VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Classification of Breast cancer on prediction of diagnosis

Abstract:

Breast cancer is the most common cancer in women around the world, and it is on the rise in developing countries, where the majority of cases are discovered late. Any advancement in cancer illness prediction and detection is critical for a healthy life. As a result, high accuracy in cancer prediction is critical for keeping patient treatment and survivability standards up to date. Furthermore, accurate benign tumour classification can save patients money by preventing them from receiving unnecessary therapy. Machine learning approaches have the potential to make a significant contribution to breast cancer prediction and early detection. On the Breast cancer Wisconsin Diagnostic dataset, we used five machine learning algorithms: Support Vector Classifier (SVC), Logistic Regression, Sequential Classifier, K-Nearest Neighbours classifier (KNN), and Decision Tree algorithm. After obtaining the results, we performed a performance evaluation and comparison between these different classifiers. The researchers used multiple functions for this algorithm and included extra features such as bagging and boosting to strengthen its usefulness. The results showed that the Sequential classifier obtained high accuracy, around 99 percent.

Keywords: Breast cancer classification, Breast cancer prediction, benign, malignant, KNN, Support Vector Classifier, Sequential Classifier, Logistic Regression, Decision Tree Classifier.

Introduction:

Many academics have recently stated that breast cancer has increased the mortality rate in women. According to the World Health Organization (WHO), over 627,000 women died in 2018. This group also expects that by 2030, the number will have risen to 2.7 million worldwide. The low survival rate is primarily due to the disease's late detection and complicated procedures. As a result, early identification of breast cancer is critical for reducing the risk of cancer spreading to other tissues and ensuring adequate treatment. Cancer is a growth of abnormal cells that arises from a genetic change in these cells and spreads throughout the body; late diagnosis and treatment result in mortality. Cancer cells in the blood or lymph system can migrate to other regions of the body, causing breast cancer to spread. Breast cancer is caused by DNA mutations and changes. There are several types of breast cancer, the most common of which are ductal carcinoma in situ (DCIS) and invasive carcinoma. There are numerous algorithms for categorizing breast cancer outcomes. Fatigue, headaches, pain, numbness (peripheral neuropathy), bone loss, and osteoporosis are all side symptoms of breast cancer. There are numerous methods for breast cancer classification and prediction. The performance of five classifiers is compared in this paper: SVC, Logistic

Regression, Sequential Classifier, kNN, and Decision Tree are some of the most used data mining algorithms. Mammography or a portable cancer diagnostic instrument can be used to detect it early during a screening test. Cancerous breast tissues change as the disease progresses, and this can be connected to the cancer stage. The stage of breast cancer (I–IV) indicates how far cancer has spread in a patient. Statistics such as a tumour, lymph node metastasis, and distant metastases, among other things, are used to identify stages. The research's purpose is to detect and categorize malignant and benign patients, as well as to figure out how to parametrize our classification techniques to reach high accuracy.

Literature Review:

1. According to Mohammed Amine Naji, Saana El Filali, Kawtar Aarika, El Habib Benlahmar, Rachida Ait Abdelouhahid, Olivier Debauche on Machine Learning algorithms for Breast cancer prediction and Diagnosis. In this paper they have used the Wisconsin Breast cancer dataset and calculate, Analyse, and evaluate different results acquired based on confusion matrix, accuracy, sensitivity, precision, and AUC to determine the best machine learning algorithms using SVM, Random Forests, Logistic Regression, Decision Tree, and K-NN algorithm that is exact, dependable, and finds the highest level of accuracy. After a thorough assessment of the models, they discovered that Support Vector Machine exceeds all other methods with 97.2 percent efficiency, 97.5 percent precision, and 96.6 percent AUC. Finally, the Support Vector Machine has proved its efficacy in the prediction and detection of breast cancer, achieving the highest accuracy and precision.

2. Apoorva V, Yogish H K, and Chayadevi M L proposed a study in 2021 on Breast cancer prediction using various Machine Learning and Neural Network algorithms such as K-Nearest Neighbor (KNN), Decision Tree (CART), Support Vector Machine (SVM), and Nave Bayes for numerical datasets and Convolution Neural Networks for image datasets. The proposed CNN beats estimations in recognizing and requesting breast cancer for picture datasets, according to their findings. In the analysis and prediction of cancer with numerical datasets, SVM has demonstrated to outperform CART, NB, and KNN.

3. Nikita Rane, Jean Sunny, Rucha Kanade, and Prof. Sulochana Devi presented a comparison study on Breast cancer utilizing machine learning algorithms employing various methodologies such as ensemble methods, data mining algorithms, and blood analysis in February 2020. On the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which is extracted from a digitized image of an MRI, they used various Machine learning algorithms such as Naive Bayes (NB), Random Forest (RT), Artificial Neural Networks (ANN), Nearest Neighbour (KNN), Support Vector Machine (SVM), and Decision Tree (DT). The data for the idea we offered was gathered through research on nine papers. We will be able to classify and forecast whether a cancer is benign or malignant using machine learning algorithms.

4. Ramik Rawal conducted a study on Breast Cancer Prediction Using Machine Learning in 2020. Various Machine learning algorithms such as SVM, Logistic Regression, Random Forest, and KNN that predict the breast cancer outcome have been compared in the article using diverse datasets with the goal of reducing error and increasing accuracy. SVM was able to demonstrate its power in terms of efficacy and efficiency based on accuracy and recall, according to Ramik.

Methodology:

Dataset: The dataset used in this study is taken from “kaggle.com”.The Dataset deals with Breast Cancer Diagnosis are Benign and Malignant. The dataset contains 12 attributes with 569 datapoints.

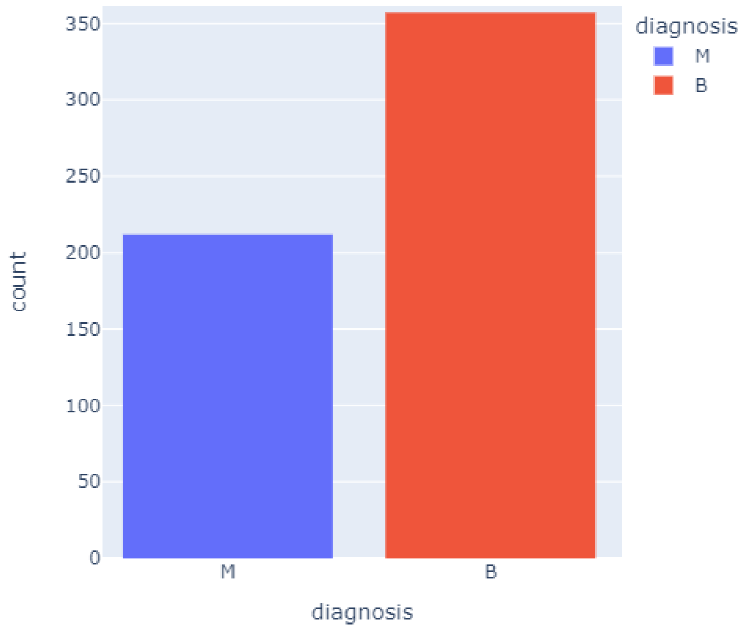
- **Benign tissue**: It is possible that a benign tumor is harmless (not cancerous). Unless it presses on a nearby structure or causes other symptoms, a benign tumor is usually not a major issue. The boundaries of a benign tumor are smooth and regular. Although a benign tumor can grow to be fairly large, it will not infiltrate surrounding tissue or spread to other parts of your body.
- **Malignant Tissue:** A malignant tumor can be cancerous. A malignant tumor has uneven borders and grows quicker than a benign tumor. It's possible for a malignant tumor to spread to other sections of your body. The medical word for this form of spread is metastasis. Metastasis can occur anywhere in the body, but it is most common in the liver, lungs, brain, and bone.

Plots:

Univariate Analysis :

In this we have done univariate analysis that which is simplest form of analysing data. It analyses each variable separately.

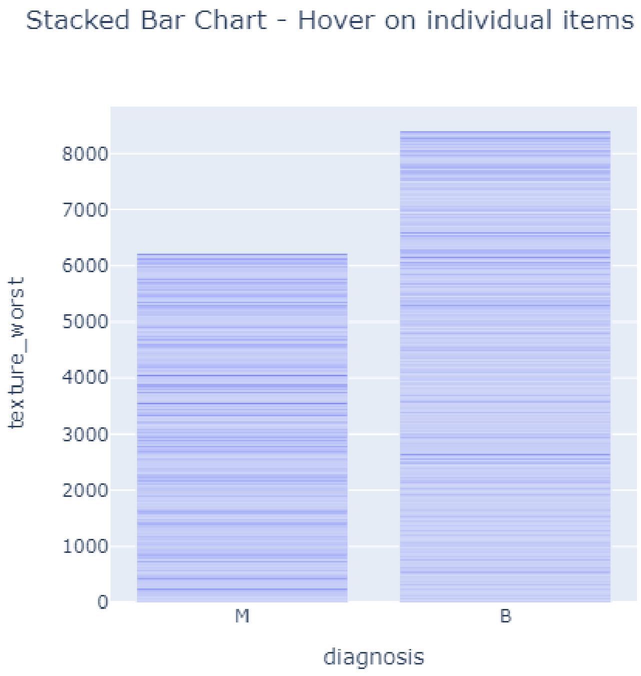
1. **Histogram:**



Inference: In this we have Benign and Malignant diagnoses. From the univariate analysis graph, we have concluded that Benign is more diagnosed than Malignant

2. **Staked bar chart:**

   Staked bar chart that uses bars to show the comparisons between categories of the data. Here we are comparing the texture worst and diagnosis as x and y values.

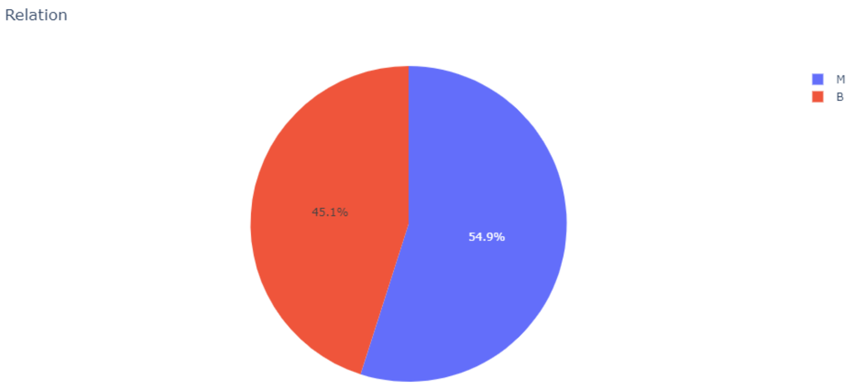


Inference:

In the graph we have taken x as diagnosis and y as texture worst. Here M consider as Malignant and B consider as Benign. We have compared diagnosis and texture worst. From that graph we have concluded as texture worst for benign is higher than texture worst for malignant.

3. Pie Chart:

In this, we have checked the relation between diagnosis and compactness worst through a pie chart



Inference:

In this, we have taken that malignant as brown and benign as blue. From the above graph, we got the percentage of compactness worst is highest for malignant (54.9%) than benign (45.1%).

Multivariate Analysis:

For the graphs, we have done a multivariate analysis, in which multiple measurements are made, this is one of the most useful methods to determine relationships and analyze patterns among large sets of data.

Fig (1)

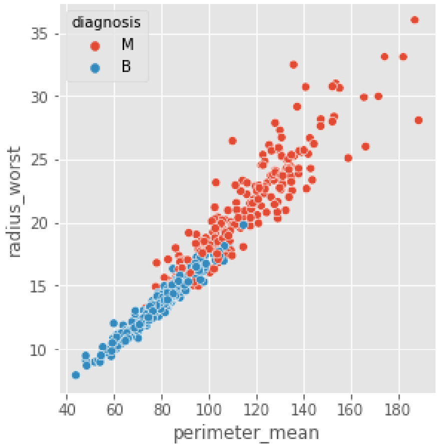


Fig (2)

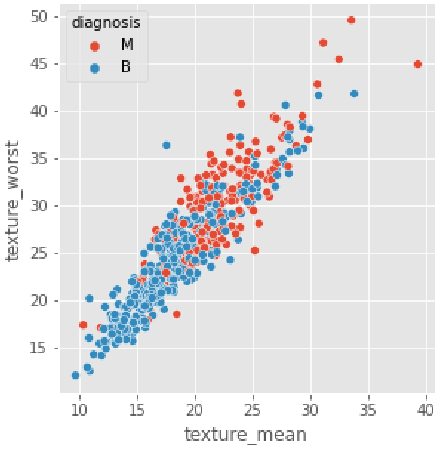


Fig (3)

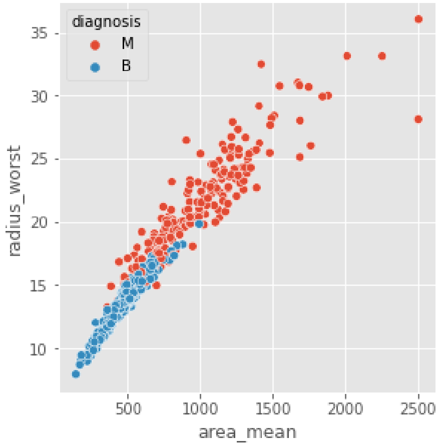
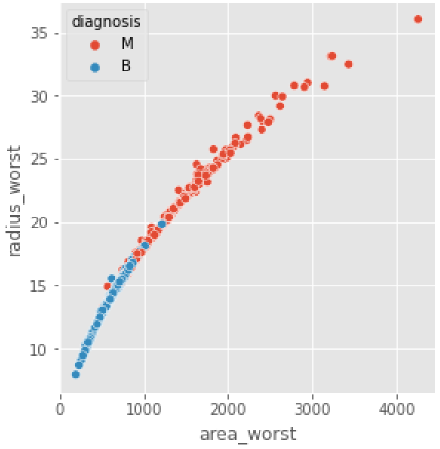


Fig (4)



Inference:

In this we have done a positive correlation that is, it shows the relationship between two variables that move in the same direction

From fig1:

From fig1, we compared the perimeter mean and radius worst**.** This graph shows perimeter mean and radius worst is higher for malignant.

From fig2:

From fig2, compared the texture mean and texture worst after comparing the graph texture mean and texture worst is lower for benign than the malignant.

From fig3:

From fig3, we have compared the radius worst and the area mean**.** From that we got area mean is highest for malignant than the benign

From fig4:

From fig4, we have compared radius worst and area worst form that we got area worst is lower for benign more than the malignant.

1. Plot:

For this graph, we have done a negative correlation, in this, it shows the relationship between two variables such that as the value of one variable increases, the other value decreases.

Fig (1)

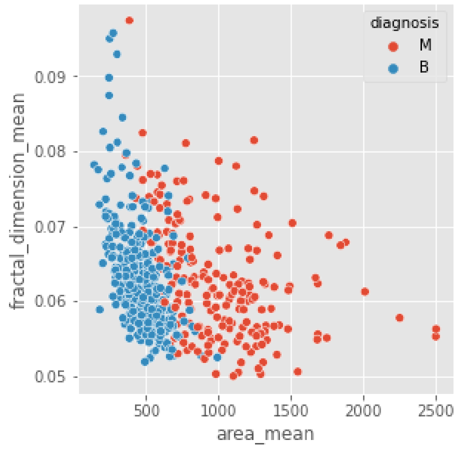


Fig (2)

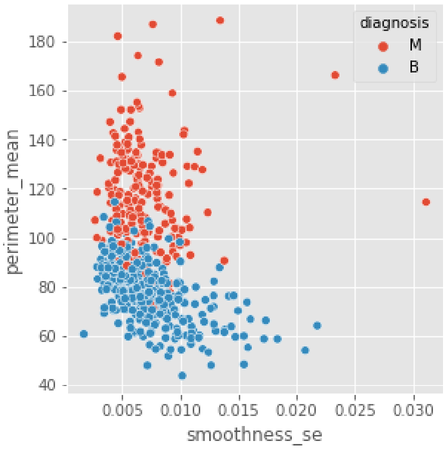


Fig (3)

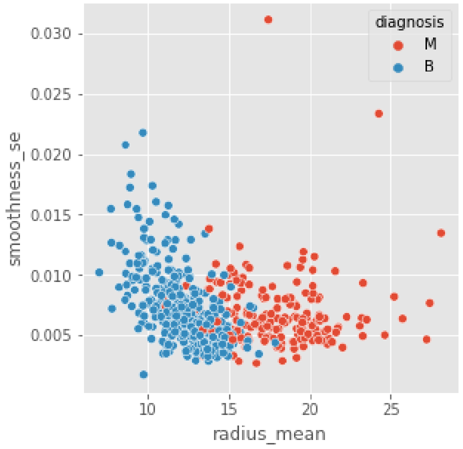
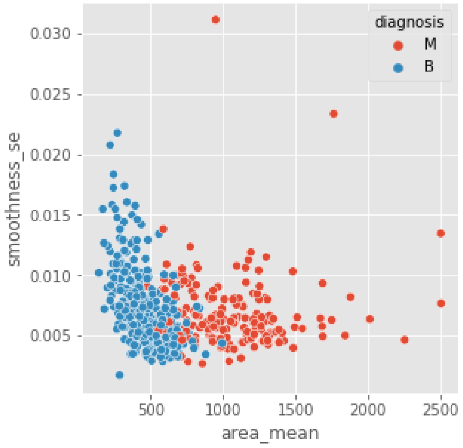


Fig (4)



Inference:

From fig 1:

In this, we have compared the area mean and fractal dimension mean from that graph we conclude that the area mean and fractal dimension is higher for malignant.

From fig 2:

In this, we have compared smoothness se and perimeter mean in this benign is lower than malignant.

From fig 3:

From fig3, in this, we have compared radius mean and smoothness se is higher for malignant.

From fig 4:

From fig5 we have compared area mean and smoothness From that, we have concluded malignant is higher.

Classification:

The act of classification concepts and things involves identifying, comprehending, and organizing them into predetermined classes or "sub-populations." machine learning systems classify future datasets into categories using pre-categorized training datasets and a range of techniques. In machine learning, classification algorithms utilize input training data to predict whether following data will fall into one of the established categories.

There are five most common classification algorithms:

- logistic regression
- decision tree
- k-nearest neighbours
- support vector machines
- Sequential classification

In this paper we are going to classify the breast cancer for prediction of diagnosis

I. Logistic regression:

A technique called logistic regression is used to predict a binary outcome: either something happens, or nothing happened. This can take the form of yes/no, pass/fail, alive/dead, and so on. The binary outcome is determined by analysing independent factors, with the findings falling into one of two groups. Although the independent variables might be numeric or categorical, the dependent variable is always categorical. Logistic regression is a classification problem-solving approach that uses predictive analysis and it comes under supervised learning technique.

Logistic regression is a popular classifier technique, the probability idea is used in a machine learning method. This regression employs a sophisticated cost function known as the "sigmoid function," which is also known as the "logistic function." the probability of a logistic regression hypothesis is between 0 and 1. To convert predictive values to probability, the sigmoid function is used. In the medical field the curve of logistic function informs us whether the cells occur cancer or not. It forms as curve in sigmoid function. All the dependent variables must be categorical in nature.

$$f(x) = \frac{1}{1+e^{-x}}$$

$f(x)$ output lies between 0 and 1

x is the input to the function

II. Decision tree:

Decision tree is a type of supervised learning technique that can be used for classification and regression problems, but mostly it is preferred using for solving classification problems. There are two nodes, they are the decision node and leaf node. The objective of decision nodes is to to make any decision and it can have multiple branches and the objective of leaf nodes are to get the output of decisions that are made by decision node and they do not contain any further branches.the decisions or the test are performed on the basis of features of the given dataset. The process of seperating the decision node/root node into sub-nodes based on the provided circumstances is known as splitting.pick a best attribute and split the dataset using that attribute.the main algorithm decision tree follows is id3 using top down approach and the other important step is calculating information gain.decison tree is very easy to use and the outliers are avoided.this algorithm can make use of categeorical as well as numerical data.

entropy:

$$E = \sum_{i=1}^{N} - p_i \, log_2 p_i$$

n=dataset with n classes,

$p_i = probability\ of\ randomly\ selected\ example$

information gain= 1- entropy

gini index: $1 - \sum_{i=1}^{N} (p_i)^2$

III. K nearest neighbor:

K nearest neighbor is the supervised learning approach and one of the simplest classification algorithms. K-nn assigns a classification to a new datapoint based on the similarity of the stored data, this means that fresh data gets sorted into distinct sets using knn algorithms. The knn approach can be used for both regression and classification, but it is more typically utilized for classification tasks. Knn is a non-parametric algorithm since it makes no assumptions about the underlying data. It is a lazy learner algorithm as it does not learn from training set, instead it saves the dataset and uses it at the time of classification. During the training phase, the knn algorithm simply saves the dataset, and as it receives new data, it updates the dataset.

IV. Support vector classifier:

Support vector classifier is a supervised machine learning algorithm and it is a dependable algorithm where it works only with limited amount of data. The main advantage is it has higher speed and performance. It takes the dataset as input and draws the decision boundary by classifying them as two binary targets. It separates the two data with the help of hyperplane. Kernel function is used, the main objective is to find a hyperplane in “n” dimensional space that classifies the datapoints. It is a binary classification problem and in python we import the dataset and the algorithm and plot them to visualize the results and check the linear line that separates the datapoints and find the accuracy of the model. It is a powerful algorithm.

V. Sequential classification:

Pattern recognition may be approached in a number of ways, one of which is sequential categorization. Sequential (multistage) classifiers are designed to break down a difficult decision into smaller chunks.a compilation of a number of simpler decisions typically, such classifiers are expressed as decision tree.one of their key advantages is that various nodes of the tree have varied properties.objects are evaluated against subsets of classes and, in most cases, various subsets of classes.the characteristics are utilised. It's challenging to get both a high identification rate and a short tree size while building these classifiers. The majority of decision tree building techniques are top-down methods, in which classes are sorted into groups at each stage. The design of a decision tree in these techniques is as following tasks: determining whether a node is tenninal, determining the decision rule for each node or not, and lastly, providing a class label to each tenninal node. The most typical technique to formulate the decision rule is to consider a variety of alternative splits and select the best one in some way. Entropy reduction, gini index, and other split quality indicators are well-known. To complete this work, we have chosen to employ the fisher criteria.

The fisher criteria, a well-known feature extraction approach for two classes of objects with normal distributions, is a well-known linear discriminant analysis method. For a variety of classes, certain expansions of the criteria have been proposed.

Proposed model:

This paper proposes a study to analyse breast cancer diagnosis which is malignant and benign based on the accuracy of machine learning algorithm. The proposed study starts with collection of datasets that targets breast cancer. The data is pre-processed, split into train data and test data and classified using machine learning algorithm. The dataset is visualized by plotting graphs with differentiable points the test result is predicted. A confusion matrix is created to check the accuracy of the result.

Performance evaluation:

1. **Confusion matrix:**

Confusion matrix is basically a performance evaluation process in machine learning algorithms and the output of the confusion matrix can be of two or more classes. It is a combination of actual values and predicted values. It is mainly used for measuring accuracy. The actual values contain 2 classes called as true and false whereas the predicted values contain positive and negative classes. The important terms are true positive (tp) which is malignant , true negative (tn), false positive (fp), false negative (fn) which is benign.

$$[TP\ FP\ FN\ TN\ ]$$

2. Accuracy score:

Accuracy score is also a performance evaluation metric in machine learning where it divides number of correct predictions to the total number of input samples to obtain a percentage between 1 to 100.

$$\text{accuracy}=\frac{TP+TN}{TP+TN+FP+FN}$$

3. Classification report:

Classification score is also a performance evaluation metric in machine learning. It mainly provides the precision, recall, f1 score and support score of the classifier algorithm. It checks whether the predictions are right or wrong and also checks percentage of correct prediction (precision). It also checks the ability of an algorithm whether it correctly fins the positive values (recall). It also deals with the weighted harmonic mean of precision and recall which is called f1 score. Support score deals with the actual occurrences of a specific class in a specified dataset.

Result and discussion:

Method	Accuracy
logistic regression	0.9649122807017544 96%
decision tree	0.9298245614035088 93%
k-nearest neighbours	0.956140350877193 0.96%
support vector machines	0.9649122807017544 96%
Sequential classification	0.9883040935672515 99%

Breast cancer is the one most common cancer among women over the age of 50 and it takes a long time in the diagnosis process. Machine learning algorithm is helpful in the process of diagnosis, and it provides effective results. In this paper we are analysing breast cancer using promising machine learning algorithms such as logistic regression, decision tree, k-nearest neighbour, support vector classifier and sequential classification. This paper focuses on predicting the performance of machine learning algorithms with the help of accuracy. The result shows that sequential classification shows the highest accuracy. Support vector machines and logistic regression show the equal and second highest accuracy. K-nearest neighbours is also almost nearest to the logistic regression and support vector machines. Decision trees shows low accuracy than any other classifier.