BREAST CANCER CLASSIFICATION

ABSTRACT

According to global cancer data, more than 2 million women are diagnosed with breast cancer annually, accounting for the bulk of all new cancer cases and fatalities associated with them. This raises serious public health concerns. Fortunately, it is a cancer that can be cured when it is early on. The prognosis and survival of patients with breast cancer are improved by early diagnosis and prompt access to effective treatment. There are considerable chances of inaccuracy and false diagnosis when classifying malignancies, which must be addressed. Correct classification can spare patients from needless procedures. Therefore, it's critical to correctly diagnose patients and divide them into benign and malignant groups. The goal of this study, which is based on machine learning (ML) methods, is to review Python methodology and its use in the diagnosis and prognosis of breast cancer by developing a straightforward machine learning model. Machine learning has a distinct edge since it can identify important traits in challenging breast cancer datasets. The technique is frequently used for pattern classification and forecast modeling. The main information used in this investigation was taken from the Wisconsin breast cancer database (WBCD). It is a benchmark database that compares outcomes using various techniques.

TABLE OF CONTENTS

List of figures

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Breast cancer (BC) is one of the most common malignancies in women. Early diagnosis of Breast Cancer and metastasis among the patients based on an accurate system can increase survival of the patients to >86%. Total number of women dying in 2021 is approximately 963,000, according to the World Health Organization (WHO). Still, the organization predicts that the number could reach 2.9 million globally. Breast cancer can occur in women and rarely in men.

Cancer is the creation of abnormal cells that come into these cells genetically and mutated. Spreads throughout the body, leading to death in diagnosis and treatment. Breast cancer is a type of cancer that starts in the breast cells, the fatty tissue or the fibrous connective tissue within the breast. This disease infects the women's chest and specifically glands and milk ducts, the spread of breast cancer to other organs is frequent and could be through the bloodstream. There are two types of breast cancer, Malignant and Benign. The first is classified as harmful, has the ability to infect other organs and is cancerous, Benign is classified as non-cancerous. Different techniques are used to capture breast cancer such as Ultrasound Sonography, Computerized Thermography, Biopsy (Histological images).

Machine learning and Data mining techniques are straightforward and effective ways to understand and predict data. Because of its unique advantages in critical features detection from complex breast cancer datasets, machine learning (ML) has been applied in medical diagnosis in a large number of applications and is widely recognized as the methodology of choice in breast cancer pattern classification and forecast modeling.

Here in this project, we have used the Wisconsin Breast Cancer Dataset (WBCD), which is a widely used dataset provided by UC Irvine machine learning repository. We will then compare the classification algorithms using an assembled approach suitable for demonstration and direct interpretation of their results. We have used seven traditional methods including Logistic Regression, Nearest Neighbor (k-NN), Support Vector Machines (SVM), Kernel Support Vector Machines (KSVM), Naive Bayes, Decision Tree and Random Forest Algorithm, and a machine learning method to predict whether the case is benign or malignant.

1.2 PROBLEM DEFINITION

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. Thus, the correct diagnosis of breast cancer and classification of patients into malignant or benign groups is the subject of much research. Radiologist examines and analyzes himself and then he / she decides the result after participating with other experts. This process takes time and the results depend on the knowledge and experience of the staff. In addition, experts are not available in every field of the world. Therefore, the research community proposed an automatic system called CAD (Computer Aided Diagnosis) for better classification of tumors, accurate results and faster Implementation without the need for radiologists or specialists.

Hence, our project aims to build a predictive system, which can predict cancer as benign or Malignant. Various Machine Learning algorithms are implemented and compared. Model with highest accuracy has been considered for the predictive system.

CHAPTER 2
LITERATURE SURVEY

2. LITERATURE SURVEY

Breast cancer prediction system based on a hybrid approach was published by Lavanya et al. it utilized a classification and regression trees (CART) classifier with feature selection and bagging technique for increased classification accuracy and improved diagnosis. They combined the hybrid approach with feature selection to eliminate irrelevant variables that had no bearing on the classification job and improve the accuracy of breast cancer classification. The Bagging signifies that the data was accurately categorised using Bootstrap aggregation. Data were gathered from the UCI machine learning repository where three breast cancer datasets (Breast Cancer, Breast Cancer Wisconsin (original), and Breast Cancer Wisconsin) were experimented with (diagnostic). The Original Dataset had 699 Instances and 11 Attributes; the Breast Cancer Dataset had 286 Instances and 10 Attributes. While the Diagnostic Dataset contained 569 Instances and 32 Attributes, all previews dataset with two classes.

Ebru Aydndag Bayrak and colleagues from the Department of Computer Engineering at Istanbul University in Turkey talked about two well-liked machine learning methods for classifying Wisconsin breast cancer. SVM (Sequential Minimal Optimization Algorithm) demonstrated the best performance in the accuracy of 96, 9957 percent for the diagnosis and prediction from the WBC dataset, according to the performance metrics of the applied machine learning approaches.

In a review article titled Applications of Machine Learning in Cancer Prediction and Prognosis, Joseph A. and colleagues found several trends in the types of machine learning techniques applied, the types of training data integrated, the types of endpoint predictions made, the types of cancers studied, and the general effectiveness of these techniques in predicting cancer susceptibility or outcomes. They came to the conclusion that the usage of machine learning classifier will probably become much more widespread in many clinical and medical settings if the calibre of investigations keeps rising.

In a review article titled Applications of Machine Learning in Cancer Prediction and Prognosis, Joseph A. and colleagues found several trends in the types of machine learning techniques applied, the types of training data integrated, the types of endpoint predictions made, the types of cancers studied, and the general effectiveness of these techniques in predicting cancer susceptibility or outcomes. They came to the conclusion that the usage of machine learning classifier will probably become much more widespread in many clinical and medical settings if the calibre of investigations keeps rising.

Utilizing five modelling methods with Greedy Search and K-fold Cross-validation, Meraryslan Meraliyev and others have worked on problems with breast cancer prediction and developed solutions. A number of algorithms, including support vector machines (SVM), decision tree classifiers, logistic regression, and neural networks, were taken into consideration. The outcomes of the modelling demonstrated that the most effective algorithms for predicting breast cancer are SVM and KNN.

In their research on Breast Cancer Diagnosis by Different Machine Learning Methods Using Blood Analysis, Muhammet Fatih Aslan and others. In terms of feature type, this dataset was unique from previous datasets. Machine learning) was used to study the importance of data in the identification of breast cancer (learning) strategies. Four different ML (machine learning) techniques were used for the analysis. Extreme Learning Machine (ELM) interfaces for Artificial Neural Networks (ANN) were developed. Additionally, using the hyper parameter optimization methodology, the hyper parameter values with the lowest error rates for the Artificial Neural

Networks (ANN), Extreme Learning Machine (ELM), KNearest Neighbor (KNN), and SVM (support vector machines) methods were identified. According to these values, accuracy rates and training times were obtained. The results showed that Standard Extreme Learning Machine had the highest accuracy rate and the shortest training time (ELM). They demonstrated that when there are many samples, using the Standard Extreme Learning Machine (ELM) is more favourable in terms of time.

In order to detect breast cancer, Abien Fred M. Agarap applied machine learning algorithms to the Wisconsin Diagnostic Dataset. Six machine learning techniques are utilized in this article to find cancer. On the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, GRU- SVM, Linear Regression, Multilayer Perceptron (MLP), Nearest Neighbor (NN) search, Softmax Regression, and Support Vector Machine (SVM) are used for the diagnosis of breast cancer. These models are evaluated for classification test accuracy as well as sensitivity and specificity values. The aforementioned dataset is made up of features that were calculated from digitalized images of FNA tests performed on a breast mass. The dataset was divided into two parts for the sake of ML algorithm implementation: 70% for training and 30% for testing. According to their findings, all of the ML algorithms given performed well when categorizing carcinomas into benign or malignant tumors, which is a binary classification problem. In light of this, the statistical methods used to solve the categorization problem were likewise successful. A CV method like k-fold crossvalidation should be utilized to further validate the findings of this study. The application of such a method would not only give a more precise measurement of model prediction performance, but it will also help identify the most important hyper-parameters for ML algorithms.

CHAPTER 3 SCOPE OF THE PROJECT

3. SCOPE OF THE PROJECT

The project is completely developed on software grounds using machine learning techniques. All the codes are written using python language as it is best fit for machine learning and Artificial intelligence based projects. It is simple and consistent, easy to understand and code, flexible, platform independent and has access to various libraries and frameworks for machine learning and also has a worldwide community. The major advantage of using machine learning algorithms is their ability to improve over time and can never be outdated. This is a re-engineered project as in the previous works, many models have been proposed which use different feature sets and methods of machine learning to diagnose breast cancer. The scarcity of large datasets and inequality between negative and positive classes are the main challenges in the research area of breast cancer prediction. Hence have major applications in the reliability aspect.

CHAPTER 4
METHODOLOGY
AND
IMPLEMENTATION

4.1 BLOCK DIAGRAM

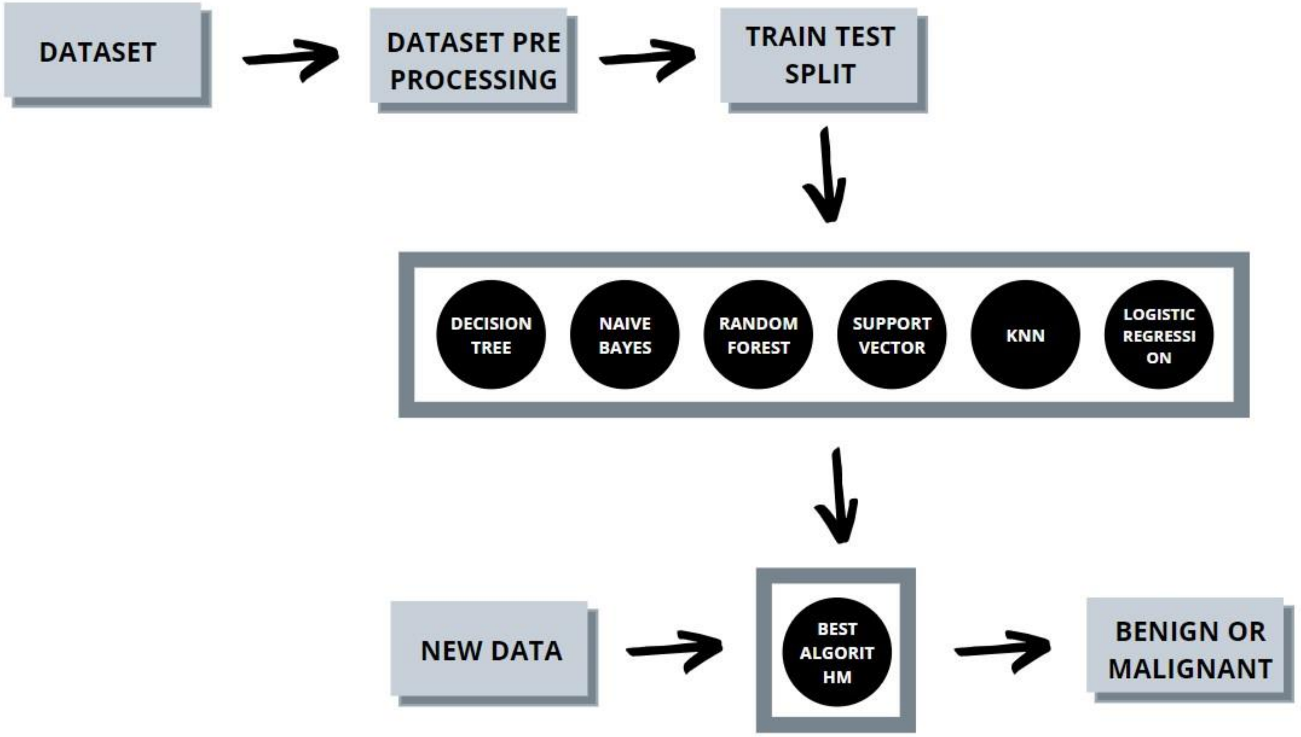


Figure 1: Block diagram of project

The first step here is to collect the information, so we gather the information from the Wisconsin dataset. Then we process the data as the raw data cannot be fed directly to the machine learning model.The data preprocessing steps are covered at this stage. Once the data is preprocessed, the dataset is split into two groups, that is training data and testing data. Training data is used to train the algorithms used in this model and testing data is used to test the performance of individual algorithms. Around 80% of the data is used for training and 20% of the data is used for testing the algorithms. The algorithms chosen here are binary classifiers since the required output is ‘0’ or ‘1’. Using the accuracy score the best algorithm is chosen. To which the new data is entered and based on the training of the algorithm, the model predicts the output to be either ‘Malignant’ or ‘Benign’.

4.2. TOOLS AND LIBRARIES USED

We have used Google Colaboratory for executing our project in Python. Following tools and libraries were also used.

1. GOOGLE COLABORATORY: Colab is a cloud-based Jupyter notebook environment that is free to use. Most importantly, it doesn't require any setup, and the notebooks you create can be modified simultaneously by your team members, much like Google Docs projects. Many common machine learning libraries are supported by Colab and can be quickly loaded into your notebook.

2. PYTHON: Python is a powerful programming language that comes in helpful when dealing with statistical problems that need the use of machine learning techniques. It has several utility functions that help with pre-processing. Processing is rapid and can be done on almost any platform. It's easy to connect with C++ and other image libraries, and it includes built-in techniques and libraries for storing and manipulating various forms of data. It includes the pandas and numpy frameworks for data manipulation.

3. NUMPY: This is a Python library for computation. It provides multidimensional array objects and utilities with outstanding performance. TensorFlow receives its input via Numpy. Numpy connects with a wide range of databases quickly and effortlessly.

4. PANDAS: Pandas is based on two key Python libraries: matplotlib for data visualisation and NumPy for arithmetic operations. Pandas acts as a wrapper around these libraries, allowing you to use fewer lines of code to access many of matplotlib's and NumPy's methods. Pandas'.plot(), for example, combines numerous matplotlib methods into a single method, allowing you to plot a chart in only a few lines.

5. SCIKIT-LEARN: It's a Python machine learning library. It includes matplotlib, numpy, and a large number of algorithms. The API is user-friendly and simple to comprehend. It has a number of functions for data analysis and graphing. To generate a solid feature set, many of its feature reduction, feature importance, and feature selection capabilities can be applied. Its algorithm can be used to tackle problems related to classification and regression, as well as their subclasses.

4.3 DATASET DESCRIPTION

The breast cancer dataset we used from the UCI Machine Learning Repository is Wisconsin dataset, which is a frequently used dataset in the study. Dr. William H. Wolberg created it at the University Of Wisconsin Hospital in Madison, Wisconsin, in the United States. The dataset was created by Dr. Wolberg utilizing fluid samples from patients who had solid breast masses and Xcyt, an intuitive graphical program that can examine cytological parameters based on digital scans. The computer extracts 10 features using a curve-fitting method from each sample cell, computes the mean value, the extreme value, and the standard error of each feature for the image, and then outputs a 30 real-valued vector.

A fine needle aspirate (FNA) of a breast lump is used to scan the image and compute the features. They outline the features of the cell nuclei visible in the photograph. Here, 357 benign and 212 malignant tumours from 569 patients were used for analysis, with 30 attributes for each instance including the diagnosis and features.

Attribute Information:

1) ID number

2) Diagnosis (M = malignant, B = benign)

Ten real-valued features are computed for each cell nucleus:

a) Radius (mean of distances from center to points on the perimeter)

b) Texture (standard deviation of gray-scale values)

c) Perimeter

d) Area

e) Smoothness (local variation in radius lengths)

f) Compactness (perimeter^2 / area - 1.0)

g) Concavity (severity of concave portions of the contour)

h) Concave points (number of concave portions of the contour)

i) Symmetry

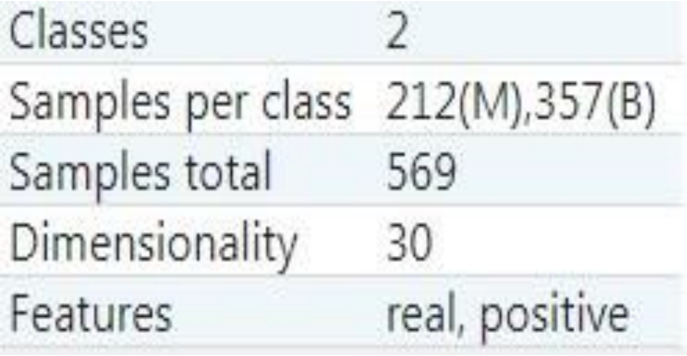
j) Fractal dimension

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, and field 23 is Worst Radius.

All feature values are recorded with four significant digits.

Missing attribute values: none

Class distribution: 357 benign, 212 malignant



Classes	2
Samples per class	212(M),357(B)
Samples total	569
Dimensionality	30
Features	real, positive

Figure 2: Data distribution

4.4. DATA PREPARATION AND CLEANING

The programming language used in our project is ‘Python’, implemented in google collaboratory. To begin with, data exploration and cleaning is done via following steps.

1) Import libraries: Required libraries viz pandas, numpy, pyplot, seaborn, matplotlib, and sklearn datasets are imported.

2) Load data: Breast cancer dataset has been loaded from sklearn datasets.

3) Data exploration:

● We examined the data set using the pandas’ head() and tail() method.

● Found the dimensions of the data set using the panda dataset ‘shape’ attribute. There are 569 rows and 31 columns which represent 569 patients with 31 data points or features for individual patient in this dataset.

● info( ) method is done to check the number of non-null values and datatypes of the column. In case of any categorical data, the data is changed in dummy numeric.

● This is done because categorical data will mislead the interpretation of data. The data type of columns is listed by using pandas dtypes.

● Data sets are not perfect and DataFrame from the dataset might contain some missing values in any column or row which can mislead the interpretation. For every missing value, pandas add NaN in its place. So, fetch the count of all the columns and rows from the dataset that contain empty values viz. NaN, NAN or na. NaN occurs in case of any missing value.

● Describe() method is used to generate statistics that summarize the central tendency, dispersion, and shape of the data frame excluding null values.

4) Count number of patients with malignant and benign tumor: The number of patients diagnosed with either malignant (M) or benign (B) tumor is counted by using pandas value_counts function and is visualized using countplot as shown in fig 11 and 12 respectively.

5) Dropping or addition of column/row: To gather the similar features label column is dropped and is stored in another variable. Which is followed by obtaining the statistical measures and correlation of the new data. Hence number of columns dropped to 30.

4.5. DATA VISUALIZATION

Data visualization provides a good, organized pictorial representation of the data which makes it easier to understand, observe, and analyze. Visualization of data is an imperative aspect of data science and machine learning projects. It helps to understand data and also to explain the data to another person. Python has several interesting visualization libraries such as Matplotlib, Seaborn, etc. In our project, we used panda's visualization which is built on top of matplotlib and seaborn library to find the data distribution of the features.

Visualization makes it easier to detect patterns, trends, and outliers, and provides clear, better and reliable result. It is implemented in this project by following visualization methods.

a) COUNTER PLOT: It is used to show the counts of observations in each categorical bin using bars. The plot will represent the class distribution of diagnosed malignant and benign patients. Here we have 212 malignant diagnosed patients i.e. around 38% of the data and, 357 i.e. 62% of patients diagnosed with benign tumors.

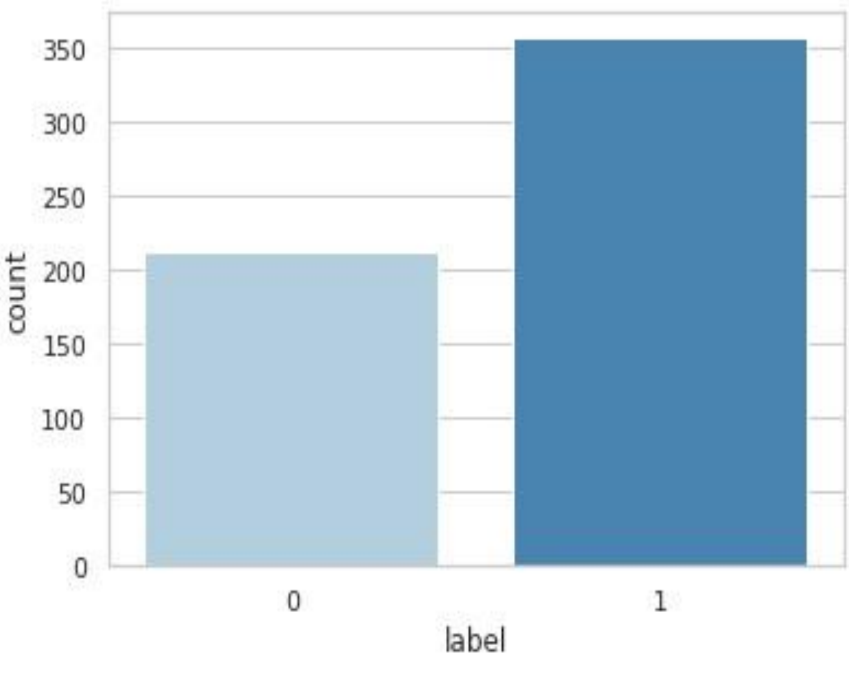


Figure 3: Count plot

Count plot visualization of above data in python is done by using seaborn countplot function.

In the below counterplot max samples mean radius is equal to 1.

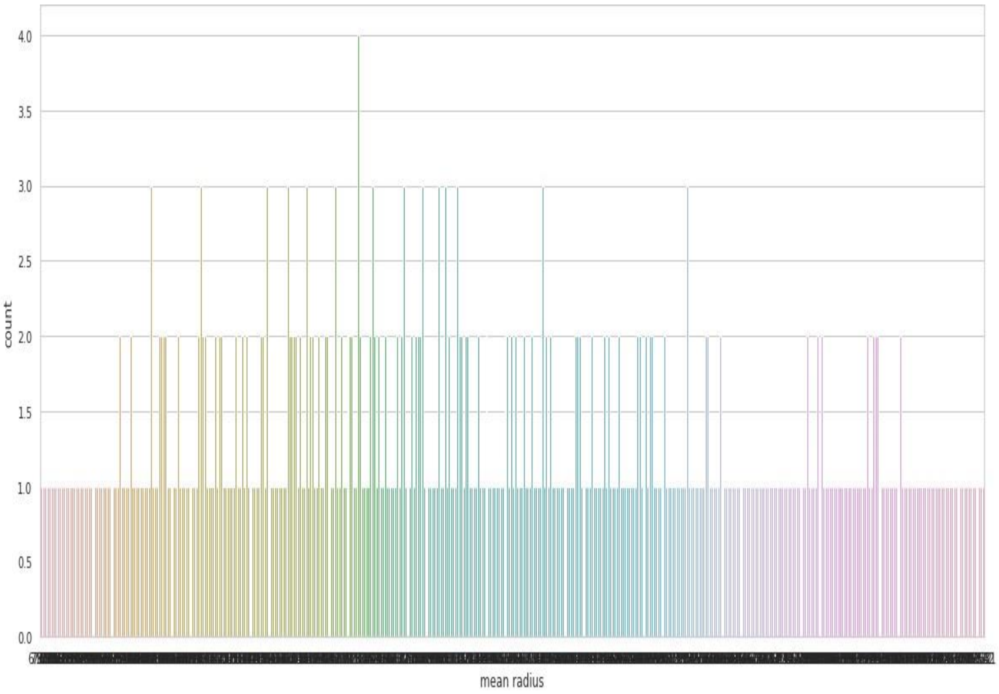


Figure 4: Counter plot of mean radius

b) **PAIR PLOT:** A pair plot is used to better visualize the data since the dataset contains numerous variables and the relationships between each and every variable need to be examined. It displays the data as a group of points. The values of two variables are matched based on where they are located in the same data row. Each value represents a location on the vertical or horizontal axis, indicating the correlation between the values. Both distribution of a single variable and relationships between two variables are permitted. It is a useful technique for locating patterns for examination. Python uses seaborn and creates pair plots using the seaborn pairplot method.

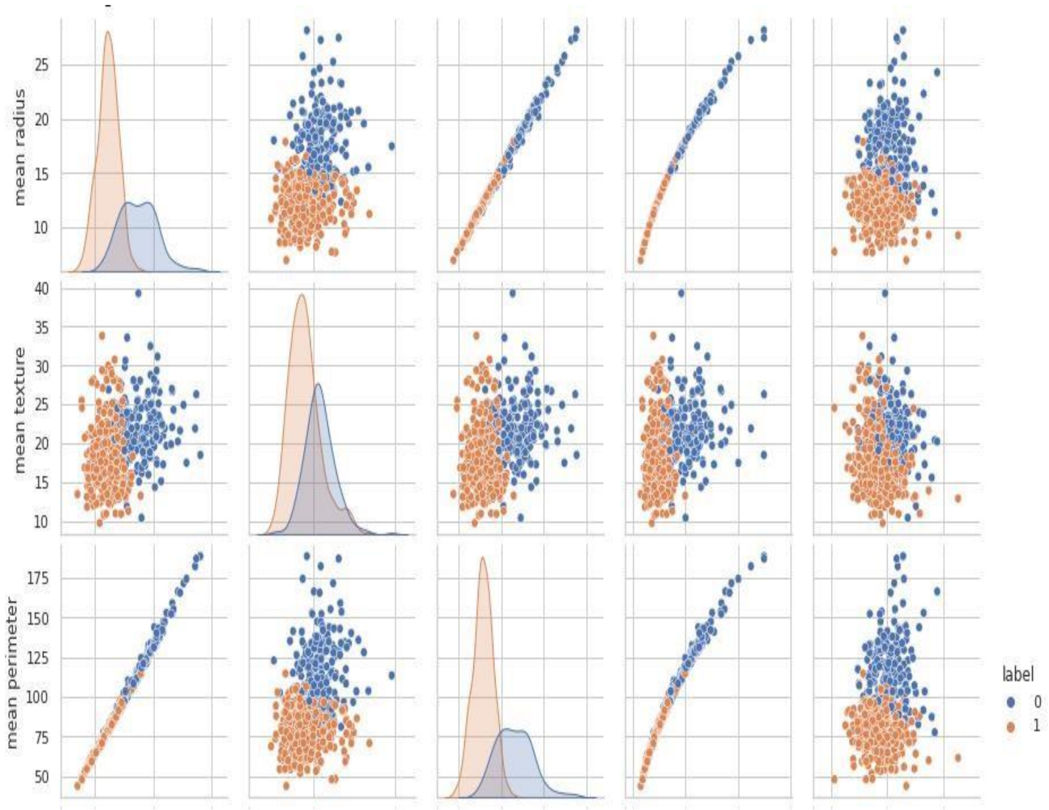


Fig.15 Pair plot

The pair plot shows malignant and benign tumor data distributed in two classes. It is easy to differentiate in the pair plot.

c) **SCATTER PLOT:** A set of dots plotted on a horizontal and vertical axis is known as a scatter plot. Because they may demonstrate the degree of connection, if any, between the values of observed quantities or events, scatter plots are crucial in statistics (called variables).

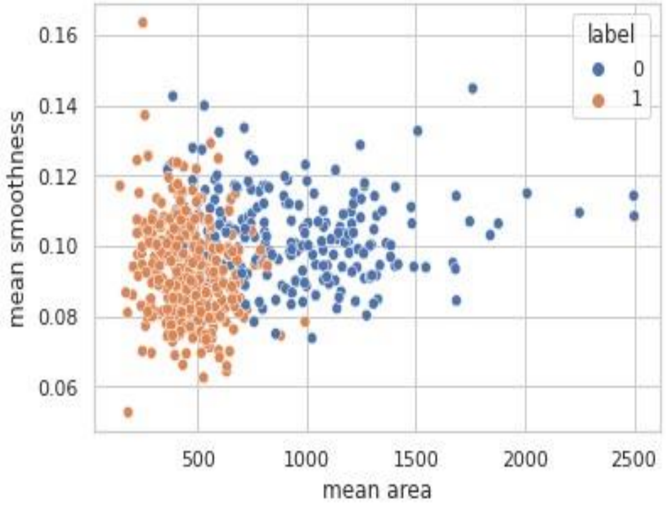


Figure 6: Scatter plot

d) HEATMAP: A heatmap (or heat map) is a graphical representation of data where values are depicted by color. The worth of several features can be seen in the heatmap below. Mean area and worst area are valued higher than other features, whereas mean perimeter, area error, and worst perimeter are valued slightly lower but higher than other features.

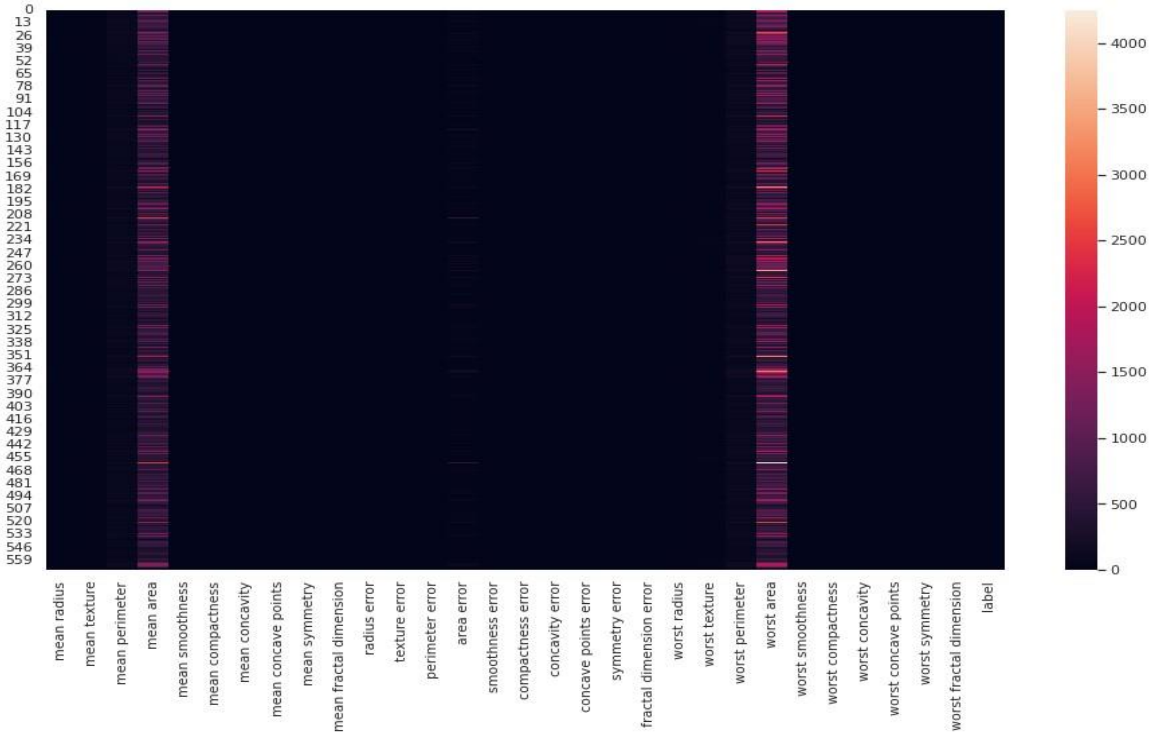


Figure 7: Heat map of data frame

The correlation analysis of the columns with respect to the pair plot is shown, which is useful for analysing the relationship between the attributes of each patient individually.

Heatmaps make correlation visualisation more dependable and simple. Because numerical data may be difficult to understand, using colours improves visualization. A heatmap gives a clear visual summary of the data. Therefore, using a correlation matrix, a heatmap is shown to determine the correlation between each characteristic and diagnosis. The seaborn heatmap function in Python is used to create heatmaps.

The bright and dark spots on the heatmap represent the main points of interest. It demonstrates the power of the correlation. The light region indicates strong association between the features or characteristics, whereas the dark area indicates poor correlation. The real correlation values given as an annotation to the heatmap make it simple to draw conclusions.

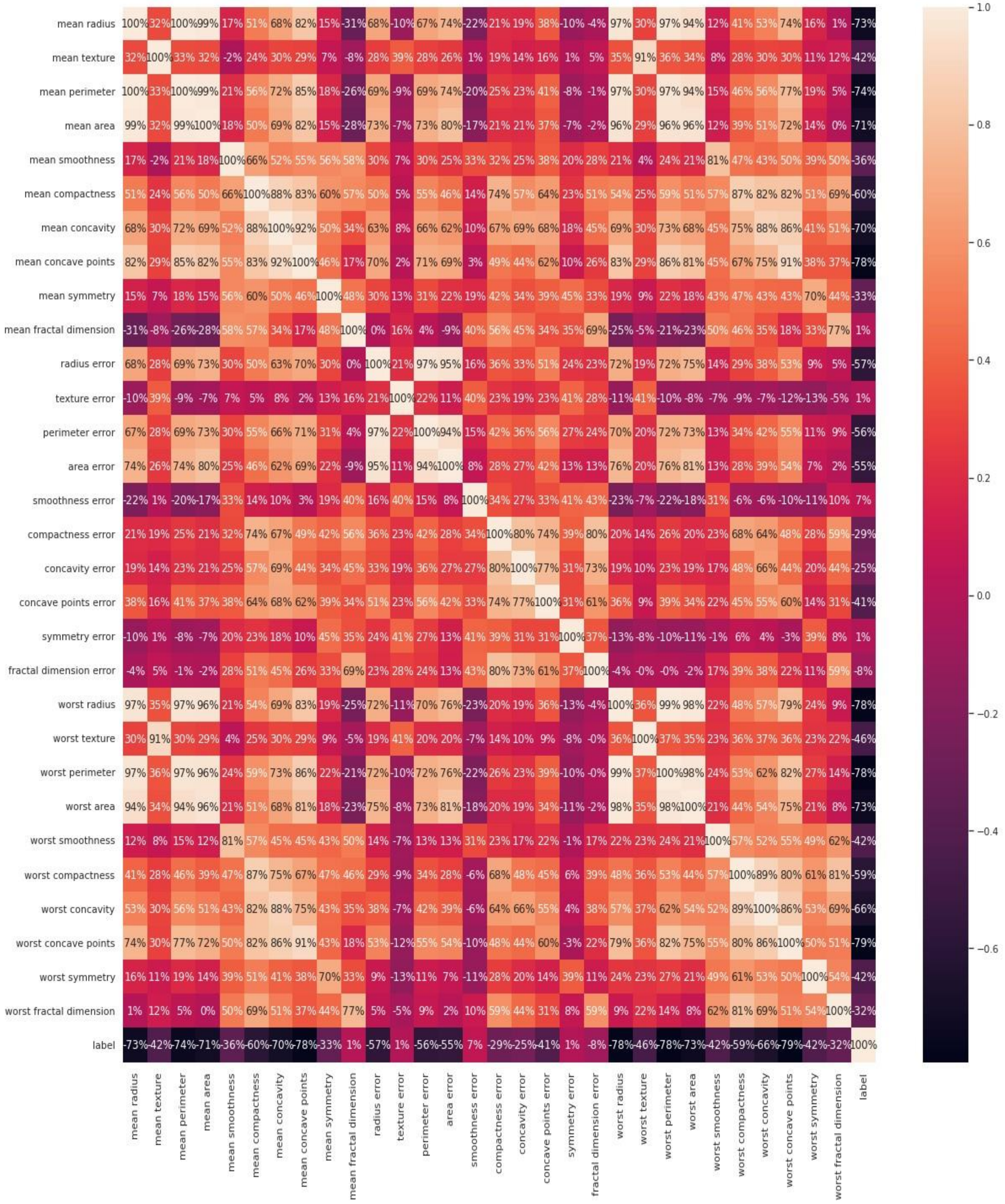


Figure 8: Heat map of corelation

e) **BARPLOT:** In a dataset, pairwise relationships are plotted using a pair plot. Each variable in the data will be shared in a single row and column on the y-axis and a single column on the xaxis thanks to the pair plot function's creation of a grid of axes. Plots are produced as in the example below. Only the feature "smoothness error" in the correlation barplot above has a stronger positive association with the target than the other features. Mean factor dimension, texture error, and symmetry error have very little positive correlation, whereas the remaining features have a lot of negative correlation.

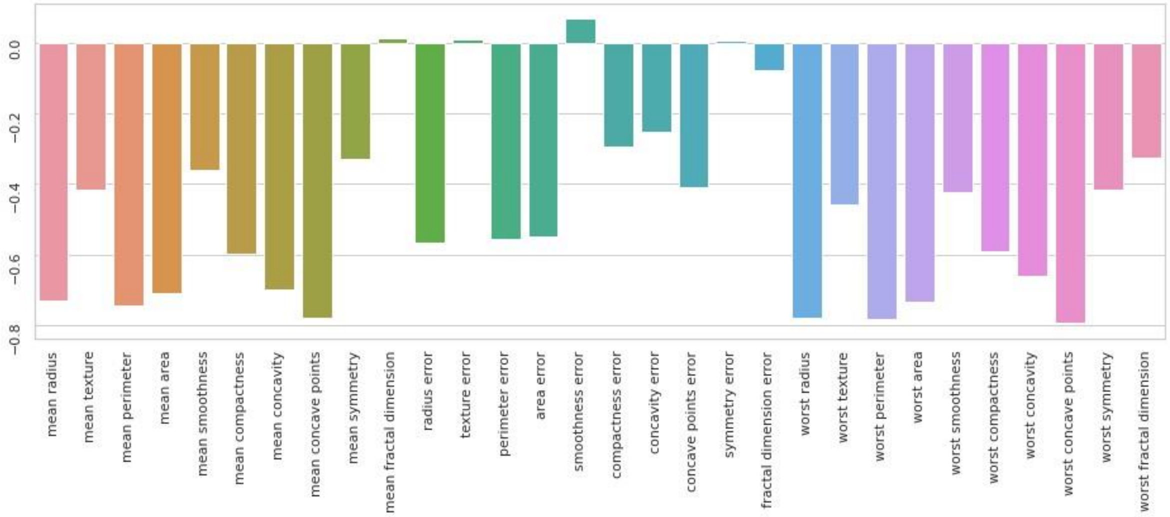


Figure 9: Bar plot

f) SWARMPLOT:A swarm plot is a type of scatter plot that is used for representing categorical values and avoids the overlapping of points. They are used to determine the most distinct clusters or the best combination of characteristics to describe a connection between two variables. Creating some straightforward linear separations or basic lines in our data set also helps to create some straightforward classification models.

i) The below fig.10 shows the data points of all the mean features of the data frame in swarmplot.

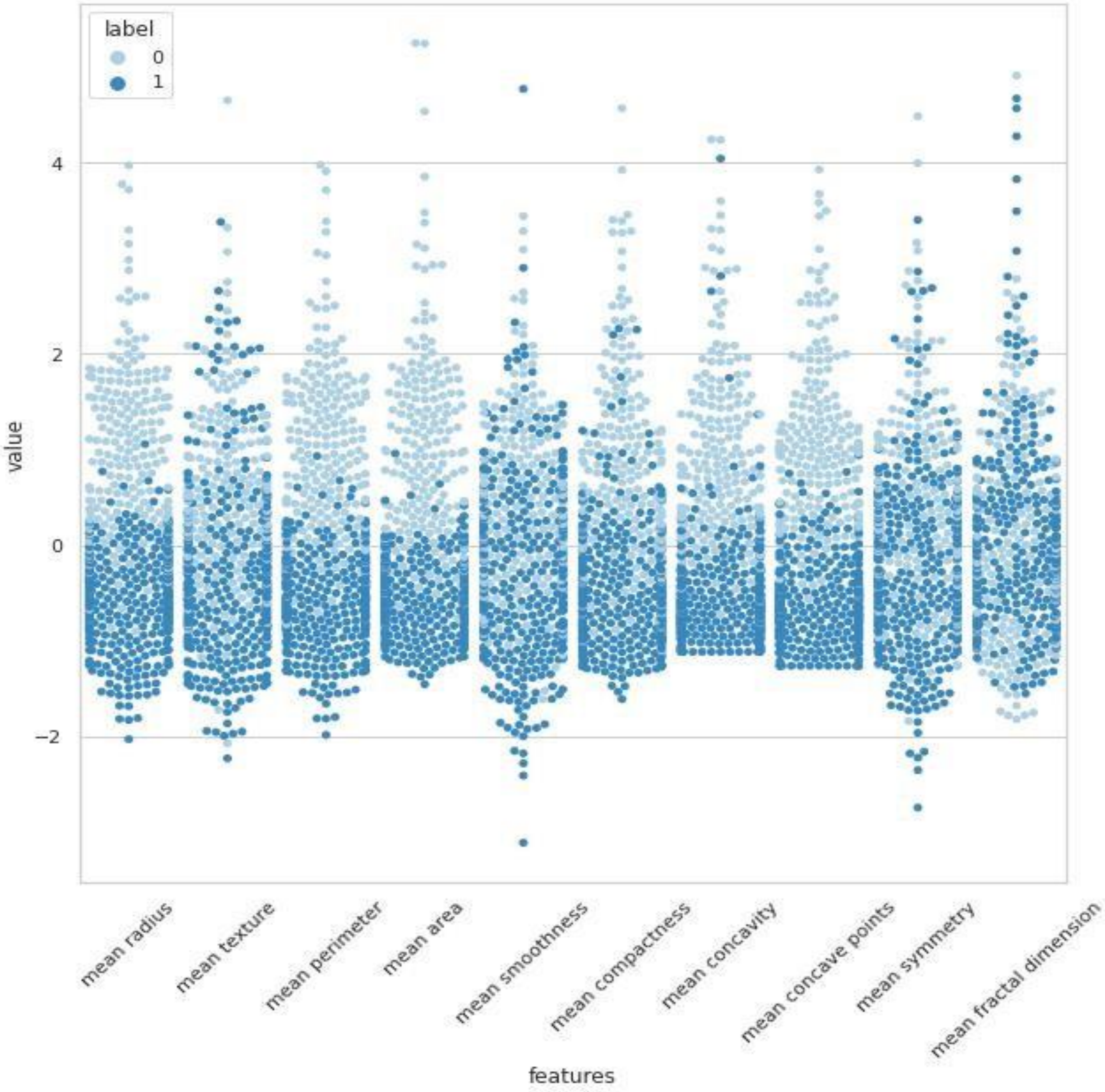


Figure 10: Swarm plot of mean features

ii) The below swarmplot shows the data points of all the error features of the data frame.

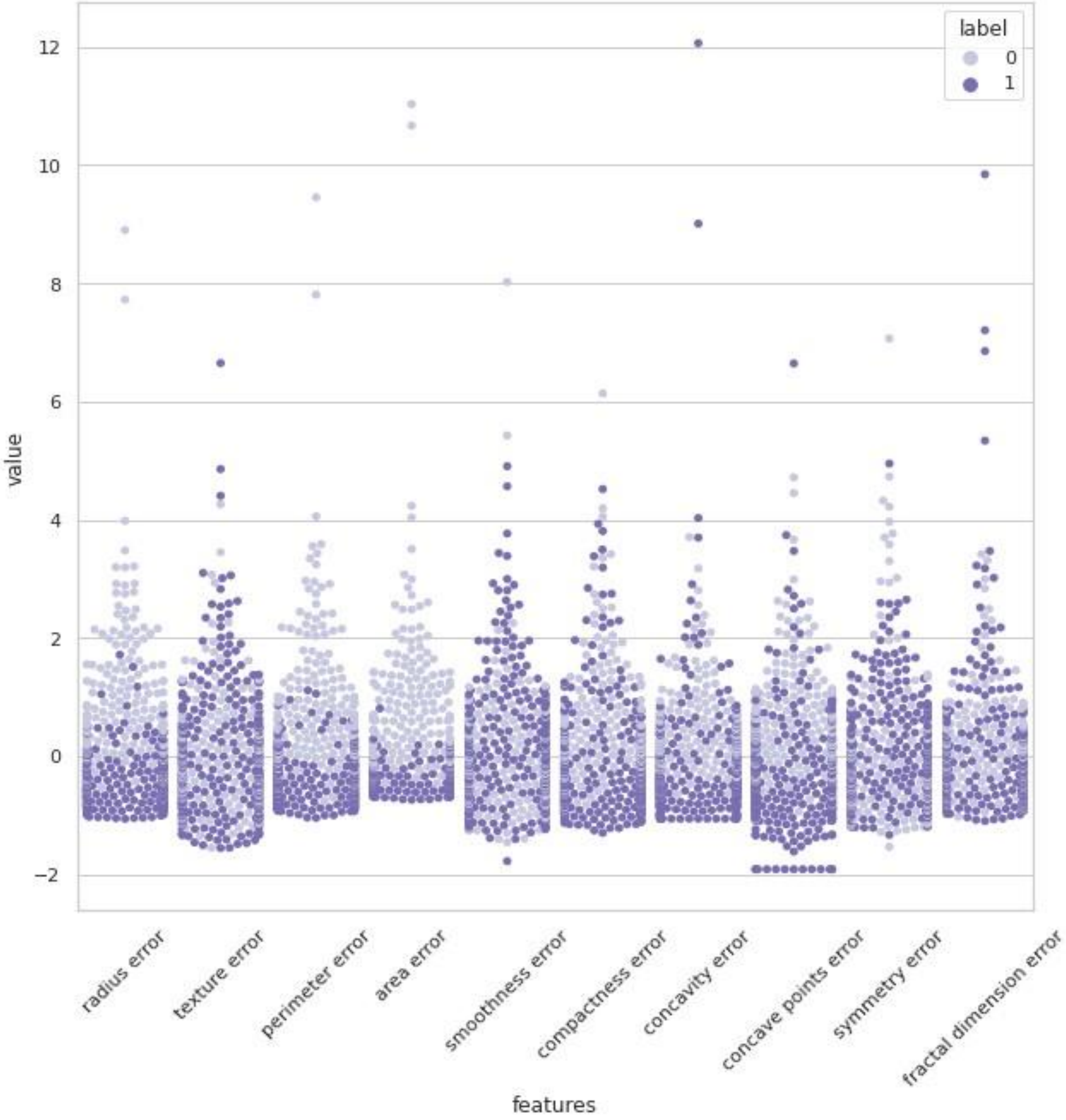


Figure 11: Swarm plot of standard errors

iii) The below swarmplot shows the data points of all the worst case features of the data frame.

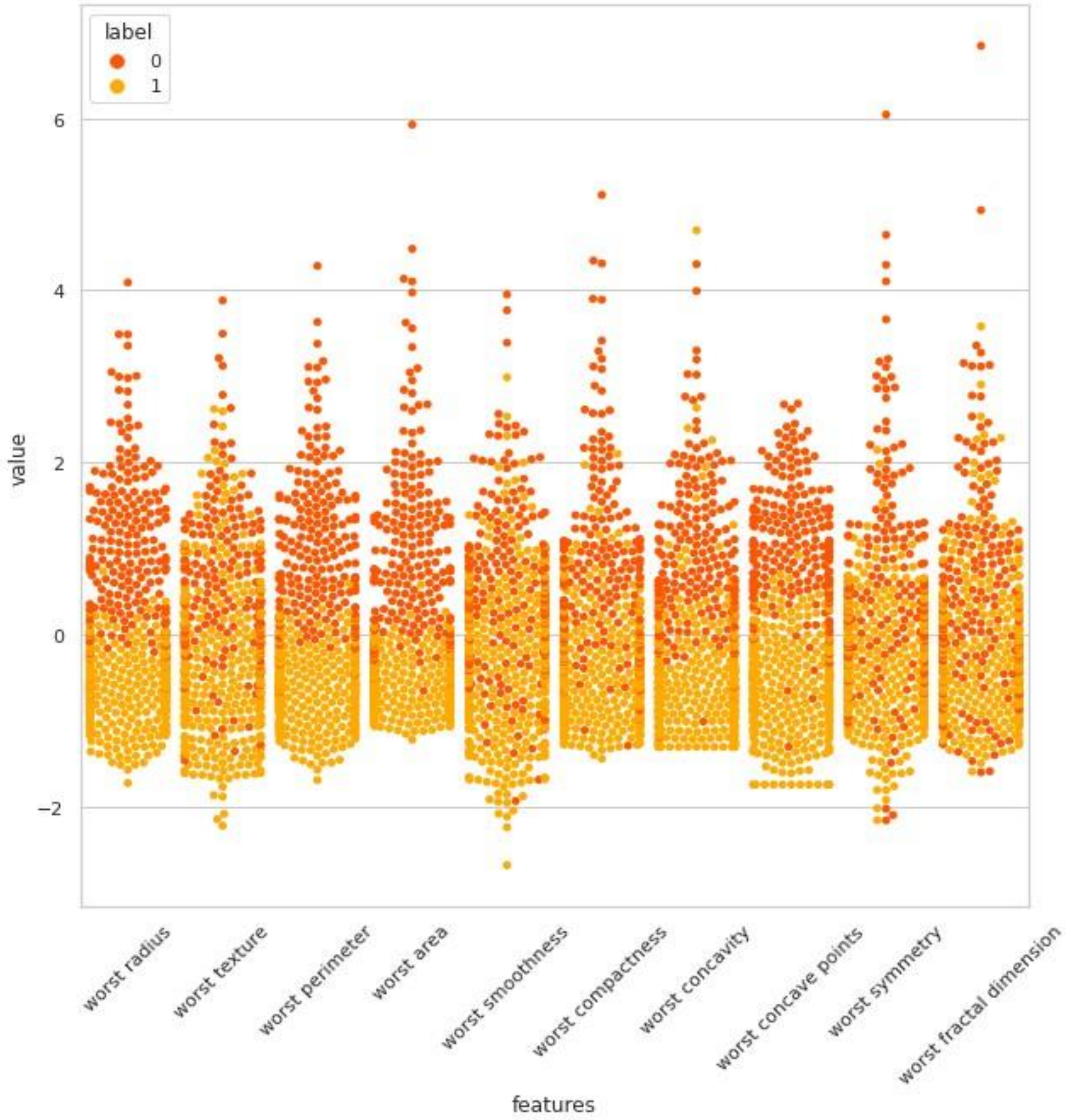


Figure 12: Swarm plot of worst features

4.6. MODEL SELECTION

Model selection is the process of choosing different machine learning algorithms. More than one kind of machine learning techniques can be used. It is also known as Algorithm selection for Predicting the best results.

The algorithms are classified into three groups:

1. **SUPERVISED LEARNING:** Method in which machine is trained on data in which the input and output are labeled. The model can learn the training data and process the future data to predict outcome. It is again divided into two groups viz.

   Regression: It is used when the result is continuous or real.

   Classification: It is used when the result is a category.

2. **UNSUPERVISED LEARNING:** Method in which the machine is trained from the unlabelled or unclassified data making the algorithm work without providing any directions.

3. **REINFORCEMENT LEARNING:** With a set of actions, parameters, and end values given to a machine learning algorithm, reinforcement learning focuses on structured learning procedures. The machine learning algorithm then attempts to explore several alternatives and possibilities after creating the rules, monitoring and analysing each output to decide which is the best. Trial and error is taught to the machine through reinforcement learning. It draws lessons from previous mistakes and starts to modify its strategy in reaction to the circumstance in order to get the optimal outcome.

In our dataset we have the outcome variable or Dependent variable i.e Y having only two set of values, either M (Malign) or B(Benign). So we will use Classification algorithm of supervised learning.

The different types of classification algorithms used in this project are:

1. Logistic Regression

2. K Nearest Neighbor
3. Support Vector Machine (Linear Classifier)
4. Support Vector Machine (RBF Classifier)
5. Gaussian Naive Bayes
6. Decision Tree Classifier
7. Random Forest Classifier

4.6.1. SPLITTING THE DATASET

We often separate the data we utilize into training data and test data. The model learns on this data in order to subsequently generalize to additional data because the training set comprises a known outcome. To evaluate the predictions of our model on this subset, we have the test dataset (or subset). We used the SciKit-Learn library in Python using the train_test_split method.

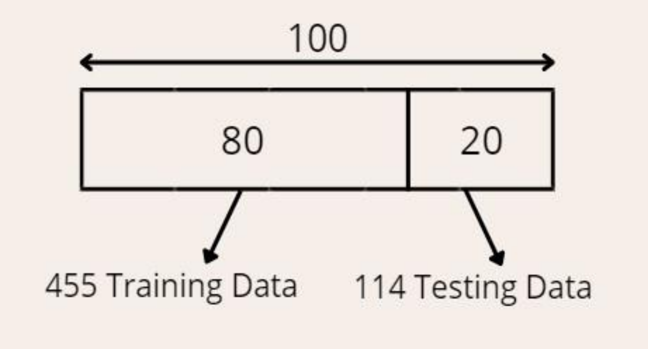


Figure 13: Train/Test splt

Train/Test is a method to measure the accuracy of your model. It is called Train/Test because you split the data set into two sets: a training set and a testing set. 80% for training, and 20% for testing. The model generated predicts the unknown results i.e. test set. Thus the dataset is divided into train and test set in order to check accuracies and precisions by training and testing it on it.

4.6.2 FEATURE SCALING

The size, units, and range of characteristics in a dataset can occasionally alter. However, because the majority of machine learning algorithms employ the Euclidean distance between two data points, it is crucial to scale all characteristics to the same magnitude. The feature/independent data will then be produced within a certain range. For illustration, 0-100 or 0-1.

This is accomplished using the StandardScaler method from sklearn.preprocessing in Python.

4.6.3. MODEL CREATION

Initially, function is created to hold all the models used in dataset to make classification. In python the algorithm is applied by sklearn library to import all the methods of classification algorithms. After importing library, and then Logistic Regression Algorithm is used to the Training Set, K Neighbors classifier Method of neighbors class to use Nearest Neighbor algorithm, SVC linear and RBF method of svm class to use Support Vector Machine Algorithm, GaussianNB method of naïve_bayes class to use Naïve Bayes Algorithm, Decision Tree classifier of tree class to use Decision Tree Algorithm, and Random Forest Classifier to use Random Forest Classification.
Within this function, accuracy of each model on the training data is also printed.

Further a model is created that contains all the models and the accuracy score on the training data for each model is observed.

To test the model, accuracy of testing data is used. To check the accuracy, the confusion_matrix method of matric class has been imported. It summarizes the performance of a classification algorithm. Confusion matrix calculates and provides what classification model is getting right and what types of errors it is making. The numbers of correct and incorrect predictions are summarized with count values and are broken down by each class. It provides insights to error being made by classifier and the types of error that are being made.

Confusion matrix accuracy of each matrix is displayed. Also, here in case of breast cancer, type II error is more dangerous error and would cause greater consequence. Thus, considering both type I and II errors in each model with respect to its accuracies, model 0 i.e. Logistic Regression is comparatively more accurate than other mentioned model for breast cancer detection.

4.6.4. CONFUSION MATRIX

Confusion matrix is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.

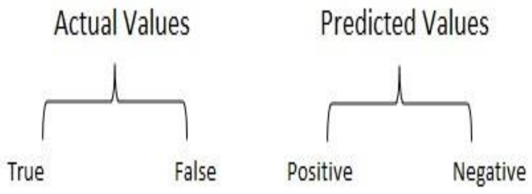


Figure 14: Predicted and actual values of confusion matrix

It reveals the number of patients with cancer who were incorrectly diagnosed as not having cancer (referred to as false negatives) and the number of patients who did not have cancer but were incorrectly diagnosed as having cancer (referred to as false positives), as well as the number of correct diagnoses, or true positives and true negatives.

False Positive (FP)(Type 1 error) = A test result which incorrectly indicates that a particular condition or attribute is present.

True Positive (TP) = Sensitivity (also called the true positive rate, or probability of detection in some fields) measures the proportion of actual positives that are correctly identified as such.

True Negative (TN) = Specificity (also called the true negative rate) measures the proportion of actual negatives that are correctly identified as such.

False Negative (FN) (Type 2 error)= A test result that indicates that a condition does not hold, while in fact it does. For example a test result that indicates a person does not have cancer when the person actually does have it.

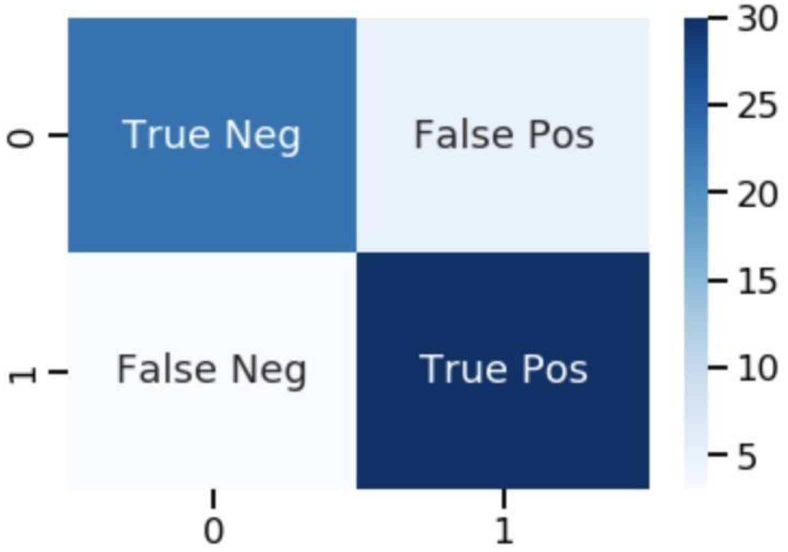


Figure 15: Confusion matrix

Recall, Precision, Specificity, Accuracy, and AUC-ROC curves are measured using confusion matrix.

i) Recall: The recall equation can be explained by saying, from all the positive classes, how many we predicted correctly.

$$\textbf{\textit{Recall}} = \frac{TP}{TP + FN}$$

ii) Precision: Precision equation can be explained by saying, from all the classes we have predicted as positive, how many are actually positive. Precision should be high as possible.

$$\textbf{\textit{Precision}} = \frac{TP}{TP + FP}$$

iii) Accuracy: From all the classes (positive and negative), how many of them we have predicted correctly.

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

iv) F-measure: It is difficult to compare two models with low precision and high recall or vice versa. So to make them comparable, we use F-Score. F-score helps to measure Recall and Precision at the same time. It uses Harmonic Mean in place of Arithmetic Mean by punishing the extreme values more.

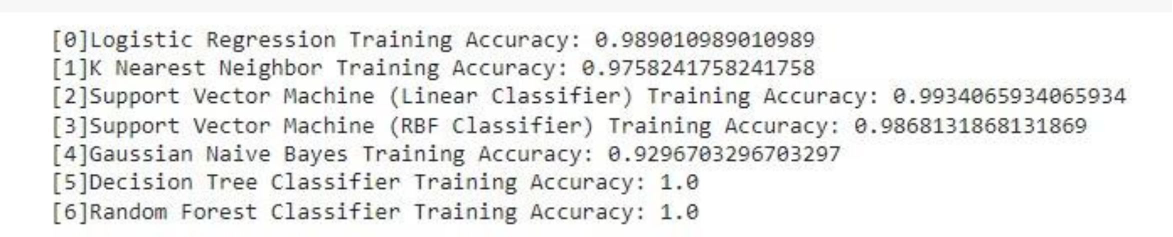
$$\textit{F - measure} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}}$$

CHAPTER 5

RESULT AND DISCUSSION

5.1 RESULT ANALYSIS

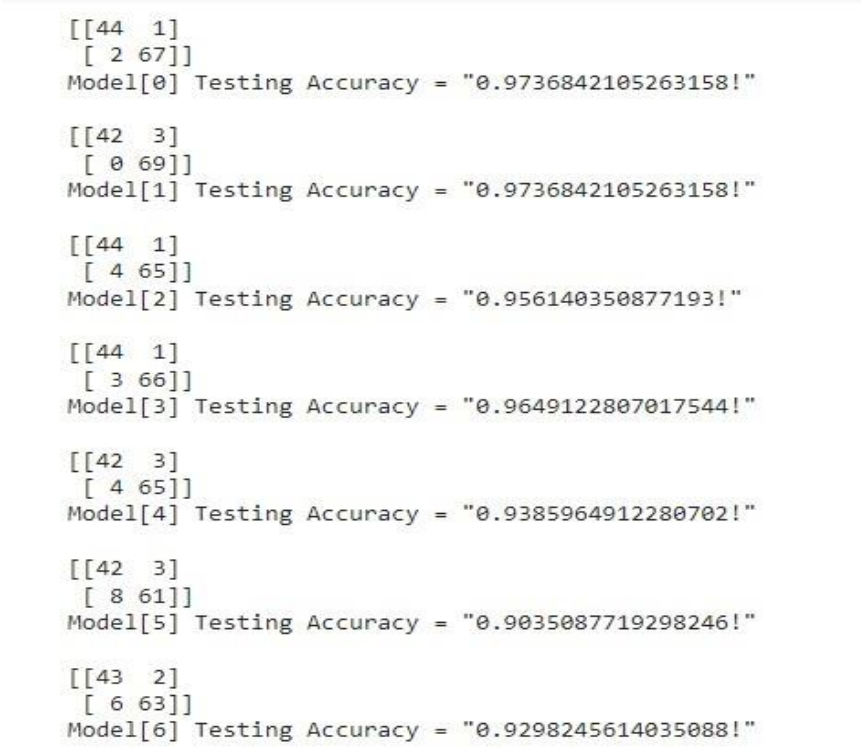
The final step in our project is building the predictive system. Training and testing accuracy of each model is considered, which is followed by the selection of the model with highest accuracy.



```
[0]Logistic Regression Training Accuracy: 0.989010989010989
[1]K Nearest Neighbor Training Accuracy: 0.9758241758241758
[2]Support Vector Machine (Linear Classifier) Training Accuracy: 0.9934065934065934
[3]Support Vector Machine (RBF Classifier) Training Accuracy: 0.9868131868131869
[4]Gaussian Naive Bayes Training Accuracy: 0.9296703296703297
[5]Decision Tree Classifier Training Accuracy: 1.0
[6]Random Forest Classifier Training Accuracy: 1.0
```

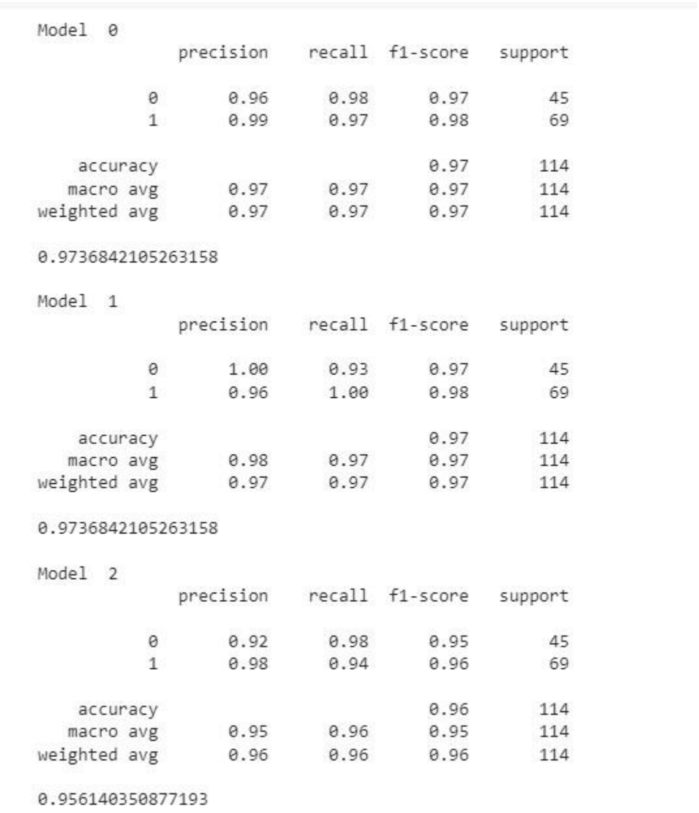
Figure 16: Training accuracy of the models

Figure depicts the training accuracy of each model, in which SVM(Linear) has the highest accuracy. Accuracy of test data and their respective confusion matrix is also considered. It summarizes the performance of a classification algorithm. Figure shown below is the confusion matrix of test data.

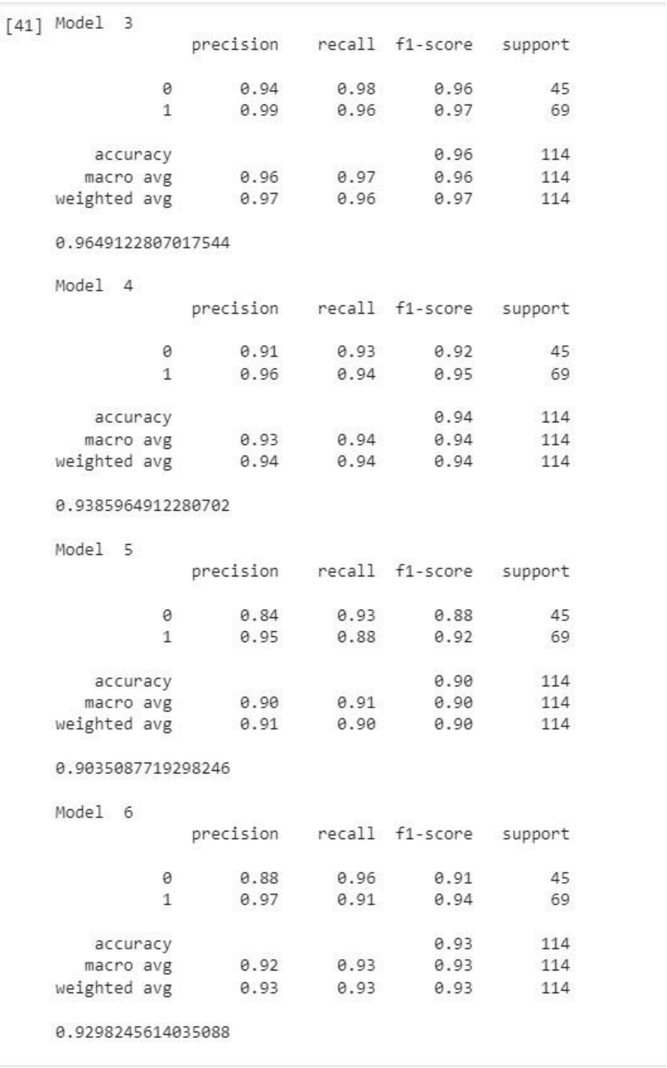


```
[[44  1]
 [ 2 67]]
Model[0] Testing Accuracy = "0.9736842105263158!"

[[42  3]
 [ 0 69]]
Model[1] Testing Accuracy = "0.9736842105263158!"

[[44  1]
 [ 4 65]]
Model[2] Testing Accuracy = "0.956140350877193!"

[[44  1]
 [ 3 66]]
Model[3] Testing Accuracy = "0.9649122807017544!"

[[42  3]
 [ 4 65]]
Model[4] Testing Accuracy = "0.9385964912280702!"

[[42  3]
 [ 8 61]]
Model[5] Testing Accuracy = "0.9035087719298246!"

[[43  2]
 [ 6 63]]
Model[6] Testing Accuracy = "0.9298245614035088!"
```

Figure 17: Testing accuracy of models



```
Model  0
              precision    recall  f1-score   support

           0       0.96      0.98      0.97        45
           1       0.99      0.97      0.98        69

    accuracy                           0.97       114
   macro avg       0.97      0.97      0.97       114
weighted avg       0.97      0.97      0.97       114

0.9736842105263158

Model  1
              precision    recall  f1-score   support

           0       1.00      0.93      0.97        45
           1       0.96      1.00      0.98        69

    accuracy                           0.97       114
   macro avg       0.98      0.97      0.97       114
weighted avg       0.97      0.97      0.97       114

0.9736842105263158

Model  2
              precision    recall  f1-score   support

           0       0.92      0.98      0.95        45
           1       0.98      0.94      0.96        69

    accuracy                           0.96       114
   macro avg       0.95      0.96      0.95       114
weighted avg       0.96      0.96      0.96       114

0.956140350877193
```

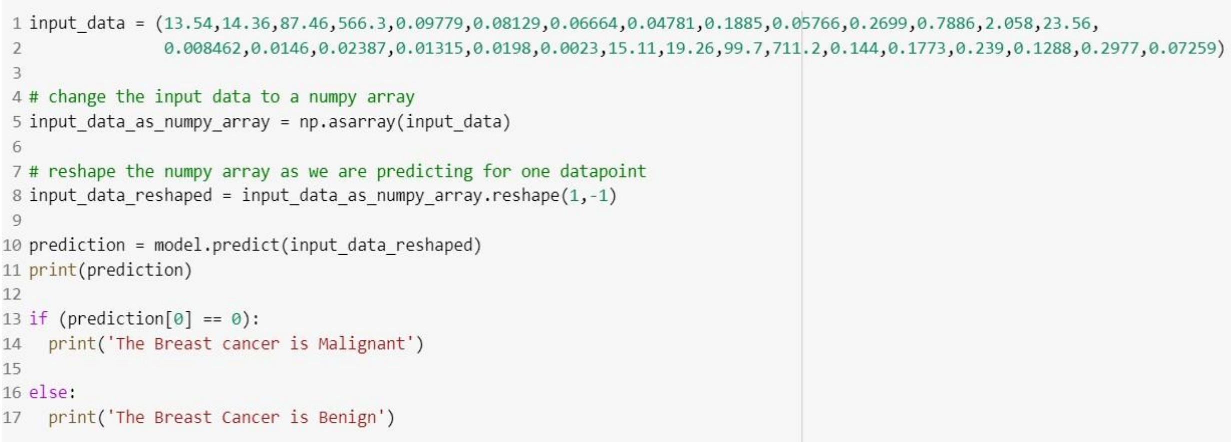


```
[41] Model  3
              precision    recall  f1-score   support

           0       0.94      0.98      0.96        45
           1       0.99      0.96      0.97        69

    accuracy                           0.96       114
   macro avg       0.96      0.97      0.96       114
weighted avg       0.97      0.96      0.97       114

0.9649122807017544

Model  4
              precision    recall  f1-score   support

           0       0.91      0.93      0.92        45
           1       0.96      0.94      0.95        69

    accuracy                           0.94       114
   macro avg       0.93      0.94      0.94       114
weighted avg       0.94      0.94      0.94       114

0.9385964912280702

Model  5
              precision    recall  f1-score   support

           0       0.84      0.93      0.88        45
           1       0.95      0.88      0.92        69

    accuracy                           0.90       114
   macro avg       0.90      0.91      0.90       114
weighted avg       0.91      0.90      0.90       114

0.9035087719298246

Model  6
              precision    recall  f1-score   support

           0       0.88      0.96      0.91        45
           1       0.97      0.91      0.94        69

    accuracy                           0.93       114
   macro avg       0.92      0.93      0.93       114
weighted avg       0.93      0.93      0.93       114

0.9298245614035088
```

Figure 18: Confusion matrix of models

Taking the confusion matrix into consideration and analysing the accuracies and other parameters, it is observed that the ‘Logistic Regression’ model with training accuracy 98% and testing accuracy 97.36% is the best fit for the predictive system.

Predictive system detects whether the patient is suffering from benign or malignant tumours. Initially input data is taken from the Breast Cancer Dataset, which is then converted into a NumPy array and fed into prediction model. If prediction is equal to zero, the system predicts the tumour as ‘Malignant’ else ‘Benign’. The figure below shows the system which predicts the type of tumour for the given input.



```python
input_data = (13.54,14.36,87.46,566.3,0.09779,0.08129,0.06664,0.04781,0.1885,0.05766,0.2699,0.7886,2.058,23.56,
              0.008462,0.0146,0.02387,0.01315,0.0198,0.0023,15.11,19.26,99.7,711.2,0.144,0.1773,0.239,0.1288,0.2977,0.07259)

# change the input data to a numpy array
input_data_as_numpy_array = np.asarray(input_data)

# reshape the numpy array as we are predicting for one datapoint
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0] == 0):
  print('The Breast cancer is Malignant')

else:
  print('The Breast Cancer is Benign')
```

Figure 19: Predictive system

CHAPTER 6
CONCLUSION

6. CONCLUSION

This project aims to evaluate different supervised machine learning algorithms and choose the most reliable model for breast cancer classification. The effort concentrated on improving predictive models using Python to forecast outcomes with greater precision. The study of the results shows that feature scaling, data integration, and various classification methods and analyses deliver a remarkably effective tool for prediction. Additionally, it has been noted that the model occasionally misdiagnosed people with cancer when they didn't have it and vice versa. The model is accurate, but when it comes to dealing with people's lives, further study is needed to produce the most exact and precise model. This will help classification approaches work better and achieve accuracy as near to 100 percent as feasible. As a result, in order to create a model that is more trustworthy, it is important to tune each of the existing models.

BIBLIOGRAPHY

[1] J. Sivapriya, A. Kumar, S. Siddarth Sai, and S. Sriram, “Breast cancer prediction using machine learning,” International Journal of Recent Technology and Engineering (IJRTE), vol. 8, 2019.

[2] Y. Khourdifi and M. Bahaj, “Applying best machine learning algorithms for breast cancer prediction and classification,” in 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS). IEEE, 2018, pp. 1-5.

[3] N. K. Sinha, M. Khulal, M. Gurung, and A. Lal, “Developing a web based system for breast cancer prediction using xgboost classifier,” International Journal of Engineering Research Technology (IJERT), vol. 9, 2020.

[4] R. Dhanya, I. R. Paul, S. S. Akula, M. Sivakumar, and J. J. Nair, “A comparative study for breast cancer prediction using machine learning and feature selection,” in 2019 International Conference on Intelligent Computing and Control Systems (ICCS). IEEE, 2019, pp. 1049-1055.

[5] M. M. Islam, H. Iqbal, M. R. Haque, and M. K. Hasan, “Prediction of breast cancer using support vector machine and k-nearest neighbors,” in 2017 IEEE Region 10 Humanitarian Technology Conference (R10- HTC). IEEE, 2017, pp. 226-229.

[6] M. S. Yarabarla, L. K. Ravi, and A. Sivasangari, “Breast cancer prediction via machine learning,” in 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, 2019, pp. 121-124.

[7] V. Chaurasia, S. Pal, and B. Tiwari, “Prediction of benign and malignant breast cancer using data mining techniques,” Journal of Algorithms & Computational Technology, vol. 12, no. 2, pp. 119-126, 2018.

[8] N. Fatima, L. Liu, S. Hong, and H. Ahmed, “Prediction of breast cancer, comparative review of machine learning techniques, and their analysis,” IEEE Access, vol. 8, pp. 150360-150376, 2020.

[9] A. Toprak, “Extreme learning machine (elm)-based classification of benign and malignant cells in breast cancer,” Medical science monitor: international medical journal of experimental and clinical research, vol. 24, p. 6537, 2018.

[10] D. S. Jacob, R. Viswan, V. Manju, L. PadmaSuresh, and S. Raj, “A survey on breast cancer prediction using data miningtechniques,” in 2018 Conference on Emerging Devices and Smart Systems (ICEDSS). IEEE, 2018, pp. 256-258.

[11] K. L. Kashyap, M. K. Bajpai, and P. Khanna, “Breast cancer detection in digital mammograms,” in 2015 IEEE international conference on imaging systems and techniques (IST). IEEE, 2015, pp. 1-6.

[12]https://www.ijert.org/breast-cancer-classification-and-prediction-using-machinelearning

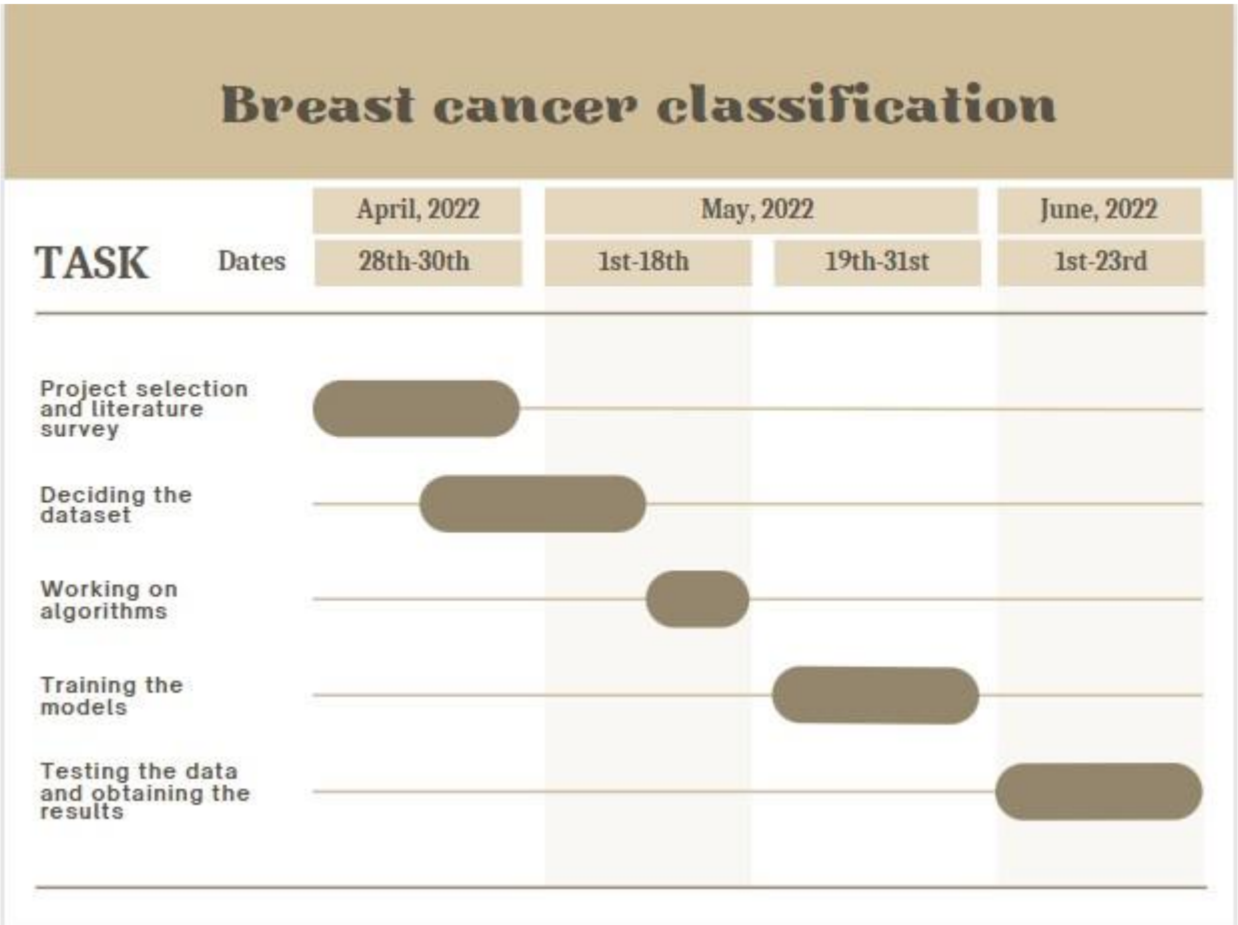
[13]https://www.irjet.net/archives/V8/i2/IRJET-V8I2129.pdf

[14]https://www.ijraset.com/research-paper/breast-cancer-detection-with-machine-learning

[15]https://www.researchgate.net/publication/333092636_Breast_Cancer_Diagnosis_and_Predict io n_Using_Machine_Learning_and_Data_Mining_Techniques_A_Review

[16]https://www.researchgate.net/publication/341508593_BREAST_CANCER_PREDICTION_ U SING_MACHINE_LEARNING

GANNT CHART


Breast cancer classification
April, 2022
May, 2022
June, 2022
TASK
Dates
28th-30th
1st-18th
19th-31st
1st-23rd
Project selection and literature survey
Deciding the dataset
Working on algorithms
Training the models
Testing the data and obtaining the results