widespread screening mammography [3]. However, it is well known that expert radiologists can miss a significant proportion of abnormalities.

In addition, a large number of mammographic abnormalities turn out to be benign after biopsy. Conventional methods of monitoring and diagnosing the diseases rely on detecting the presence of particular signal features by a human observer [4]. Due to large number of patients in intensive care units and the need for continuous observation of such conditions, several computer aided-diagnosis approaches for automated diagnostic systems have been developed in the past ten years to attempt to solve this problem. Such techniques work by transforming the mostly qualitative diagnostic criteria into a more objective quantitative feature classification problem.

Despite the fact that these traditional approaches are fundamental clinical and laboratory procedures, routine screening is not universally feasible; this can lead to missed or delayed diagnoses. Due to this, it is critical to quickly identify those who are at high risk to improve timely clinical treatment of these patients, but it is still unclear how to identify their risk.

The early diagnosis of breast cancer can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. As a result, this work aims to use machine learning technique to develop a predictive model to help detect benign and malignant tumors onset in women with early-stage cancer.

1.2 Research Objective

This project seeks to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in the model selection. The goal is to classify whether the breast cancer is benign or malignant.

1.3 Methods Used

In order to achieve the goal of this project, the following machine learning techniques were employed:

(a) Logistic Regression

(b) K Nearest Neighbors Classifier

(c) Support Vector Machine

(d) Naive-Baye's Classifier

(e) Discriminant Analysis

(f) Random Forest Classifier

(g) Decision Tree Classifier

(h) Artificial Neural Network

(i) Gradient Boosting Classifier

1.4 Relevance of the Study

(a) The model would act as a major predictive indication to assist medical professionals in identifying the prevalence of breast cancer in women from the onset, allowing them to make quick and informed decisions for patients before the problem worsens.

(b) The study's selected attributes would be an important tool and could offer vital data that could serve as a logical supplement to help with the diagnosis and clinical trials of breast cancer treatments.

1.5 Organisation of Report

This report is subdivided into five parts. The first chapter presents the problem statement, research objectives, and methodologies employed to attain the set objectives. The second chapter presents important details about the study and other research works that have been carried out concerning the concept. Chapter three discusses the methods used to build up the project. The sequential approach to how the models were formulated, the project outcomes, analysis, and discussions are presented in Chapter four. The final chapter provides the conclusions and recommendations for further studies.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

In this section, review of works of several authors on the diagnosis of breast cancer using machine learning techniques were discussed.

2.2 Previous Research Concerning the Concept

The goal of using machine learning techniques to aid in breast cancer detection has been the subject of extensive research to date. Many researchers have made progress in their studies of breast cancer by utilizing a variety of datasets, including the SEER dataset, mammogram pictures as a dataset, Wisconsin Dataset, and dataset from different hospitals. By exploiting these dataset authors extract and select various features and complete their research. These are some significant works.

Kiyan and Yildirim. [5] has discussed that statistical neural networks can be used to perform breast cancer diagnosis effectively. The scholar has compared statistical neural network with Multi Layer Perceptron on WBCD database. Radial basis function(RBF), General Regression Neural Network(GRNN), Probabilistic Neural Network(PNN) were used for classification and their overall performance were 96.18% for RBF, 97% PNN, 98.8% for GRNN and 95.74% for MLP. Hence it is proved that these statistical neural network structures can be applied to diagnose breast cancer.

Maglogiannis *et al.* [6] have presented an article on An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers with Bayesian classifiers and ANN for prognosis and diagnosis of breast cancer disease. Wisconsin diagnostic breast cancer datasets were used to implement SVM model to provide distinction between the malignant and benign breast masses. These datasets involve measurement taken according to Fine Needle Aspirates (FNA). The article provides the implementation details along with the corresponding results for all the assessed classifiers. Several comparative studies have been carried out concerning both the prognosis and diagnosis problem demonstrating the superiority of the proposed SVM algorithm in terms of sensitivity, specificity and accuracy.

Sahran *et al.* [7] developed a computerized breast cancer diagnosis by combining genetic algorithm and Back propagation neural network which was developed as faster classifier model to reduce the diagnose time as well as increasing the accuracy in classifying mass in breast to either benign or malignant. In these two different cleaning processes was carried out on the dataset. In Set A, it only eliminated records with missing values, while set B was trained with normal statistical cleaning process to identify any noisy or missing values. At last Set A gave 100% of highest accuracy percentage and set B gave 83.36% of accuracy. Hence the author concluded that medical data are best kept in its original value as it gives high accuracy percentage as compared to altered data.

Delen *et al.* [8] used three prevalent information mining methods named Decision Trees, Artificial Neural Networks and Support Vector Machines alongside the most normally utilized measurable examination systems, for example, Logistic Regression to develop forecast models for prostate malignant growth survivability. The informational collection encased around 120,000 records and 77 factors. A K-Fold cross-validation process was executed in model structure, assess and examinations. The outcome showed that SVMs was the most precise (with a test set accuracy of 92.85%) for this zone, trailed by ANNs and DTs.

Ngadi *et al.* [9] used support vector machine algorithm to test different classification methods including RBF, Poly, and Linear. Then they compared the results with other classification methods such as Naïve Bayes, decision tree, K nearest neighbour, support vector machines, random forest, and Adaboost. Random forest had the best performance result with 93% accuracy. This proves that neighboring Support Vector Classier (NSVC) was better than the other methods.

Salma [10] selected two different data sets from WBCD and KDD also they used Fast Modular Artificial Neural Network (FM-ANN) for both of them. They compared the results with other techniques (RBF, Feed forward Neural Network (FNN), and Modular Neural Network (MNN)). After training and testing KDD achieved better accuracy of 99.96% due to the number of features were more. Comparing the results FM- ANN proved to be more accurate.

Asri *et al.* [11], demonstrated that Support vector Machine (SVM) proves its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of precision and low error rate with an accuracy of 97.13%. In other work, Osman [12] proposed a solution for the diagnosis of Wisconsin breast cancer with a prediction of 99.10% found by the SVM algorithm by combining a clustering algorithm with an efficient probabilistic vector support machine.

Following the review of literature on the suggested topic, this project is focused on assessing machine learning algorithms and approaches in order to conclude the best methodology for breast cancer prediction. The actual investigation seeks to assess the effectiveness and performance of the methods for the data under consideration.

CHAPTER 3
METHODOLOGY

3.1 Overview

This chapter outlines the various methods and procedures considered in this project. A brief discussion of the machine learning algorithms were presented. These methods were implemented to achieve the objectives of the research.

3.2 Machine Learning Techniques

ML provides a set of methods/tools for identifying patterns in large databases. Those patterns could be then used in prediction or decision making. Machine learning is a multi-disciplinary (statistics, mathematics, computer science, etc.) provides algorithms that are designed to learn from data.

The subsections below gives a brief description of some of the ML techniques used in this project.

3.2.1 Logistic Regression

Logistic Regression is a classification technique used in machine learning. It uses a logistic function to model the dependent variable. The dependent variable is dichotomous in nature, i.e. there could only be two possible classes (eg. either the cancer is malignant or not). As a result, this technique is used while dealing with binary data.

In logistic regression, in order to map the predicted values to probabilities, sigmoid function is used. The sigmoid function/logistic function is a function that resembles an “S” shaped curve when plotted on a graph. It takes values between 0 and 1 and “squishes” them towards the margins at the top and bottom, labeling them as 0 or 1. The equation for the sigmoid function is $y = 1/(1 + e^{-\mathrm{x}})$.

The decision for converting a predicted probability into a class label is decided by the parameter known as Threshold. A value above that threshold indicates one class while the one below indicates the other.

3.2.2 K-Nearest Neighbourhood (KNN)

KNN classification model is a simple algorithm that uses the entire dataset in its training phase. Whenever a prediction is required for an unseen data instance, it searches through the entire training dataset for K-most similar instances and the data with the most similar instance is returned as the prediction. This technique suggests that if you are similar to your neighbours, then you are one of them. It is a simple algorithm which stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions).

In KNN, K is the number of nearest neighbors. The number of neighbors is the core deciding factor. K is generally an odd number if the number of classes is 2. When K=1, then the algorithm is known as the nearest neighbor algorithm. This is the simplest case. Suppose P1 is the point, for which label needs to predict. First, you find the one closest point to P1 and then the label of the nearest point assigned to P1.

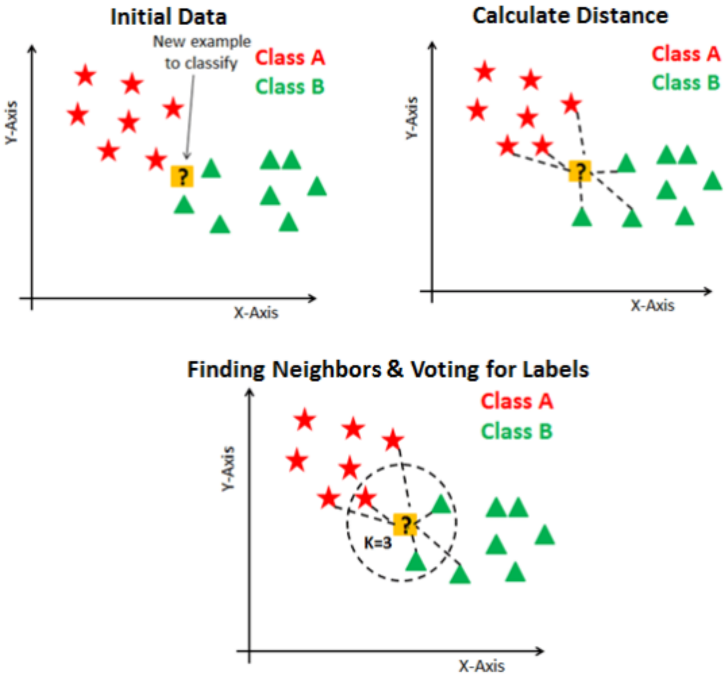


Figure 3.1: Illustration of KNN Classification Algorithm

Suppose P1 is the point, for which label needs to predict. First, you find the K closest point to P1 and then classify points by majority vote of its K neighbors. Each object votes for their class and the class with the most votes is taken as the prediction. For finding closest similar points, you find the distance between points using distance measures such as Euclidean distance, Hamming distance, Manhattan distance and Minkowski distance. KNN has the following basic steps: Calculate distance, Find closest neighbors and Vote for labels.

3.2.3 Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.

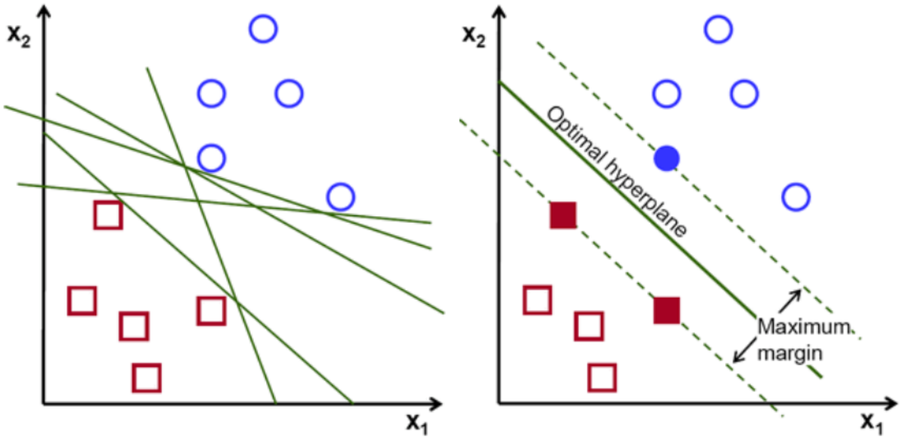


Figure 3.2: Possible Hyperplanes for a SVM Classifier

To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane. It becomes difficult to imagine when the number of features exceeds 3.

3.2.4 Naive Bayes Classifier

Naive Bayes classifier is a probabilistic machine learning model that is used for classification task. The crux of the classifier is based on the Bayes theorem. The Bayes theorem is given as P(A|B) = (P(B|A)P(A))/(P(B)).

Using Bayes theorem, we can find the probability of A happening, given that B has occurred. Here, B is the evidence and A is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other. Hence it is called naive.

3.2.5 Gradient Boosting

Boosting is a special type of Ensemble Learning technique that works by combining several weak learners (predictors with poor accuracy) into a strong learner (a model with strong accuracy). This works by each model paying attention to its predecessor's mistakes.

In Gradient Boosting, each predictor tries to improve on its predecessor by reducing the errors. But the fascinating idea behind Gradient Boosting is that instead of fitting a predictor on the data at each iteration, it actually fits a new predictor to the residual errors made by the previous predictor.

Gradient Boosting has three main components:

(a) Loss Function - The role of the loss function is to estimate how good the model is at making predictions with the given data. For classification,it helps us understand how accurate our model is at classifying people who did or didn't like certain movies.

(b) Weak Learner - A weak learner is one that classifies our data but does so poorly, perhaps no better than random guessing. In other words, it has a high error rate.

(c) Additive Model - This is the iterative and sequential approach of adding the trees (weak learners) one step at a time. After each iteration, we need to be closer to our final model. In other words, each iteration should reduce the value of our loss function.

3.2.6 Neural Network

The theory of neural network computation provides interesting techniques that mimic the human brain and nervous system. Neural network is an information technology, capable of representing knowledge based on massive parallel processing and pattern recognition based on past experience or examples. The artificial neural networks have been extensively studied and used in time series forecasting. The pattern recognition ability of a neural network makes it a good alternative classification and forecasting tool in business applications. In addition, a neural network is expected to be superior to traditional statistical methods in forecasting because a neural network is better able to recognize the high-level features, such as serial correlation, if any, of a training set.

Artificial Neural Network has been developed as generalizations of mathematical models of human cognition or neural biology, based on the assumptions that:

(a) Information processing occurs at several simple elements that are called neurons.
(b) Signals are passed between neurons over connection links.
(c) Each connection link has an associated weight, which, in a typical neural net multiplies the signal transmitted.
(d) Each neuron applies an activation function (usually nonlinear) to its net input (sum of weighted input signals) to determine its output signal.

Through replicate learning process and associative memory, the ANN model can accurately classify information as pre-specified pattern. A typical ANN consists of a number of simple processing elements called neurons, nodes or units. Each neuron is connected to other neurons by means of directed communication links. Each connection has an associated weight. The weights are the parameters of the model being used by the net to solve a problem. ANNs are usually modelled into one input layer, one or several hidden layers, and one output layer.

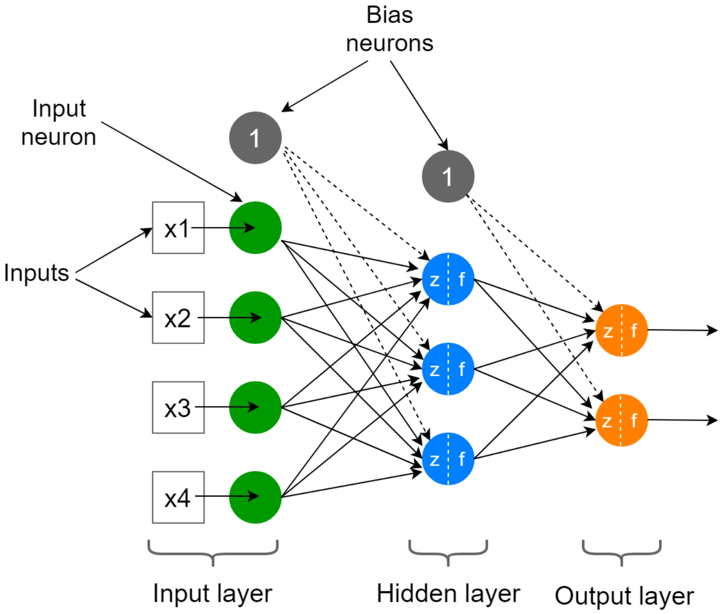


Figure 3.3: Architecture of Artificial Neural Network

The Components of ANNs are neurons, connections and weight, and propagation function. ANNs are composed of artificial neurons which are derived from biological neurons. Each artificial neuron has inputs and produces a single output which can be sent to multiple other neurons. The inputs can be the feature values of a sample of external data, such as images or documents, or they can be the outputs of other neurons.

3.2.7 Decision Tree

Decision Tree is a model that presents classifications as a tree. The data set is broken to small sub-data, then to smaller ones. As a result, the tree is developed and at the last level, the result is revealed. In a tree structure, the leaves characterize the class labels whereby the branches characterize conjunctions of feature leading to the class labels Hence, DT is not sensitive to noise.

The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem based on given conditions

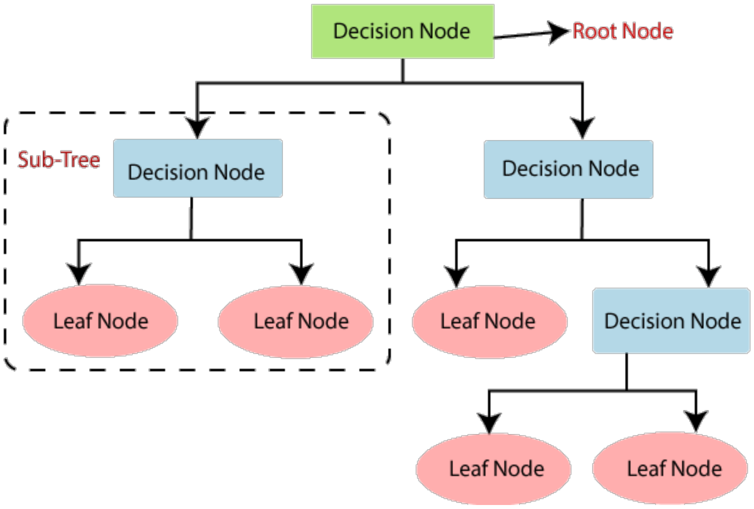


Figure 3.4: Architecture of a Decision Tree

Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets. Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node. Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions. Branch/Sub tree is formed by splitting the tree. Pruning is the process of removing the unwanted branches from the tree. The root node of the tree is called the parent node, and other nodes are called the child nodes.

3.2.8 Random Forest

RF algorithm is used at the regularization point where the model quality is highest, variance and bias problems are compromised [14]. RF builds numerous numbers of DTs using random samples with a replacement to overcome the problem of DTs. Each tree classifies its observations, and majority votes decision is chosen. RF is used in the unsupervised mode for assessing proximities among data points.

3.2.9 Discriminant Analysis

Discriminant analysis (DA) is a multivariate technique used to separate two or more groups of observations based on K variables measured on each experimental sample and find the contribution of each variable in separating the groups. DA works by finding one or more linear combinations of the K selected variables.

Furthermore, prediction or allocation of new observations to previously defined groups can be investigated with a linear or quadratic function to assign each individual to one of the predefined groups. The end result of DA is a model that can be used for the prediction of group memberships. This model allows us to understand the relationship between the set of selected variables and the observations. Furthermore, this model will enable one to assess the contributions of different variables.

Linear discriminant analysis (LDA) is a simple classification method, mathematically robust, and often produces robust models, whose accuracy is as good as more complex methods. LDA assumes that the various classes collecting similar objects (from a given area) are described by multivariate normal distributions having the same covariance but different location of centroids within the variable domain.

Quadratic discriminant analysis (QDA) is a general discriminant function with quadratic decision boundaries which can be used to classify data sets with two or more classes. QDA has more predictability power than LDA but it needs to estimate the covariance matrix for each class.

3.3 Cross-Validation (CV)

Cross-validation is a crucial evaluation technique in machine learning used to evaluate the generalization ability of a model.

It assesses a model's capacity to predict new data that was not used in the estimate process in order to detect flaws like over-fitting or selection bias, as well as to predict how the model would generalize to a different data set. It is a re-sampling strategy in which different portions of the data are used to test and train a model on successive iterations. It is mostly used in situations when the goal is to anticipate how well a predictive model will perform in practice. In machine learning, there are several types of CVs. However, the K-fold CV was used in this study.

K-Fold Cross-Validation

In K-fold cross-validation, the training data of size N is randomly partitioned into K equal subsets. Out of these K subsets, we use K-1 subsets as the training set and the remaining as our test set. This process is repeated for K iterations. In each iteration, a different fold is kept for testing, and the remaining K-1 is used for training.

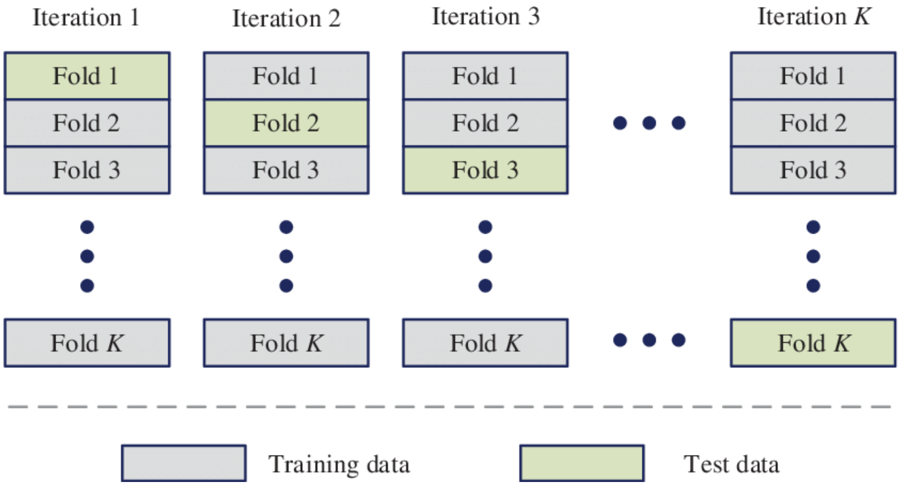


Figure 3.5: Illustration of the K-Fold Cross Validation

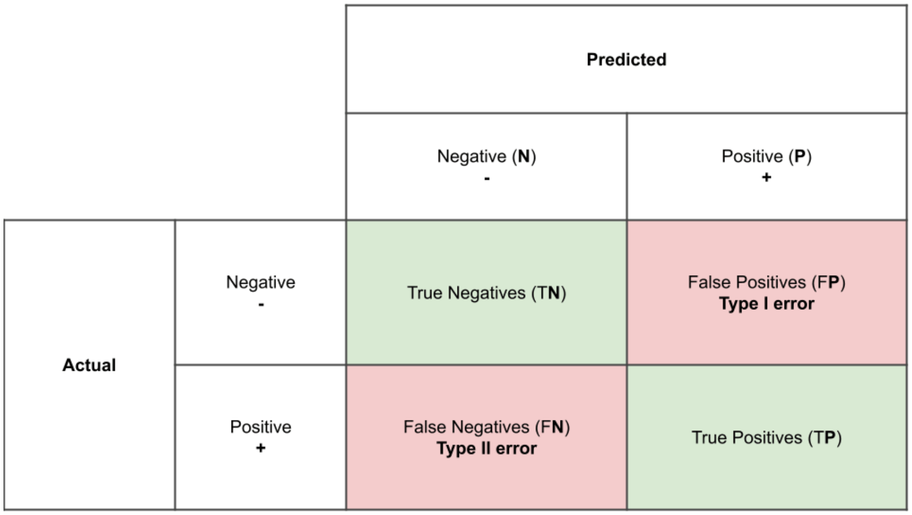
The mean of the values computed in the loop becomes the performance metric from the K-fold cross-validation. In this project, we used the 10-fold CV (i.e., setting K=10).

3.4 Performance Metrics

In machine learning model, we desire to know how it performs, this performance is measured with metrics. Until the performance is good enough with satisfactory metrics, the model is not worth deploying, we have to keep iterating to find the sweet spot where the model is not under-fitting nor over-fitting (a perfect balance).

There are different metrics for measuring the performance of a machine learning model. The metrics used are explained below.

Confusion Matrix is the way to measure the performance of a classification problem where the output can be of two or more type of classes. A confusion matrix is a table with two dimensions viz. "Actual" and "Predicted" and furthermore, both the dimensions have "True Positives", "True Negatives", "False Positives", and "False Negatives". It shows how well the model is performing, what needs to be improved, and what error it is making.



		Predicted	
		Negative (**N**) **-**	Positive (**P**) **+**
Actual	Negative **-**	True Negatives (T**N**)	False Positives (F**P**) **Type I error**
	Positive **+**	False Negatives (F**N**) **Type II error**	True Positives (T**P**)

Figure 3.6: Layout of a Confusion Matrix

Where, true positive is the correctly predicted positive class outcome of the model, true negative represents the correctly predicted negative class outcome of the model, false positive is the incorrectly predicted positive class outcome of the model and false negative the incorrectly predicted negative class outcome of the model.

Sensitivity (or Recall): Sensitivity is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could be made by the model. Mathematically, it is given as TP/(TP+FN).

Specificity: Also known as true negative rate, it measures the proportion of correctly identified negatives over the total negative prediction made by the model. Specificity = TN / (TN + FP).

Precision: this quantifies the number of correct positive predictions made out of positive predictions made by the model. Precision calculates the accuracy of the True Positive. Precision = TP/(TP + FP).

F1-Score: F1-Score keeps the balance between precision and recall. It is often used when class distribution is uneven, but it can also be defined as a statistical measure of the accuracy of an individual test. F1-Score = 2 ∗ (Precision ∗ Recall) / (Precision + Recall).

Balanced Accuracy: It is the arithmetic mean of sensitivity and specificity, its use case is when dealing with imbalanced data, i.e. when one of the target classes appears a lot more than the other. Balanced Accuracy = (Sensitivity + Specificity)/2.

CHAPTER 4
RESULTS, ANALYSIS AND DISCUSSIONS

4.1 Overview

This chapter examines and analyzes the data in depth. It began with some description and information for the data considered. R statistical software was used for all analyses.

4.2 Data Source and Description

The data for this project was a secondary data from the UCI Machine Learning repository, and it consists of ten attributes (including the target variable). The table below gives a description of the attributes and their corresponding domain.

Table 4.1: Description of Attributes in the Dataset

	Attribute	Domain
1	Clump Thickness (CT)	1 - 10
2	Uniformity of Cell Size (UCSi)	1 - 10
3	Uniformity of Cell Shape (UCSh)	1 - 10
4	Marginal Adhesion (MA)	1 - 10
5	Single Epithelial Cell Size (SECS)	1 - 10
6	Bare Nuclei (BN)	1 - 10
7	Bland Chromatin (BC)	1 - 10
8	Normal Nucleoli (NN)	1 - 10
9	Mitoses	1 - 10
10	Class	0 (B) or 1 (M)

In this data, the dependent variable (target) is the Class which is either benign or malignant. The other nine variables (CT, UCSi, UCSh, MA, SECS, BN, BC, NN and Mitoses) are the predictors considered in this work.

In all, the data consists of 699 instances of which 458 are Benign and 241 are Malignant.

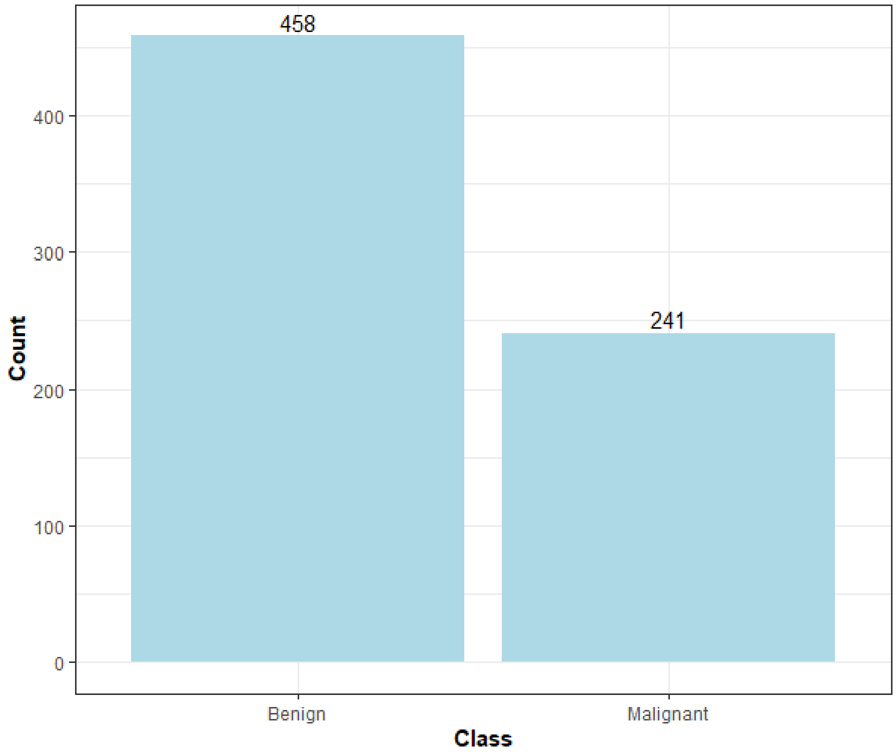


Figure 4.1: Distribution of the Classes in the Data

To build the machine learning models, the data was split into 80% training and 20% testing.

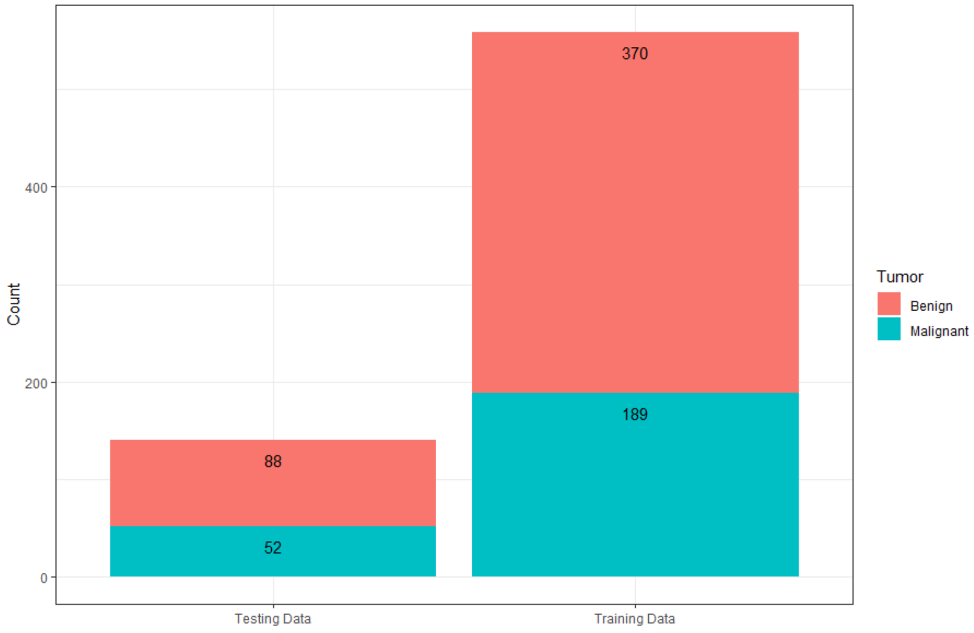


Figure 4.2: Distribution of the Classes Based on 80% - 20% Splitting

The training data sets has 370 Benign and 189 Malignant instances. The testing set, on the other hand, consist of 88 Benign and 52 Malignant instances.

4.3 Analysis and Discussions

In order to predict the testing data, parameter tuning was carried out on the methods to achieve their optimal parameters for a better accuracy. This section shows the cross-validation results for the models.

4.3.1 Logistic Regression

In this section, logistic model with subset selection was built to find the optimal model, i.e., the model with predictors that are statistically significant to the prediction of breast cancer based on Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

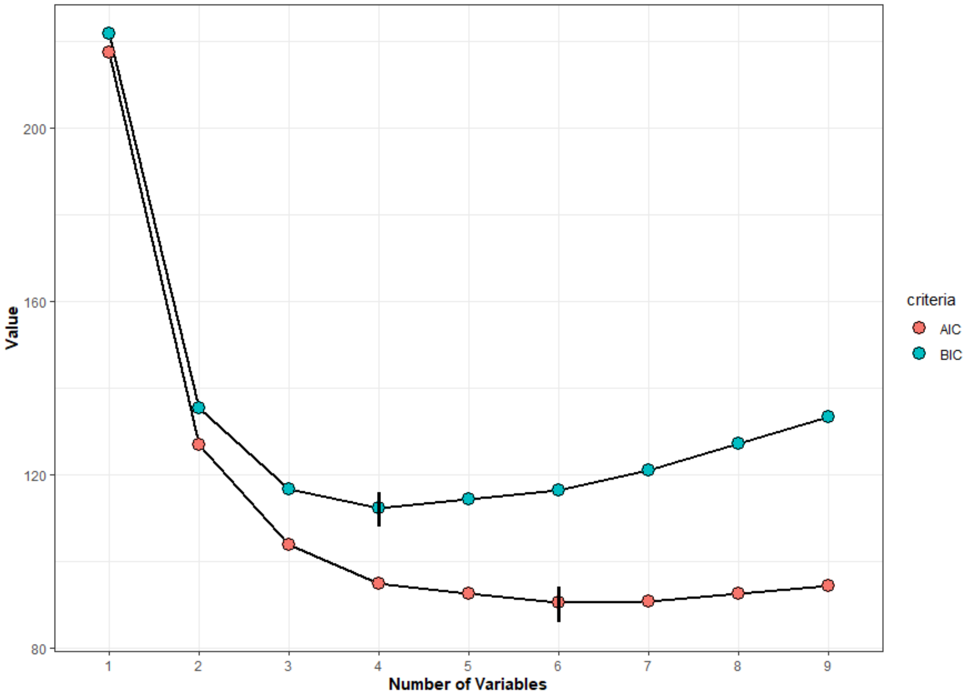


Figure 4.3: Plots of AIC and BIC Values for p Subsets Logistic Models

From Figure (4.3), the best subsets logistic model with the least AIC has 6 predictors which include CT, UCSh, BN, BC, NN and Mitoses. Also, the best subsets logistic model with the least BIC has 4 predictors which include CT, UCSh, BN and BC.

Therefore, the respective predictors were considered in predicting the logistic models based on AIC and BIC.

4.3.2 KNN Classification

Here, different KNN models were fitted for K from 1 to 10 and their training and testing accuracy are represented in Figure (4.4) below.

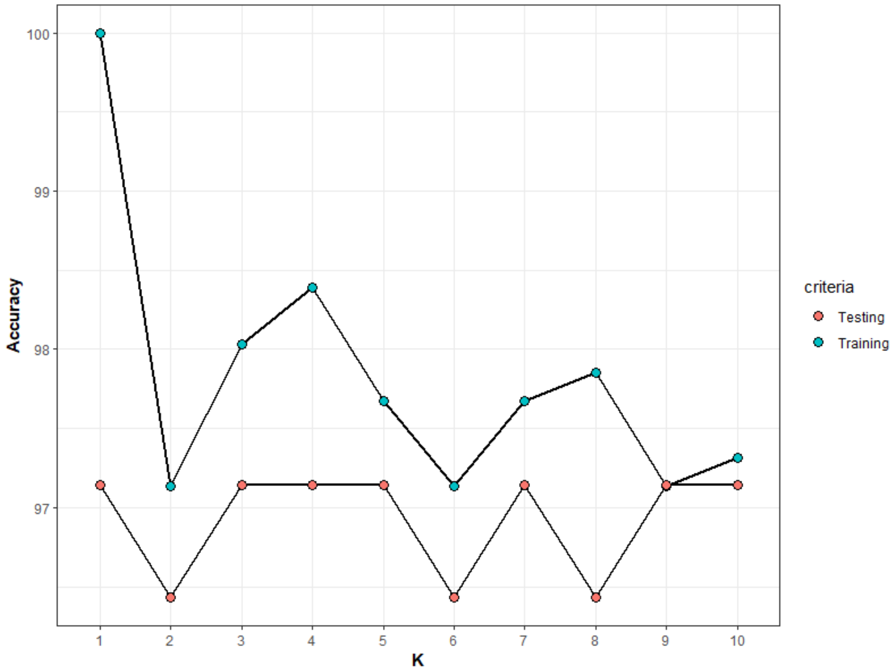


Figure 4.4: K Values and their Training and Testing Errors

To ensure that there is no issue of over-fitting, K value with a good testing accuracy and a minimal difference between the training accuracy and the testing accuracy is considered. From the Figure above, K=9 gave a high accuracy (with 97.143% training accuracy and 97.143% testing accuracy). Therefore, we used K=9 in the final prediction.

4.3.3 Random Forest

Here, we built a Random forest model from the randomForest package in R using the train data. Different "mtry" values were considered. The Out-of-Bag (OOB) error for each mtry value was evaluated.

The aim of this was to use the mtry value with the lowest OOB error (best model) to create a variable importance plot to decide on the importance of the variables and to make predictions for testing data set.

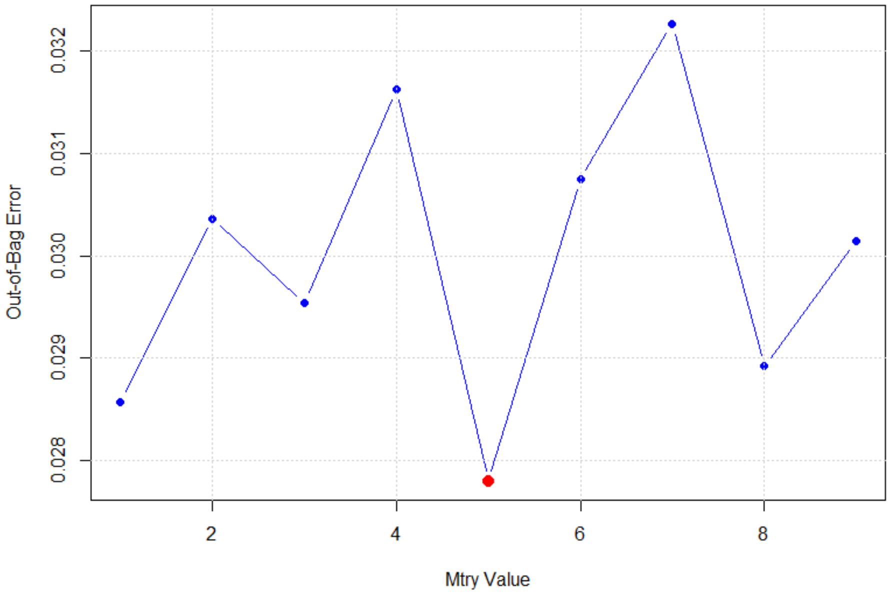


Figure 4.5: Plot of mtry Values and their Out-of-Bag Errors

Figure (4.5) shows that the best mtry value with the least OOB error occurred at 5 (in red). Therefore, mtry=5 was used to fit a random forest model to examine its performance.

Again, the importance of the predictors were examined in Figure (4.6) below. Each value for a specific variable is the amount of accuracy lost by omitting that variable from the model. The more the accuracy declines, the more critical the variable becomes for classification success. As a result, the greater the value, the more important the variable is.

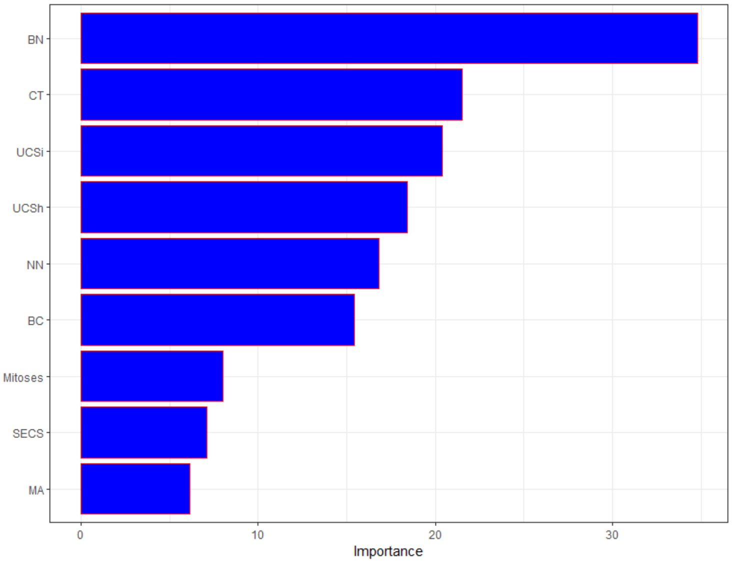


Figure 4.6: Variable Importance Plot from the Random Forest Model

From the importance plot, the most important variable is BN, and the least important is MA. It ranked the variables from the most important to the least important variable. After training the random forest model and examining the importance of the variables, the predictive power of the model was tested by predicting the potential class of each instance.

4.3.4 Artificial Neural Network

To find the optimal parameters for the neural network model, different values were considered for the number of hidden nodes and the weight decay. Figure (4.7) below shows that the highest cross-validation accuracy was attained at a weight decay of 0.1 and hidden neurons of 3. As a result, these optimal parameters were used for predicting the testing dataset.

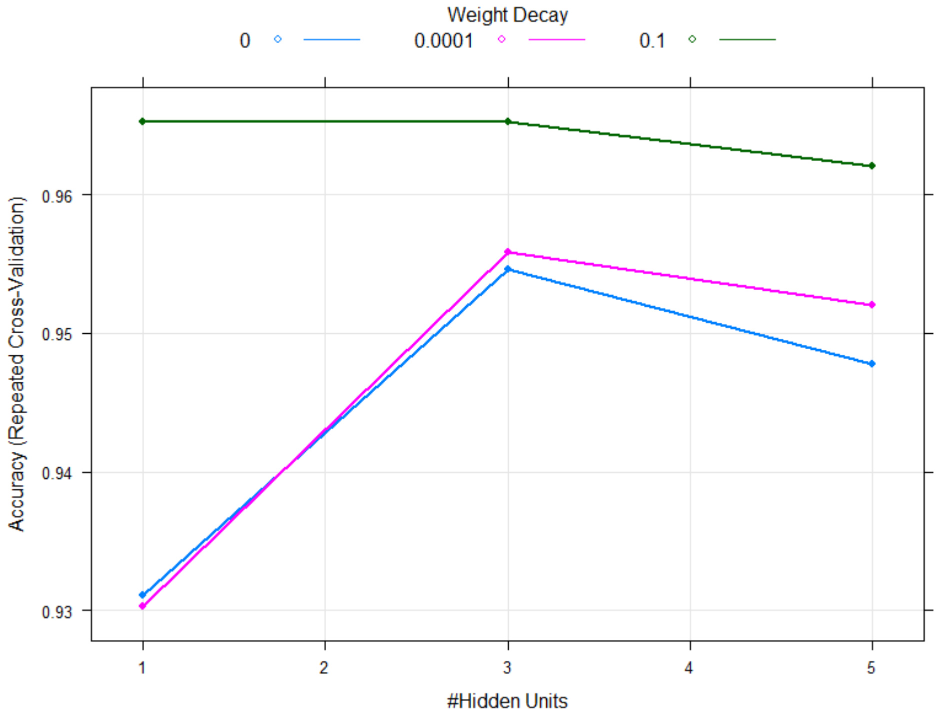


Figure 4.7: Parameter Tuning for the Neural Network Model

4.3.5 Decision Tree

In decision tree model, the complexity parameter (cp) is a key parameter which influences the performance of the model. As a result, model tuning was performed to attain the optimal cp value with high cross-validation accuracy.

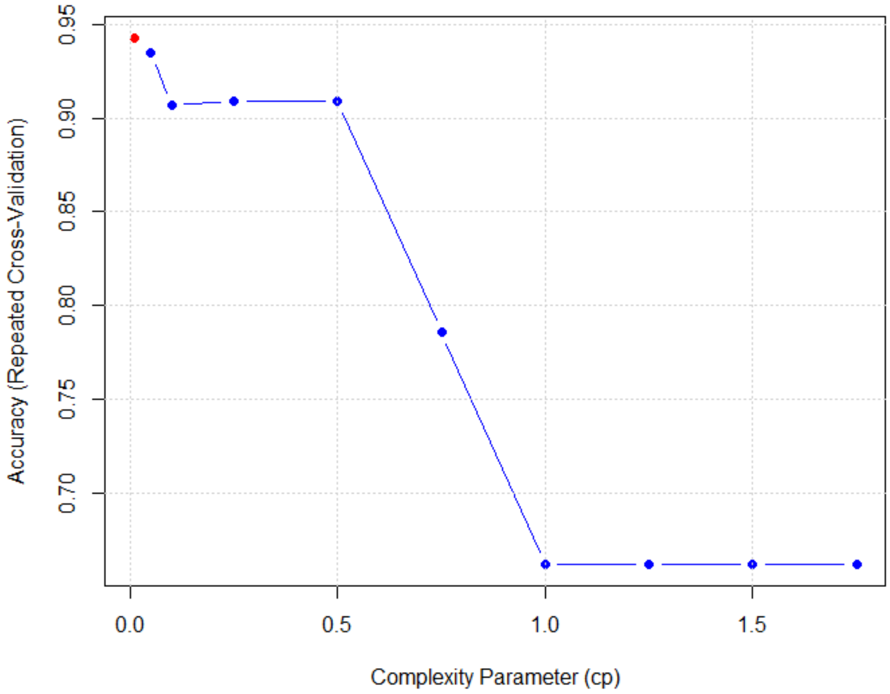


Figure 4.8: Complexity Parameter (cp) Tuning in Decision Tree

From Figure (4.8), the optimal cp was achieved at 0.01 (coloured in red). As a result, this optimal parameter was used for predicting the testing dataset.

In the decision tree model, two different splitting criteria were also considered, i.e., Information Gain and Gini Index and their respective trees are displayed below.

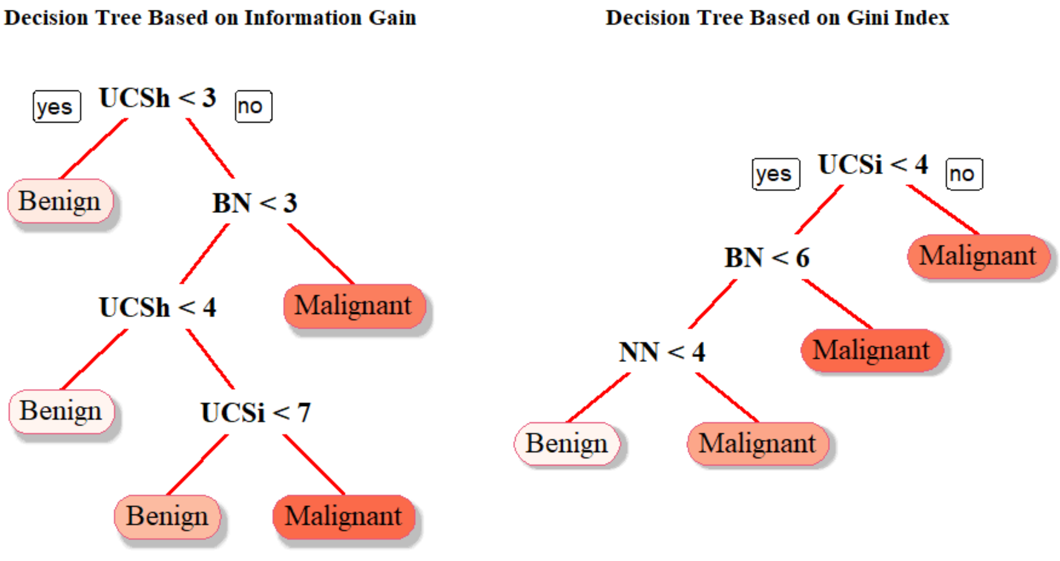


Figure 4.9: Decision Tree Model

4.3.6 Support Vector Machines (SVM)

In this model, the cost (C) is an important parameter which influences the performance of the model. As a result, model tuning was performed to attain the optimal value with high cross-validation accuracy.

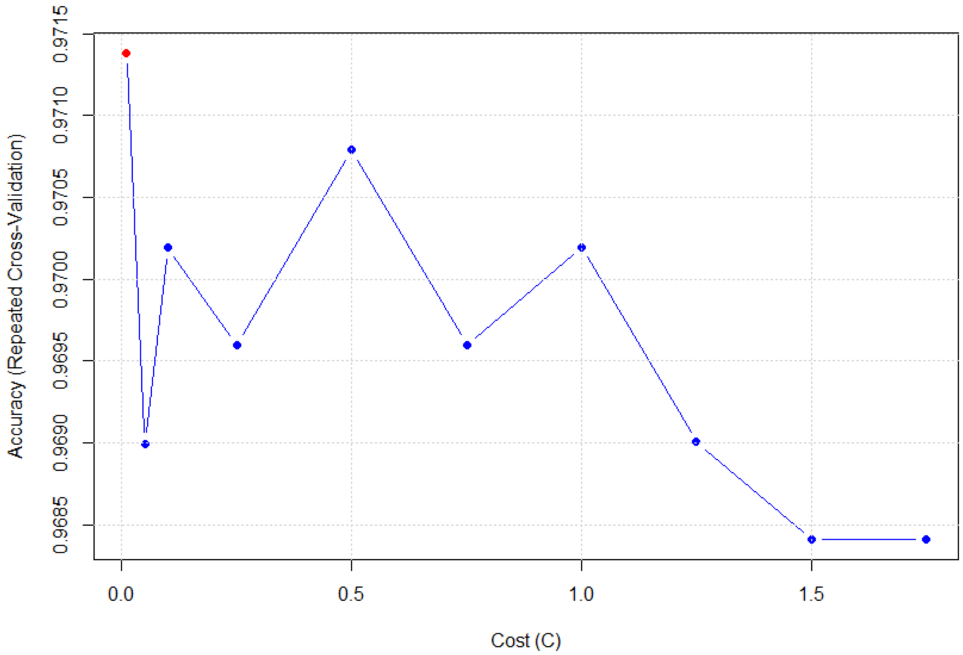


Figure 4.10: Cost (C) Parameter Tuning in SVM

4.3.7 Gradient Boosting Classifier

To find the optimal parameters for the gradient boosting method, different values were considered for the number of trees, interaction.depth, shrinkage factor and the n.minobsinnode. Figure (4.11) shows that the highest cross-validation accuracy was attained at a n.trees = 50, interaction.depth = 5, shrinkage = 0.1 and n.minobsinnode = 20. As a result, these optimal parameters were used for predicting the testing dataset.

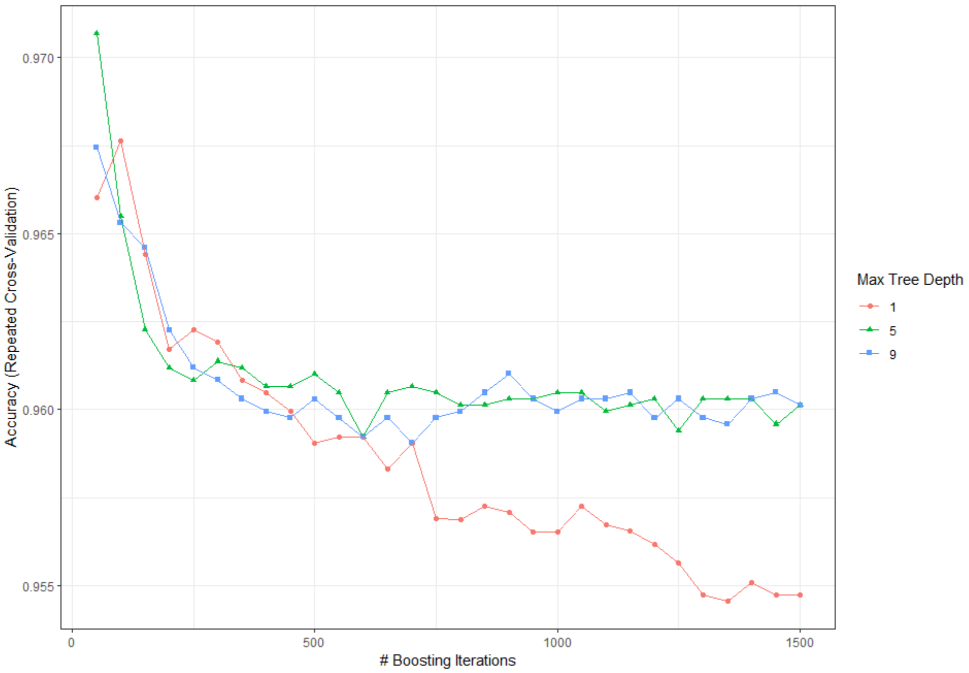


Figure 4.11: Parameter Tuning for Gradient Boosting Classifier

4.4 Performance Metrics Comparison

After applying the ML algorithms on the data. We used balanced accuracy, sensitivity, specificity, and F1 score to evaluate and compare the models to identify the best model.

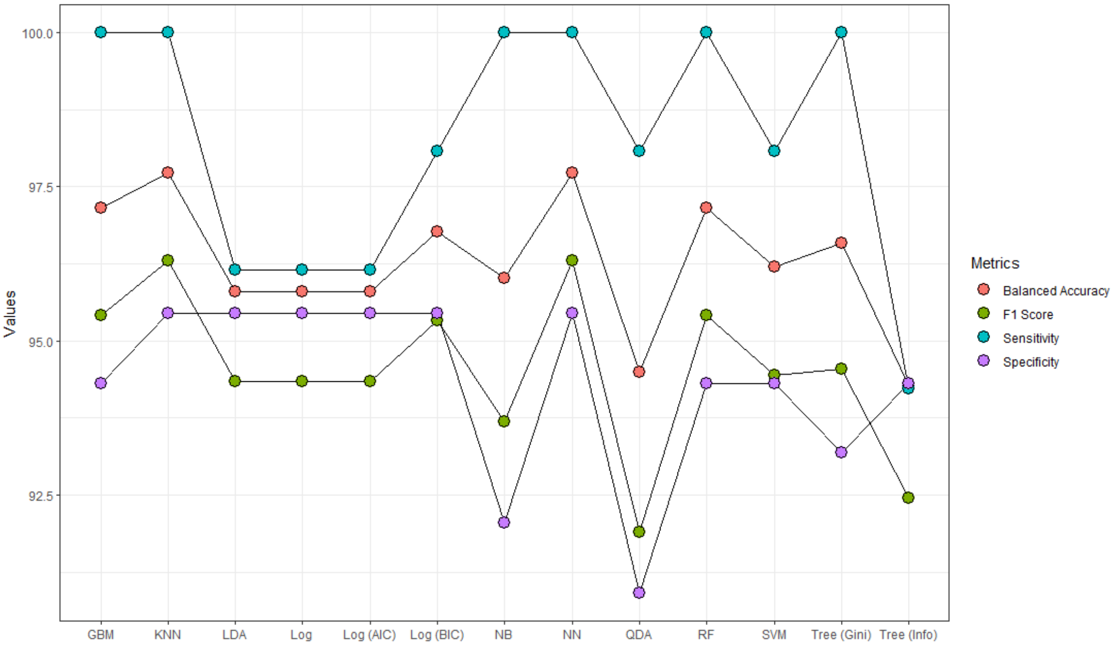


Figure 4.12: Performance Metrics for the Models Used

Table 4.2: Comparison of Performance Metrics (%) for the Methods

Method	F1 Score	Sensitivity	Specificity	Balanced Accuracy
Logistic (AIC)	94.34	96.15	95.45	95.80
Logistic (BIC)	95.33	98.08	95.45	96.77
Decision Tree (Info)	92.45	94.23	94.32	94.27
Decision Tree (Gini)	94.55	100	93.18	96.59
K-Nearest Neighbor	96.30	100	95.45	97.73
SVM	94.44	98.08	94.32	96.20
Gradient Boosting	95.41	100	94.32	97.16
Neural Network	96.30	100	95.45	97.73
Naive-Baye's	93.69	100	92.05	96.02
LDA	94.34	96.15	95.45	95.80
QDA	91.89	98.08	90.91	94.49
Random Forest	95.41	100	94.32	97.16

Out of those that actually have the Malignant tumor, the Random Forest, KNN, Decision Tree (based on Gini Index), Neural Network, Naive-Baye's and the Gradient Boosting perfectly classified all correctly with a sensitivity of 100%.

From the performance table, the logistic model based on AIC and BIC, KNN, neural network and LDA correctly classified a high proportion of Benign tumors out of the total Benign tumors in the testing data with a sensitivity value of 95.45%.

WIth regards to the F1 score and balanced accuracy, the KNN and neural network gave the same score with 96.30% (F1 score) and 97.73% (balanced accuracy). Overall, the KNN and neural network models gave the highest percentage of correct predictions.

The Figure below shows the confusion matrix for the best performing models.

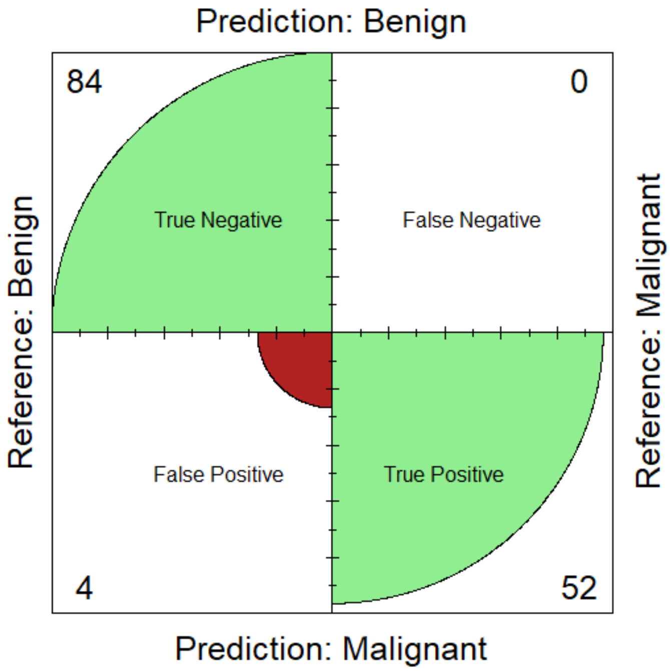


Figure 4.13: Confusion Matrix for the Best Model (KNN and Neural Network)

Based on the confusion matrix above, both models correctly predicted 84 Benign tumors out of the 88 and perfectly classified all the 52 Malignant tumor. The percentage of all correct predictions from these models is 97.14%.

The proportion of correctly classified Malignant over the total Malignant prediction made by these models is 100%. On the other hand, the proportion of correctly identified Benign over the total Benign prediction made by the models is 95.45%.

Therefore, the best models that are accurate to differentiate between the tumors are KNN and neural network.

CHAPTER 5
CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusion

In this project, breast cancer and ML were introduced as well as an in-depth literature review was performed on existing ML methods used for breast cancer detection. On the Wisconsin Breast Cancer Diagnostic data we applied nine main algorithms which are: SVM, Random Forests, Neural Network, Naive Bayes, Gradient Boosting, Logistic Regression, Decision Tree, K-NN, Discriminant Analysis and evaluated different results obtained based on confusion matrix, accuracy, sensitivity, precision and F1 score to identify the best machine learning algorithm that are precise, reliable and has the higher accuracy.

After an accurate comparison among the models, we found that K-Nearest Neighbor and Neural Network achieved a higher balanced accuracy of 97.73% and F1 Score of 96.30% and outperformed all other algorithms. In conclusion, these methods have demonstrated their efficiency in Breast Cancer prediction and diagnosis and achieve the best performance in terms of accuracy and precision.

5.2 Limitation and Recommendations

The fact that all of the results in this project are specific to the Wisconsin Breast Cancer database should be noted as a limitation of this work. It is therefore important to consider applying the same techniques in future works to other databases to validate the results obtained using this database.

Even though this project's results were positive, further in-depth study is still needed, particularly when it comes to the usage of other machine learning models that are similar to neural networks and KNN in function.

Lastly, it is important to do a comprehensive investigation even into the models that performed the best during the project.

BIBLIOGRAPHY

[1] A. Osareh and B. Shadgar, "Machine learning techniques to diagnose breast cancer," in *2010 5th international symposium on health informatics and bioinformatics*, pp. 114–120, IEEE, 2010.

[2] D. M. Parkin and L. M. Fernández, "Use of statistics to assess the global burden of breast cancer," *The breast journal*, vol. 12, pp. S70–S80, 2006.

[3] J. Ma and A. Jemal, "Breast cancer statistics," *Breast cancer metastasis and drug resistance*, pp. 1–18, 2013.

[4] R. L. Birdwell, D. M. Ikeda, K. F. O'Shaughnessy, and E. A. Sickles, "Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection," *Radiology*, vol. 219, no. 1, pp. 192–202, 2001.

[5] T. Kiyan and T. Yildirim, "Breast cancer diagnosis using statistical neural networks," *IU-Journal of Electrical & Electronics Engineering*, vol. 4, no. 2, pp. 1149–1153, 2004.

[6] I. Maglogiannis, E. Zafiropoulos, and I. Anagnostopoulos, "An intelligent system for automated breast cancer diagnosis and prognosis using svm based classifiers," *Applied intelligence*, vol. 30, no. 1, pp. 24–36, 2009.

[7] S. Sahran, A. Qasem, K. Omar, D. Albashih, A. Adam, S. N. H. S. Abdullah, A. Abdullah, R. I. Hussain, F. Ismail, N. Abdullah, *et al.*, "Machine learning methods for breast cancer diagnostic," *Breast Cancer and Surgery*, 2018.

[8] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial intelligence in medicine*, vol. 34, no. 2, pp. 113–127, 2005.

[9] M. Ngadi, A. Amine, H. Hachimi, and A. El-Attar, "A new optimal approach using nsvc for breast cancer diagnosis classification," *International Journal of Imaging and Robotics*, vol. 16, no. 4, pp. 24–36, 2016.

[10] M. U. Salma, "Fast modular artificial neural network for the classification of breast cancer data," in *Proceedings of the Third International Symposium on Women in Computing and Informatics*, pp. 66–72, 2015.

[11] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016.

[12] A. H. Osman, “An enhanced breast cancer diagnosis scheme based on two-step-svm technique,” *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 4, 2017.