A Multivariate Breast Cancer Model

Future Institute of Engineering and Management

Maulana Abul Kalam Azad University

Subhajit Majumder . Akash Halder

Head of the Department : Dr. Anirban Chakrabarty

team FUTURE

FUTURE INSTITUTE OF ENGINEERING AND MANAGEMENT
Sonarpur Station Road, Kolkata – 700150

MCA-DEPT

CERTIFICATE FOR PROJECT WORK

We do hereby certifing that the work which is being presented in the Minor Project Report entitled "A Multivariate Breast Cancer Model", in partial fulfillment of the requirements for the award of Master of Computer Applications submitted to the MCA-Dept of Future Institute of Engineering and Management, Kolkata, WB is an authentic record of our own work carried out during the period from 7/9/21 to 20/9/21 under the supervision of **Head of the Department, Dr. Anirban Chakrabarty.**

The matter presented in this thesis has not been submitted by us for the award of any other degree elsewhere.

Name & Signature of the Candidate(s)

a) Akash Halder

b) Subhajit Majumder.

This is to certify that the above statement made by the students, is correct to the best of my knowledge.

A Chakrabarty

Signature of the Supervisor

Date: 20/09/21

A Chakrabarty

Head
MCA-Dept
Future Institute of Engineering and Management
Kolkata, WB

Signature of the External Examiner & Panel Members

ACKNOWLEDGEMENT

We have taken efforts in this project , it would not have been possible without the kind support and guidance of our supervisor **Dr. Anirban Chakrabarty** sir. We would like to extend our sincere thanks to him.

We would also like to express our gratitude for constant supervision as well as for providing necessary
information regarding the project and also for his support in completing the project. His constant guidance and willingness to share his vast knowledge made us understand this project and its manifestations in great depths and helped us to complete the assigned tasks on time.

We would like to express our gratitude towards our fellow group members for their kind cooperation and
encouragement which helped us in completion of this project.

Our thanks and appreciation also goes to our teachers and fellow classmates in the department of Computer Application.

TABLE OF CONTENTS

Abstract

In the developing world, cancer death is one of the major problems for humankind. Even though there are many ways to prevent it before happening, some cancer types still do not have any treatment. One of the most common cancer types is breast cancer, and early diagnosis is the most important thing in its treatment. Accurate diagnosis is one of the most important processes in breast cancer treatment. In the literature, there are many studies about predicting the type of breast tumors. In this research paper, data about breast cancer tumors from Dr. William H. Walberg of the University of Wisconsin Hospital were used for making predictions on breast tumor types. Data visualization and machine learning techniques including logistic regression, k-nearest neighbors, support vector machine, naïve Bayes, decision tree, random forest, and rotation forest were applied to this dataset. R, Minitab, and Python were chosen to be applied to these machine learning techniques and visualization. The paper aimed to make a comparative analysis using data visualization and machine learning applications for breast cancer detection and diagnosis. Diagnostic performances of applications were comparable for detecting breast cancers. Data visualization and machine learning techniques can provide significant benefits and impact cancer detection in the decision-making process. In this paper, different machine learning and data mining techniques for the detection of breast cancer were proposed. Results obtained with the logistic regression model with all features included showed the highest classification accuracy (96.06%), and the proposed approach revealed the enhancement in accuracy performances. These results indicated the potential to open new opportunities in the detection of breast cancer.

Keywords: breast cancer, data visualization, early diagnosis, machine learning, risk assessment

Introduction

Data science has become one of the most popular research areas of interest in the world. Many datasets can be useful in different situations such as marketing, transportation, social media, and healthcare. However, only a few of them have been interpreted by data science researchers, and they believe that these datasets can be useful for predictions. Nowadays, many of the marketers have started to analyze their datasets because of the big information they have on hand, and they want to turn these data into meaningful information for future predictions. By doing that, marketers can apply some new tactics or change their goal.

Data mining and machine learning techniques are straightforward and effective ways to understand and predict future data. Dealing with large data manually is almost impossible. Therefore, data visualization is a very important step to have a general idea about given data. Data analysis techniques are popular in many companies and have an impact on different study areas. For instance, Facebook's News Feed uses machine learning by following user patterns. Another study has been made about optimizing energy consumption in large-scale buildings. Customer relationship management systems are also using machine learning techniques. In addition to all these different studies, machine learning studies in healthcare are very popular. Data mining techniques and clustering methods are used for different types of diseases to make data understandable and teach the computer to predict current data.

Cancer death is one of the major issues for the healthcare environment. It is one of the most significant reasons for women's death. Breast cancer is the most common type of cancer in women with denser breast tissue due to its physiological features. The detection of this disease in the early stages can help to avoid the rising number of deaths. According to the Globocan 2018 data, one of every four cancer cases diagnosed in women worldwide is breast cancer, and it ranks fifth among the causes of death worldwide. According to the same data, the incidence of age-related breast cancer worldwide in 2018 was 23.7 per 100,000, whereas the mortality rate due to breast cancer was reported as 6.8 per 100,000. Despite the increase in the number of medical studies and technological developments that contribute to the treatment of cancer, there are still some problems in the diagnosis of cancer. After lung cancer, breast cancer is the major cause of women's death. Breast cancer originates from breast tissue, most commonly from the inner lining of milk ducts or the lobules that supply the ducts with milk. A mutation or modification of DNA or RNA could force normal cells to transform into cancer cells, and these mutations could occur due to an increase in entropy or nuclear radiation, chemicals in the air, bacteria, fungi, electromagnetic radiation viruses, parasites, heat, water, food, mechanical cell-level injury, free radicals, evolution, and aging of DNA and RNA. It is important to make an accurate diagnosis of tumors. Most tumors are the result of benign (non-cancerous) changes within the breast, but if a malignant tumor is diagnosed as benign it will cause serious problems. Early detection of breast cancer and getting modern cancer treatment are the most important strategies to prevent deaths from breast cancer. It is easy to treat early, small, and non-spreading breast cancer successfully. The most reliable way to find breast cancer early is by having regular screening tests.

Age, family history, genetics, race, ethnicity, being overweight, drinking alcohol, and lack of exercise are risk factors associated with breast cancer. Healthcare is an open-ended environment with very rich information, yet very poor knowledge. There is a huge amount of data in healthcare systems, and it is important to discover and build relationships with hidden data. The main causes of death were classified into five broad groups according to the International Classification of Diseases (ICD), and breast cancer was included in two groups. A report from McKinsey states that the volume of data is growing at a rate of 50% per year. Currently, data science has officially become a very significant field even though the term data science was first coined in the early 1990s. A study defined the term data science as implying focus around data and, by extension, statistics, which is a systematic study about the organization, properties, and analysis of data and their role in inference. In previous data mining research about healthcare, some methods were applied to different types of diseases and genes, and the methods including analytical, collecting, sharing, and compressing methods were applied on healthcare datasets. Even though multiple disciplines can be applied to data science, machine learning methods are mostly applied to healthcare datasets. Machine learning is a data analysis technique that teaches a computer what comes as an output

with different algorithms. Decision tree, k-means clustering, and neural networks are the most common algorithms for machine learning applications. While there is no better way to diagnose breast cancer, early diagnosis can be accepted as the first step of treatment and risk assessment to minimize factors. It allows a person to control risk factors, although some breast cancer risk factors cannot be changed.

In this study, public data about breast cancer tumors from Dr. William H. Walberg of the University of Wisconsin Hospital were taken and used for data visualization, classification, and machine learning algorithms, which included logistic regression, k-nearest neighbors, support vector machine, and decision tree. Public data included samples taken from patients with solid breast masses and a user-friendly usage of graphical programs called City. This study aimed to establish an adequate model by revealing the predictive factors of early-stage breast cancer patients from a wider perspective and compare the strength of the model with accuracy measures.

The organization of the remaining sections of the current study is as follows. A literature review that contains recent related studies on breast cancer detection and diagnosis is in Section 2. In Section 3 and Section 4, methods and application of the proposed method to a dataset are given. The results, discussion, and comparative analysis are demonstrated in the last section.

Methods and application

1. Logistic Regression

Logistic regression is a technique that firstly used for biological studies in the early twentieth century. It has become widespread for social studies too. Logistic regression is also one of the predictive analyses. Logistic regression is appropriate to use when there is one binary dependent variable and other independent variables. Linear and logistic regressions are different in terms of the dependent variable. Linear regression is a more appropriate technique for continuous variables.

Logistic regression has two phases: forward propagation and backward propagation. The first step of forward propagation is multiplying weights with features. Initially, since weights are unknown, random values can be assigned. A sigmoid function assigns a probability between 0 and 1. According to a threshold value, the prediction is performed. After prediction, the predictive value is compared with the observed values, and then a loss function is generated. The loss function indicates how far the predicted value is from the real value. If the loss function value is very high, then backward propagation is applied. The aim of backward propagation is updating weight values according to cost function by taking the derivative . The sigmoid function is shown below:

$$\sigma(z) = \frac{1}{1+e^{-z}} \qquad \ldots\ldots\ldots(1)$$

2. K-Nearest Neighbor (KNN)

KNN is a supervised learning technique that means the label of the data is identified before making predictions. Clustering and regression are two purposes to use it. K represents a numerical value for the nearest neighbors. KNN algorithm does not have a training phase. Predictions are made based on the Euclidean distance to k-nearest neighbors. This technique is applied to the prediction of breast cancer dataset since it already has labels such as malignant and benign. The label is classified according to the nearest neighbor to the class labels of its neighbors. A representation of the KNN algorithm is shown below.

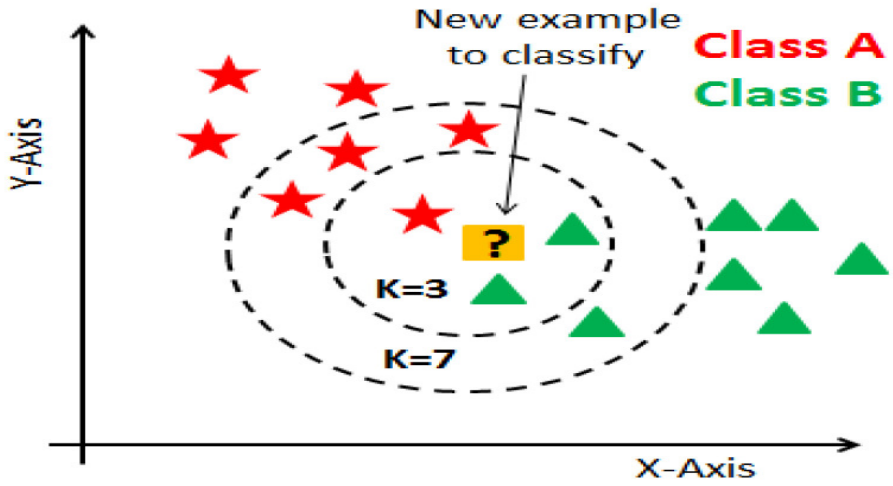

Figure - 1

3. Decision Tree

A decision tree (DT) is one of the most common supervised learning techniques. Regression and classification are two main goals to use it. It seeks to solve problems by drawing a tree figure. Features are known as decision nodes, and outputs are leaf nodes. Feature values are considered as categorical in the decision tree algorithm. At the very beginning of this algorithm, it is essential to choose the best attribute and place it at the top on tree figure and then split the tree. Gini index and information gain are two methods for the selection of features.

Randomness or uncertainty of feature x is defined as entropy and can be calculated as follows:

$$H(x)=Ex[I(x)]=-\sum p(x)\log p(x) \qquad \cdots\cdots\cdots\cdots (2)$$

Entropy values for each variable are calculated, and by subtracting these values from one, information values can be obtained. A higher information gain makes an attribute better and places it on top of the tree.

Gini index is a measure of how often a randomly chosen element would be incorrectly identified. Therefore, a lower Gini index value means better attributes. Gini index can be found with the given formula:

$$G=\sum pi*(1-pi) \text{ for } i=1,\ldots n \qquad \cdots\cdots\cdots (3)$$

A decision tree is easy to understand. However, if data contain various features it might cause problems that are called overfitting. Therefore, it is crucial to know when to stop growing trees. Two methods are typical for restricting the model from overfitting: pre-pruning, which stops growing early, but it is hard to choose a stopping point; and post-pruning, which is a cross-validation used to check whether expanding the tree will make improvements or lead to overfitting. DT structure consists of a root node, splitting, decision node, terminal node, sub-tree, and parent node. There are two main phases of the DT induction process: the growth phase and the pruning phase. The growth phase involves a recursive partitioning of the training data resulting in a DT where decision trees have a natural “if”, “then”, “else” construction that makes it fit easily into a programmatic structure

- **here we have implemented an ensemble technique, by combining the mentioned 3 classifier outcomes.**

4. Voting Classifier

A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output.

It simply aggregates the findings of each classifier passed into Voting Classifier and predicts the output class based on the highest majority of voting. The idea is instead of creating separate dedicated models and finding the accuracy for each them, we create a single model which trains by these models and predicts output based on their combined majority of voting for each output class.

Voting Classifier supports two types of votings.

Hard Voting: In hard voting, the predicted output class is a class with the highest majority of votes i.e the class which had the highest probability of being predicted by each of the classifiers. Suppose three classifiers predicted the output class(A, A, B), so here the majority predicted A as output. Hence A will be the final prediction.

Soft Voting: In soft voting, the output class is the prediction based on the average of probability given to that class. Suppose given some input to three models, the prediction probability for class A = (0.30, 0.47, 0.53) and B = (0.20, 0.32, 0.40). So the average for class A is 0.4333 and B is 0.3067, the winner is clearly class A because it had the highest probability averaged by each classifier.

Evaluation Matrix:

1.Confusion Metrics

It is a matrix of size 2×2 for binary classification with actual values on one axis and predicted on another.

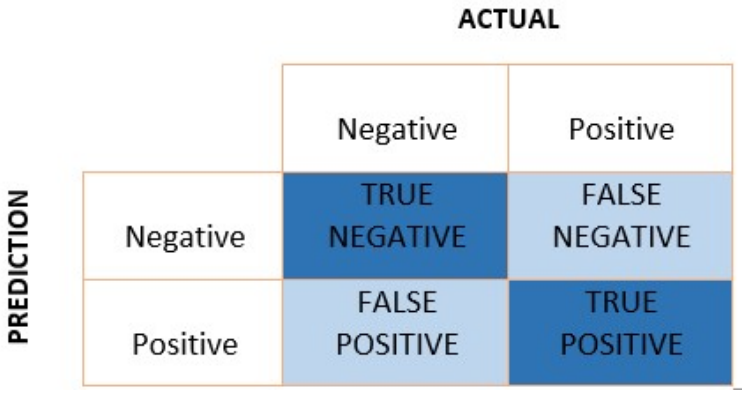

Figure - 2

Let's understand the confusing terms in the confusion matrix: **true positive, true negative, false negative, and false positive** with an example.

EXAMPLE

A machine learning model is trained to predict tumor in patients. The test dataset consists of 100 people.

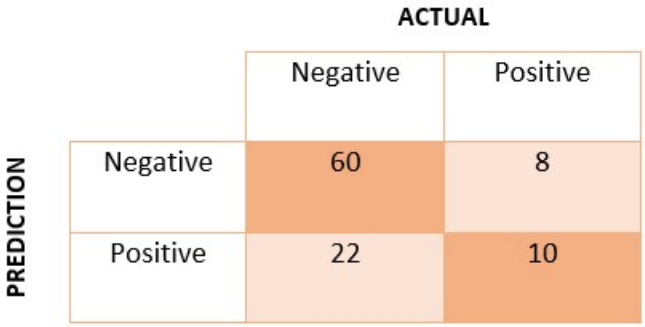

Figure - 3

2.Accuracy

Accuracy is one metric for evaluating classification models. Informally, **accuracy** is the fraction of predictions our model got right. Formally, accuracy has the following definition

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total No of Prediction}} \qquad \text{.....(4)}$$

3.Precision

Precision is a measure that tells you how meaningful a positive (target class = 1) is. This is accomplished by dividing the number of correct positive predictions over total number of positive predictions (true positive divided by sum of false positives and true positives).

$$Precision = \frac{TP}{TP + FP} \qquad \text{.....(5)}$$

4.Recall

Recall is very similar to precision. The denominator in recall, however, is composed of true positives and false negatives, while the numerator remains true positives. This shifts our focus from how much a positive score actually matters, to an understanding of how effective our model is at identifying any positive case that is present (because a false negative is actually a positive). As the number of false negatives grow, there is no increase to the numerator of true positives, and the recall gets lower and lower.

$$Recall = \frac{TP}{TP + FN} \qquad \text{.....(6)}$$

5.F1-Score

The F1-Score is the perfect way in which we can get a better sense of model performance when we have imbalanced data since accuracy alone isn't a good metric. F1-Score is a harmonic average (max value is the arithmetic mean) of the precision and recall score. Having a blend of precision and recall gives us a strong idea of how well the model actually works.

$$F1\ score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \qquad \text{.....(7)}$$

Experiment

Attribute Information :

	Id	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses	Class
0	1000025	5	1	1	1	2	1.0	3	1	1	0
1	1002945	5	4	4	5	7	10.0	3	2	1	0
2	1015425	3	1	1	1	2	2.0	3	1	1	0
3	1016277	6	8	8	1	3	4.0	3	7	1	0
4	1017023	4	1	1	3	2	1.0	3	1	1	0
5	1017122	8	10	10	8	7	10.0	9	7	1	1

Figure No- 4

In **figure no. 4** we can see the dataset upto n-th(here, n=6) position from head**.**

1. Sample code number: id number
2. Clump Thickness: 1 - 10
3. Uniformity of Cell Size: 1 - 10
4. Uniformity of Cell Shape: 1 - 10
5. Marginal Adhesion: 1 - 10
6. Single Epithelial Cell Size: 1 - 10
7. Bare Nuclei: 1 - 10
8. Bland Chromatin: 1 - 10
9. Normal Nucleoli: 1 - 10
10. Mitoses: 1 - 10
11. Class: (0 for benign, 1 for malignant)

Implementation:

```
import numpy as np
import pandas as pd
from google.colab import drive
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import VotingClassifier
from sklearn.model_selection import cross_val_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
from sklearn.metrics import f1_score

drive.mount('/content/gdrive', force_remount=True)

df=pd.read_csv('/content/gdrive/My Drive/BreastCancer.csv')

df.head(6)

# To check the count of rows and columns
df.shape

# to see datatypes of the columns
df.dtypes

df['Class'].unique()
len(df[df.Class==1])/df['Class'].value_counts().sum()
len(df[df.Class==0])/df['Class'].value_counts().sum()
plt.figure()
sns.countplot(df.Class)
```

#missing value checking

```
df.isnull().sum()
# missing value Implementation
df['Bare.nuclei'] = df['Bare.nuclei'].fillna(np.mean(df['Bare.nuclei']))
df.isnull().sum()
```

#FEATURE REDUCTION :

```
df.corr()
%matplotlib inline
```

```
pd.plotting.scatter_matrix(df, figsize =(10,10))
plt.show()

# Drop Id & Mitoses column not used in analysis
df.drop(['Id','Mitoses','Cell.shape'], 1, inplace=True)

df.head(6)

plt.figure(figsize=(20,5))
plt.hist(df.iloc[:,:-1].T)
```

#Outlayer checking

```
df.iloc[:,:-1].boxplot(figsize=(20,5))
```

#Train Test split :

```
X = np.array(df.drop(['Class'], axis=1))
y = np.array(df['Class'])
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state
= 121)
```

#NORMALIZATION

```
# Scale the data. We will use the same scaler later for scoring function
scaler = MinMaxScaler().fit(X_train)
X_train = scaler.transform(X_train)
X_test = scaler.transform(X_test)
```

#Voting Classifier: Multiple Model Ensemble

```
lr = LogisticRegression(random_state = 45)
knn = KNeighborsClassifier(n_neighbors = 3, metric = 'minkowski', p = 2)
dt = DecisionTreeClassifier(max_depth=2)
Voting_Classifier = VotingClassifier(estimators=[('dt',dt),('lr',lr),('knn',knn)], votin
g= 'hard')
Voting_Classifier.fit(X_train, y_train)


print('Train Score:',Voting_Classifier.score(X_train,y_train)) \\Score
print('Test Scxore:',Voting_Classifier.score(X_test,y_test))


# Predicting the Test set results
y_pred = Voting_Classifier.predict(X_test)
```

#Result Analysis :

```
accuracies = (cross_val_score(estimator = Voting_Classifier, X = X_train, y = y_train, c
v = 5))
```

```
#applying cross validation.
print(accuracies)
print(accuracies.mean())
print(accuracies.std())


# Making the Confusion Matrix
print(confusion_matrix(y_test, y_pred))
# Get and reshape confusion matrix data
matrix = confusion_matrix(y_test, y_pred)
matrix = matrix.astype('float') / matrix.sum(axis=1)[:, np.newaxis]

# Build the plot
plt.figure(figsize=(7,5))
sns.set(font_scale=1.4)
sns.heatmap(matrix, annot=True, annot_kws={'size':10},
            cmap=plt.cm.Greens, linewidths=0.2)

# Add labels to the plot
class_names = ['Not Malignant','Malignant']
tick_marks = np.arange(len(class_names))
tick_marks2 = tick_marks + 0.5
plt.xticks(tick_marks, class_names, rotation=25)
plt.yticks(tick_marks2, class_names, rotation=0)
plt.xlabel('Predicted label')
plt.ylabel('True label')
plt.title('Confusion Matrix for the Model')
plt.show()


#Precision score

precision_score(y_test, y_pred)

#Recall score
recall_score(y_test, y_pred)

#F1 Score
f1_score(y_test, y_pred)

#Classification Report
print(classification_report(y_test, y_pred))
```

Result Analysis

Balanced or Imbalanced Checking:

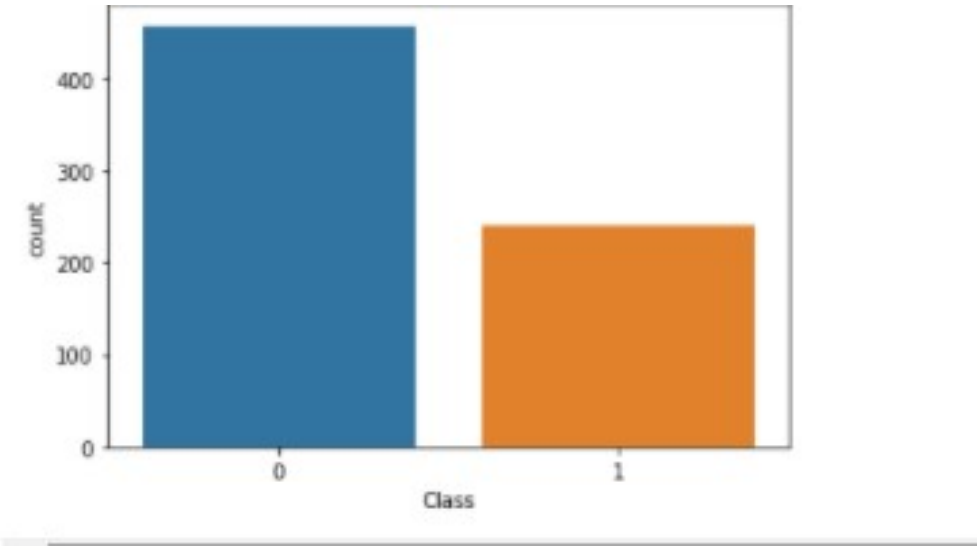

Figure-5

From **figure no -5** we can see that it is a balanced dataset.

Co-Relation :

	Id	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses	Class
Id	1.000000	-0.055308	-0.041603	-0.041576	-0.064878	-0.045528	-0.098668	-0.060051	-0.052072	-0.034901	-0.080226
Cl.thickness	-0.055308	1.000000	0.644913	0.654589	0.486356	0.521816	0.587300	0.558428	0.535835	0.350034	0.716001
Cell.size	-0.041603	0.644913	1.000000	0.906882	0.705582	0.751799	0.686801	0.755721	0.722865	0.458693	0.817904
Cell.shape	-0.041576	0.654589	0.906882	1.000000	0.683079	0.719668	0.709606	0.735948	0.719446	0.438911	0.818934
Marg.adhesion	-0.064878	0.486356	0.705582	0.683079	1.000000	0.599599	0.665049	0.666715	0.603352	0.417633	0.696800
Epith.c.size	-0.045528	0.521816	0.751799	0.719668	0.599599	1.000000	0.581261	0.616102	0.628881	0.479101	0.682785
Bare.nuclei	-0.098668	0.587300	0.686801	0.709606	0.665049	0.581261	1.000000	0.675896	0.577362	0.338740	0.816050
Bl.cromatin	-0.060051	0.558428	0.755721	0.735948	0.666715	0.616102	0.675896	1.000000	0.665878	0.344169	0.756616
Normal.nucleoli	-0.052072	0.535835	0.722865	0.719446	0.603352	0.628881	0.577362	0.665878	1.000000	0.428336	0.712244
Mitoses	-0.034901	0.350034	0.458693	0.438911	0.417633	0.479101	0.338740	0.344169	0.428336	1.000000	0.423170
Class	-0.080226	0.716001	0.817904	0.818934	0.696800	0.682785	0.816050	0.756616	0.712244	0.423170	1.000000

Figure-6

from figure **no-6** we can see –
The independent variables Cl.thickness,Cell.size,Cell.shape,Marg.adhesion,Epith.c.size,Bare.nuclei,Bl.cromatin,Normal.nucleoli have strong +correlation with the response variable class.

Scatter Matrix:

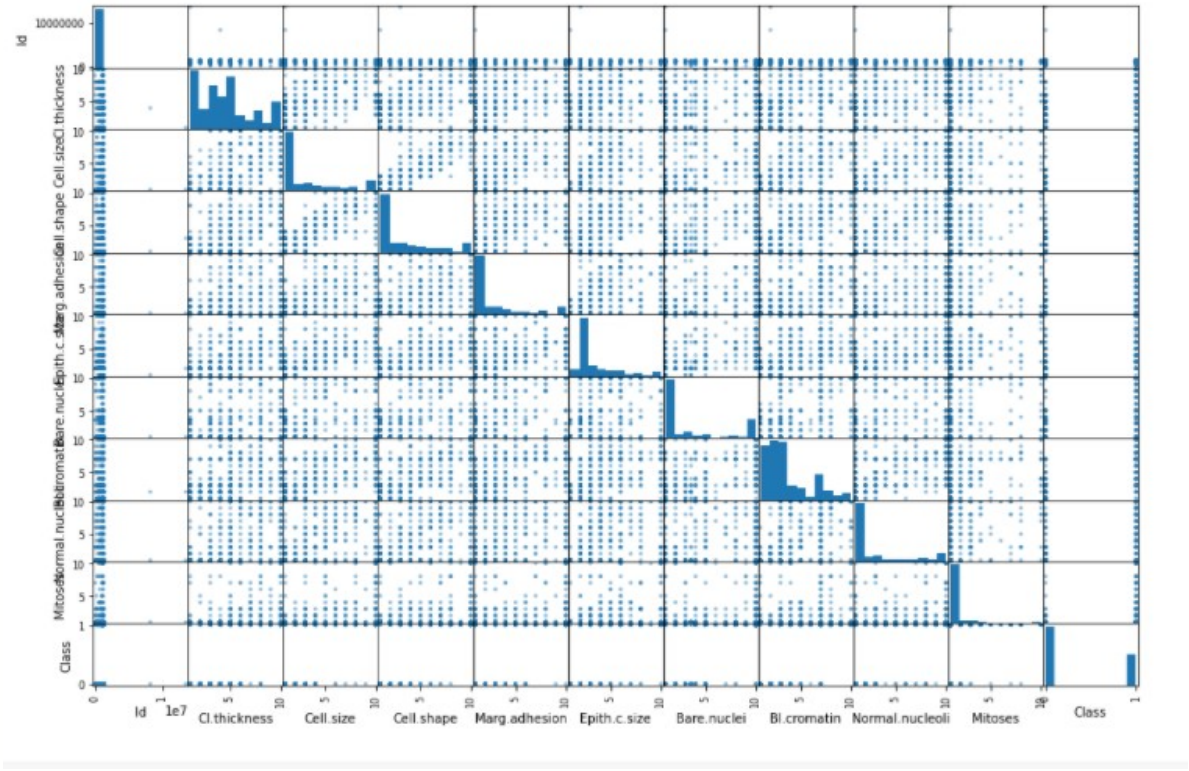

Figure-7

In **Figure no -7** from the diagonal position we can see the distribution of the input Features and from the off diagonal position we see the co-relation between them.

Labels:

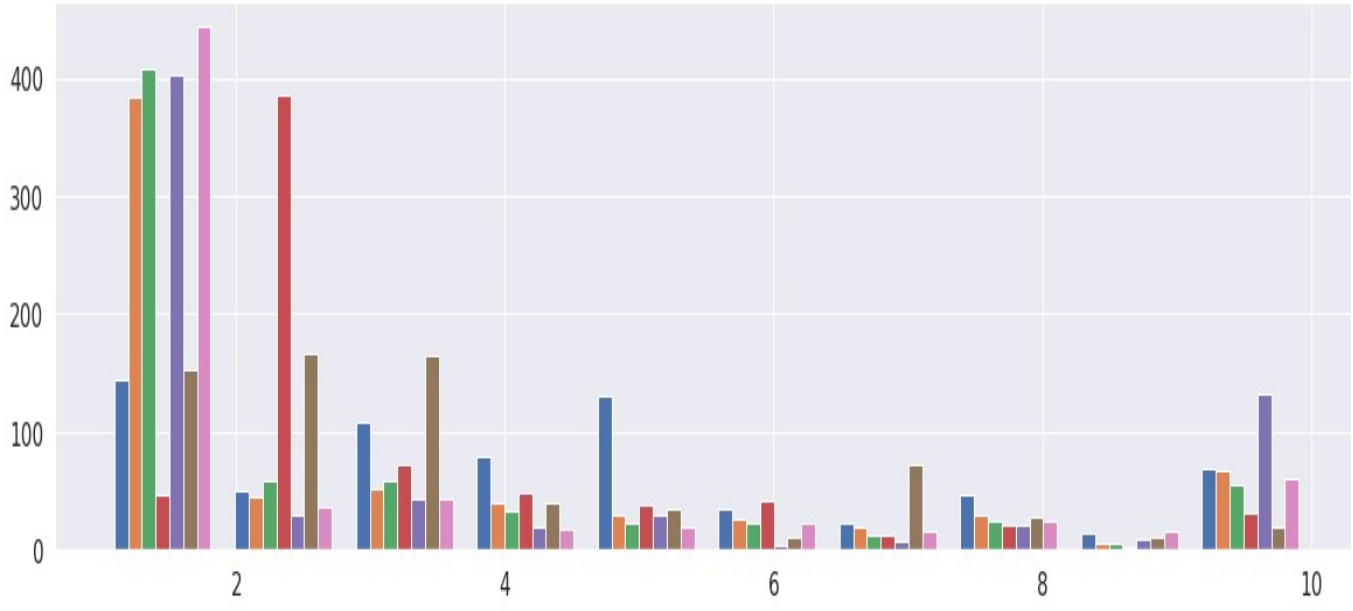

Figure-8

From **figure no-8** we can conclude that the input Features are not labeled.

Outliers checking :

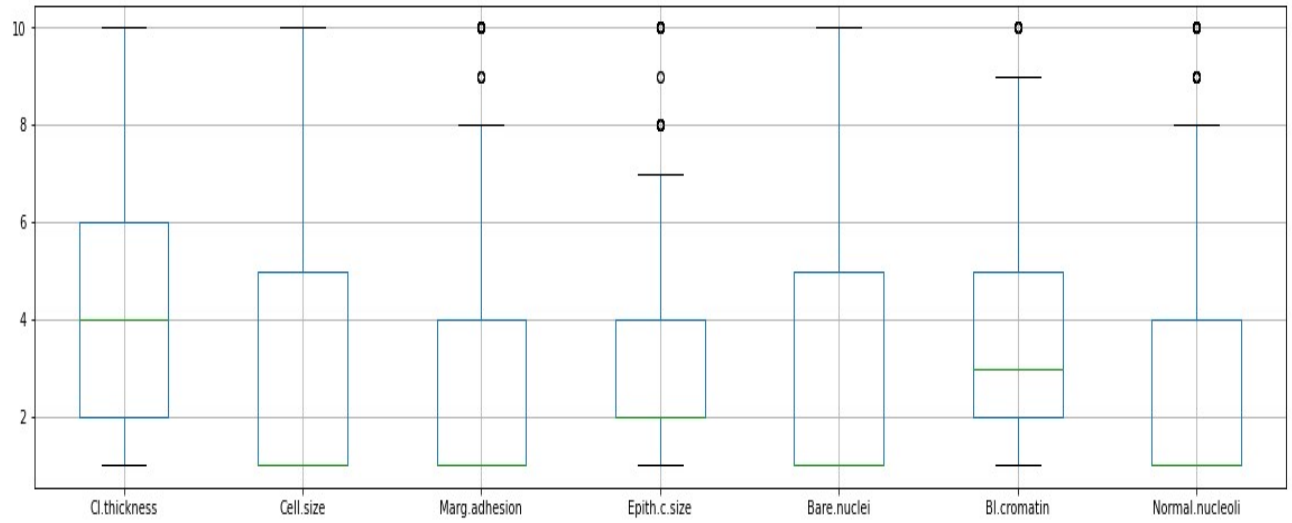

Figure-9

From **figure no-9** we can clearly see that the dataset contains several outliers.

Train Test Score :

```
Train Score: 0.9713774597495528
Test Scxore: 0.9642857142857143
```

Cross Validation :

Accuracy :

```
[0.97321429 0.94642857 0.95535714 0.97321429 0.96396396]
```

Mean Accuracy :

```
0.96243564993565
```

Standard deviation:

```
0.01040216849563941
```

Confusion Matrix :

```
[[85  2]
 [ 3 50]]
```

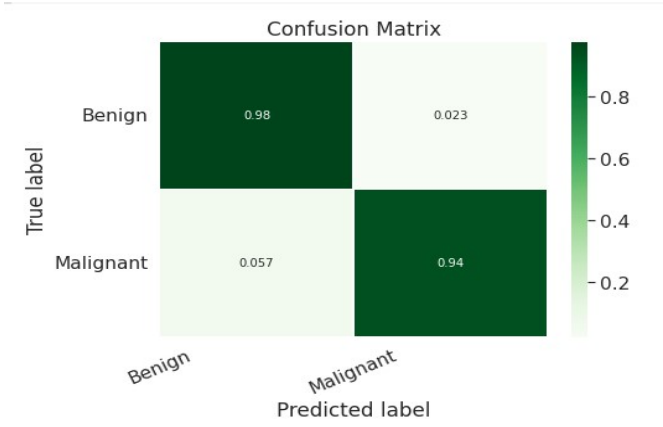

Figure-10

From **figure no-10** we can see that around 98% of Benign observations are correctly predicted as Benign Class and around 94% of Malignant observations are correctly predicted as Malignant Class.

Precision

0.9615384615384616

Recall

0.9433962264150944

f1 Score

0.9523809523809524

Classification Report :

	precision	recall	f1-score	support
0	0.97	0.98	0.97	87
1	0.96	0.94	0.95	53
accuracy			0.96	140
macro avg	0.96	0.96	0.96	140
weighted avg	0.96	0.96	0.96	140

Figure-11

From **Figure no – 11** we can see :

- For Class 0(Benign):
  - Precision-97%
  - Recall-98%
  - F1 Score-97%
- For Class 1(Malignant):
  - Precision-96%
  - Recall-94%
  - F1 Score-95%

Conclusion

Confusion matrix, precision, recall, and F1 score provides better insights into the prediction as compared to accuracy performance metrics. Applications of precision, recall, and F1 score is in information retrieval, word segmentation, named entity recognition, and many more.To choose our model we always need to analyze our dataset and then apply our machine learning model. Breast cancer if found at an early stage will help save lives of thousands of women or even men. These projects help the real world patients and doctors to gather as much information as they can. The research on nine papers has helped us gather the data for the project proposed by us. By using machine learning algorithms we will be able to classify and predict the cancer into being or malignant. Machine learning algorithms can be used for medical oriented research, it advances the system, reduces human errors and lowers manual mistakes.

References

1.https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/

2.https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3

3. http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic% 29

4.https://tatwan.github.io/How-To-Plot-A-Confusion-Matrix- In-Python/

5.https://seaborn.pydata.org/tutorial/axis_grids.html

6.https://seaborn.pydata.org/generated/seaborn.pairplot.html

7.https://seaborn.pydata.org/generated/seaborn.heatmap.html

8. "Ultrasound characterisation of breast masses", The Indian journal of radiology imaging by S. Gokhale., Vol. 19, pp. 242-249, 2009. K. Elissa, "Title of paper if known," unpublished.

9. "Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach" by Pragya Chauhan and Amit Swami, 18 October 2018

10. "On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset" by Abien Fred M. Agarap, 7 February 2019

11. "Analysis of Machine Learning Techniques for Breast Cancer Prediction" by the Priyanka Gupta and Prof. Shalini L of VIT university, vellore, 5 May 2018.

12. "Breast Cancer Diagnosis by Dierent Machine Learning Methods Using Blood Analysis Data" by the Muhammet Fatih Aslan, Yunus Celik , Kadir Sabanci and Akif Durdu, 31 December, 2018

13. "Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction", by Yixuan Li, Zixuan Chen October 18, 2018

14. "Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study" by Mumine Kaya Keles, Feb 2019

15. "Breast Cancer Prediction Using Data Mining Method " by Haifeng Wang and Sang Won Yoon, Department of Systems Science and Industrial Engineering State University of New York at Binghamton Binghamton, May 2015.

16. "Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis" by Wenbin Yue , Zidong Wang, 9 May 2018

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

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