Breast Cancer Diagnosis by exploiting data exploration techniques

Submitted in partial fulfillment for the degree of

BACHELOR OF ENGINEERING
in
COMPUTER ENGINEERING

By

Prathiksha Poojary 116A1059
Shruti Rao 116A1066
Jigar Somaiya 116A1033

UNDER THE GUIDANCE OF
Prof.Pranita Mahajan

September 16, 2020

DEPARTMENT OF COMPUTER ENGINEERING
SIES GRADUATE SCHOOL OF TECHNOLOGY
NERUL, NAVI MUMBAI – 400706
ACADEMIC YEAR 2019– 2020

CERTIFICATE

This is to certify that the project titled **"Breast Cancer Diagnosis by exploiting data exploration techniques"** is a bonafide work carried out by the following students,submitted to the University Of Mumbai in partial fulfillment of the requirement for the award of the degree of **Bachelor of Engineering** in **Computer Engineering**

1. Prathiksha Poojary	116A1059
2. Shruti Rao	116A1066
3. Jigar Somaiya	116A1033

Internal Guide	**Head of Department**	**Principal**
Prof.Pranita Mahajan	Dr.Aparna Bannore	Dr.Atul Kemkar

PROJECT REPORT APPROVAL

This project report entitled **"Breast Cancer Diagnosis by exploiting data exploration techniques"** by following students is approved for the degree of **Bachelor of Engineering in Computer Engineering**

1.	Prathiksha Poojary	116A1059
2.	Shruti Rao	116A1066
3.	Jigar Somaiya	116A1033

Name of External Examiner: ______________________________

Signature: ____

Name of Internal Examiner: ______________________________

Signature: ____

Date:

Place:

DECLARATION

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Prathiksha Poojary	116A1059	______________
Shruti Rao	116A1066	______________
Jigar Somaiya	116A1033	______________

Date:

ACKNOWLEDGEMENT

We wish to express our deep sense of gratitude to thank our project guide Prof. Pranita Mahajan for providing timely assistant to our query and guidance. We take this opportunity to thank our HOD Dr.Aparna Bannore and Principal Dr. Atul Kemkar for their valuable guidance and immense support in providing all the necessary facilities.

We would also like to thank the entire faculty of CE Department for their valuable ideas and timely assistance in this project.Last but not the least, we would also like to thank teaching and nonteaching staff members of our college for their support, in facilitating timely completion of this project.

Project Team

Prathiksha Poojary

Shruti Rao

Jigar Somaiya

CONTENTS

ABSTRACT

It is previously known that data preprocessing lays the groundwork when working with raw datasets. Before the discovery of useful information/knowledge, the target dataset must be properly preprocessed. But it is unfortunately ignored by the most researchers due to its perceived difficulty and time required to perform them.In this paper, we research the influence of data preprocessing and also the effects of over and under preprocessing.This paper aims to present comparison of the largely popular data preprocessing techniques and their effect on different data classification algorithms. The Wisconsin Diagnosis Breast Cancer data set was used as a training set to compare the performance of the various machine learning techniques in terms of key parameters such as accuracy, and precision.[2] The results obtained are very competitive and convey that not all data preprocessing methods are necessary. Experiments about some algorithms with different preprocessing methods also confirm that preprocessing has a great influence on the performance of a classifier.

List of Figures

List of Tables

List of Abbreviations

DT	Decision Tree
KNN	K-Nearest Neighbour
LDA	Linear Discriminant Analysis
LR	Logistic Regression
ML	Machine Learning
MLP	Multi-Layer Perceptron
NB	Naive Bayes
PCA	Principal Component Analysis
RF	Random Forest
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine

Chapter 1

INTRODUCTION

In 2018, it is estimated that among women 627,000 women died from breast cancer, that is approximately 15 % of all cancer deaths[14]. In order to improve the survival rate,the most important factor is early detection of the tumor. Every year, breast cancer kills more than 500,000 women around the world. In resource-poor settings, a majority of women with breast cancer are diagnosed at an advanced stage of disease; their five-year survival rates are low, ranging from 10-40%[15] The five-year survival rate for early localized breast cancer exceeds 80%, in settings where early detection and basic treatment are available and accessible.[15]For detection,the two strategies are early diagnosis and screening[16]. There are many procedures for detection of breast cancer. It can be done by Physical examination by the physician to check for lumps. It can also be done using mammograms. A mammogram is an X-ray of the breast which is used for screening of breast cancer. If an abnormality is detected on a screening mammogram, further diagnostic mammogram is recommended.Breast ultrasound uses sound waves in order to produce images of structures deep within the body. Ultrasound is used to determine if a new breast lump is a fluid-filled cyst or a solid mass.The only definitive way for the diagnosis of breast cancer is a biopsy. A specialized needle device guided by X-ray or another imaging test is used to extract a core of tissue from the region where the tumor is suspected.Biopsy samples are sent to a laboratory for analysis where experts determine whether the cells are cancerous. Using Magnetic resonance imaging (MRI),pictures of the interior of the breast are created. To create this picture, an MRI machine uses a magnet and radio waves.[16]

NEED

Breast Cancer is the most common female cancer worldwide. All most 25% of all cancers with an estimated 1.67 million new cancer cases diagnosed in 2012.[14] The life time risk of developing breast cancer in women is approximately 1/8 in USA, 1/ 12 in Europe, 1/40 in Asia [WHO 2008]. In a lot of cases, due to false-negative diagnosis,the tumor worsens or may lead to death.So to prevent this,there is a need for systems that can diagnose cancer with more accuracy[14].Detection in early stages is important as the disease is curable if detected in early stages.

ORGANIZATION OF THE REPORT

Chapter 1 gives introduction of the topic.Chapter 2 discusses the existing literature.Chapter 3 explains the proposed system. Chapter 4 discusses the Design and Methodology which explains the dataset,performance metrics and hardware and software details.Chapter 5 discusses the results of the computations.Chapter 6 explains the conclusion and future scope of the topic.

Chapter 2

LITERATURE SURVEY

In this report a brief review of the area of study –Breast Cancer Detection using Machine Learning. By doing so, the research may be guided accordingly by firstly discovering where the research is coming from, what and how much have been studied regarding the topic and what it is yet to tackle. This provides background to the research, it will provide the necessary backbone and support to the research. By reviewing the past publications and researches related to the study, the researcher will have an idea of how such a study has been done in the past. In this way, this research may be able to reflect, compare itself, learn from setbacks and produce a stronger and more efficient study. The various research papers we have referred to are as follows:

Comparison of different Machine Learning methods for Breast Cancer diagnosis[1]
Purpose / Key Findings:
1.Comparison of SVM and ANN based on accuracy,Precision,recall,ROC Area.
2.SVM gives better accuracy as compared to ANN.
Gaps / Issues:
No pre-processing of dataset done.

Breast Cancer detection using Machine Learning Algorithms[2]
Purpose / Key Findings:
1.Comparison between Random Forest,KNN,Naive Bayes.
2.KNN gives the highest accuracy,followed by Random Forest.
Gaps / Issues:
Naive Bayes gives lowest accuracy due to comparatively smaller dataset size.

Breast Cancer Classification using Machine Learning[3]
Purpose / Key Findings:
1.Comparison between Naive Bayes and KNN and evaluation of accuracy using cross validation.
2.KNN Gives Highest accuracy
Gaps / Issues:

KNN gives less acuuracy when dataset is large due to increase in time complexity.

Using Machine Learning algorithms for breast cancer risk prediction and diagnosis [4]

Purpose / Key Findings:

1.In this paper,the algorithms Support Vector Machine (SVM), Decision Tree (CART), Naive Bayes (NB) and k Nearest Neighbours (k-NN) are used.[4]
2.The data is divided into training data and testing data.
3.Compared to the other algorithms,SVM algorithm's performance was drastically bad.
3.When we standardize the input, performance of SVM drastically increases while the performance of CART and Naive Bayes is poor.[4]
4.k-NN gave the best performance overall,followed by Naive bayes and Logistic regression[4]

Gaps / Issues:

1.The SVM is only applicable when the number of class variable is binary i.e. we can't have more than 2 classes.[4]
2.No pre-processing of data is performed.

Using Data Mining Tools for Breast Cancer Prediction and Analysis [5]

Purpose / Key Findings:

1.Here the missing values are replaced by the median of the values of individual attributes.[5]
2.Here,four data mining algorithms were used: Bayes classifier (Naive Bayes, Bayesian Logistic Regression), Decision Tree (J48, simple CART).[5]
3.WEKA software was used.[5]
4.After comparison,it was found that Simple CART decision tree algorithm gave the best accuracy.[5]
5.Naive Bayes is better than Bayesian Logistic Regression and CART is better than J48

Gaps / Issues:

Simple CART takes a lot of execution time.i.e its time complexity is more.

Deep Learning Algorithms for Predicting Breast Cancer Based on Tumor Cells[6]

Purpose / Key Findings:

1.They also compared the deep learning activation functions with other classification methods like Naïve Bayes(NB), Decision tree(DT), Support Vector Machine(SVM), Vote(DT+NB+SVM), Random Forest(RF) and AdaBoost
2.Highest Accuracy was given by Exponential rectifier.
3.The best precision of malignant and benign was were given by SVM and random forest respectively.

Gaps / Issues:

1.Different methods gave better performance with respect to a particular parameter. For eg.Highest Accuracy was given by Exponential rectifier. The best precision of malignant and benign was were given by SVM and random forest respectively.
2.No method gave best performance with respect to all the parameters.

Prediction of Breast Cancer using Support Vector Machine and K-Nearest Neighbour.[7]

Purpose / Key Findings:

1.Used SVM and KNN to predict breast cancer.
2.SVM gives better performance in all parameters
3.10 fold cross validation was used. 1 chunk was used for testing,rest 9 were used for training.

Gaps / Issues:

The false discovery rate is comparatively high in K-NN rather than SVM.

Breast Cancer Diagnosis Using Genetic Algorithm for Training Feed Forward Back Propagation.[8]

Purpose / Key Findings:

1.GA is used as a mechanism to get the ideal weights for FFBPNN.
2.When GA was used, it gave better results than classic FFBPNN.

Gaps / Issues:

Pre-processing techniques have not been applied on algorithms without GA.

Breast cancer diagnosis using GA feature selection and rotation forest.[9]

Purpose / Key Findings:

1.GA feature selection and different data mining techniques, namely Logistic Regression,Random Forest,Rotation forest etc.

2.Algorithms with GA were better.(Rotation forest gave the best accuracy)

Gaps / Issues:

Preprocessing techniques have not been applied on algorithms without GA.

Breast Cancer Diagnosis Using Deep Learning Algorithm.[10]

Purpose / Key Findings:

1.They have encoded, standardized and normalized the dataset.

2.They have preprocessed the data using PCA.

3.CNN (with appropriate Preprocessing techniques) gives very high accuracy.

Gaps / Issues:

1.Only used PCA model.

2.They have not tried any other model for comparison

On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset.[11]

Purpose / Key Findings:

1.To diagnose breast cancer using Machine Learning.

2.SVM algorithm was used.

Gaps / Issues:

Inadequate preprocessing of the data.

Chapter 3

PROPOSED SYSTEM

Pre processing is the transformations that are applied on the data before feeding it to the algorithm.Data preprocessing it is the technique which is used to convert raw data into clean data so it is feasible for analysis.For achieving better results data should be in proper format. Moreover data should be in such a format that many machine and deep learning algorithms can be applied and we can choose one which gives better results.

3.1 PREPROCESSING TECHNIQUES

A)label Encoder: Usually datasets contain multiple labels in one or more than one columns, they can be in numbers or words but to make it human readable they are usually in words. Label encoding refers to process of converting labels into numeric form to make it machine readable.

B)Normalization:An attribute in a dataset may contain values with varying scale so Machine Learning algorithms always benefits by converting these values to a common scale. It is applied as a part of data preparation and its main goal is to convert numeric values in a column to a common scale without distorting differences in the range of values.

We generally use minmax scaler for this. Formula for min max scaler:

X-min/(max-min), where min is minimum value and max is maximum value in column

C)Standardization:It is a useful technique to convert attributes with Gaussian distribution and with varying mean and standard deviation to standard Gaussian distribution with mean of 0 and standard deviation of 1.

Generally standard scalar is used for this and formula for the same is:

z=x-mean/standard deviation

D)Linear Discriminant Analysis(LDA):It is the most commonly used dimension reduction technique in machine learning applications. It decomposes dataset into lower dimensions , in addition to that it reduces over fitting of data and reduces the computational cost. The approach of LDA is similar to PCA but in addition to finding component axes that maximise the variance , we find axes that maximise separation between multiple classes.

Steps for LDA:

1.Computing d-dimensional mean vectors:

For every class, every feature finding mean vector.

2.Computing the scatter matrices:

To determine the relationship of each feature with every other feature.

3.Decomposition of square matrix into eigenvector and eigenvalues.

4.Selecting linear discriminates for new feature subspace:

Sorting of eigenvectors by decreasing order of eigen values. Choosing K eigen- vectors with largest eigenvalues.

5.Transforming samples onto the new subspace: Mapping the samples to the new feature subspace.

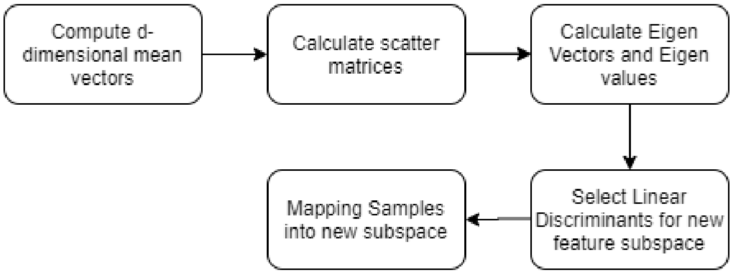
Compute d-dimensional mean vectors
Calculate scatter matrices
Calculate Eigen Vectors and Eigen values
Select Linear Discriminants for new feature subspace
Mapping Samples into new subspace

Figure 1: Flow Diagram for LDA

E)Principal Component Analysis(PCA):Principal Component Analysis is a method to reduce the number of variables in a dataset. It does by combining highly correlated variables together. PCA allows us to represent data along one axis and that axis is called principal component. It simplifies higher dimensions to lower dimensions.

Steps for PCA:

1.Standardize:

Standardize the given data.

2.Calculate Covariance:

Find covariance matrix for the given data. Covariance is a measure of how two variables move together.

3.Deduce Eigen Vectors:

Our main principal component becomes X axis and axis perpendicular to it becomes Y axis , then we need to fit our data to two axes. So we need to find out eigen vectors as eigen vectors indicates the directions of new axes.

4.Reorient data:

To reorient the data according to new axes we multiply our data with eigen vectors. This reoriented data is score.

5.Plot the Reoriented data or score.

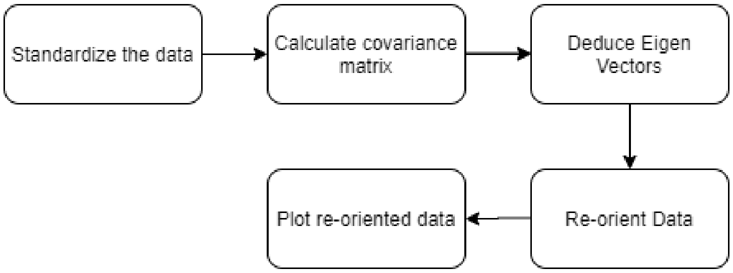

Figure 2: Flow Diagram for PCA

3.2 COMPONENTS OF THE SYSTEM

Machine learning (ML) is an application of artificial intelligence (AI) which has the ability to learn automatically and improve itself from experience without being explicitly programmed.

Machine learning aims on developing computer programs that can access data and use it for learning.Machine learning algorithms are programs which when exposed to more data, adjust themselves to perform better. Machine-learning algorithms have a specific way of adjusting its own parameters, depending on the feedback on its previous performance it makes predictions about a dataset.The main aim of these algorithms is allow the computers to learn automatically without human intervention.

A) ML algorithms are categorized as supervised, unsupervised, semi- supervised and reinforcement learning. Supervised learning - In this some data is already tagged with the correct answer and we train the machine using this labelled data. It generates a function predicting outputs based on input observations. Unsupervised learning - This uses information that is neither classified nor labeled and allows the algorithm to act on that information without guidance. Here, the machine is forced to train from an unlabeled dataset and then differentiate it on the basis of some characters. Semi-supervised learning - This algorithm is trained upon a combination of labeled and unlabeled data. This learning combines a small amount of labeled data with a large amount of unlabeled data during training. Reinforcement learning - The learning happens from the environment and this learning employs rewarding desired behaviours and punishing the undesired ones.

B) K-Nearest Neighbours (kNN) - KNN is a supervised learning algorithm that can be used to solve classification and regression problems. It assumes that similar things are in close proximity. It uses a database in which the data points are separated into several classes to predict the classification of a new sample point. It is non-parametric. Here ,there is no explicit training phase. KNN when used for classification, the output is a class membership which gives a discrete value. And when it is used in regression, the output is the value of the object which gives continuous values.

C) Naive Bayes (NB) - It is a supervised classification technique which assumes independence among predictors and it is a probabilistic algorithm . It is called naive because it assumes that all features are independent from each other, this is generally not the case in real life scenarios, but still Naïve Bayes proves to be efficient for a wide variety of machine learning problems.

D) Support Vector Machine (SVM) - SVM is a supervised machine learning algorithm.It is based on finding the hyperplane that best divides the dataset into 2 classes. In this, each data item is plotted as a point in n-dimensional space with the value of each feature being the value of a particular coordinate.

E) Decision Tree (DT) - To build a decision tree there is no need to normalize the data. DT builds classification and regression models in a tree structure where it breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The result of this is a tree with decision nodes and leaf nodes.

F) Random Forest (RF) - RF is a classification algorithm which contains several decision trees. While building each tree, it makes use of bagging and feature randomness for creation of an uncorrelated forest of trees whose prediction is more accurate than prediction of an individual tree The DT in the forest considers a random subset of features for forming questions and has only access to a random set of training data points.

G) Logistic Regression (LR) - LR is used to find a relationship between features and probability of particular outcome.Logistic Regression uses Sigmoid function. In LR, setting a threshold value is very important as classification problems depend on it. The value of recall and precision affect the decision for the value of threshold.

H) Multilayer Perceptron (MLP) - It is usually applied to supervised problems.Multilayer perceptron consists of more than one perceptron, it is a deep, artificial neural network. MLP with one hidden layer is used for approximating continuous function. They train an input-output pair and learn to model the correlation between input and output.

I) Stochastic Gradient Descent (SGD) - In SGD a single random sample is selected rather than the entire dataset for each iteration. In this, the gradient of the cost function of a single example at each iteration is found instead of the sum of the gradient of the cost function of all the examples. Here we reach the minima with a shorter training time.

J) Adaboost - Adaboost is a boosting technique which combines many weak classifiers into a single strong classifier. It is a technique that builds on top of other classifiers. Based on the results of the previous classifier, it chooses the training set for each new classifier that you train. It determines the weight that should be given to each of the classifiers proposed answer when combining the results.

Chapter 4

DESIGN AND METHODOLOGY

4.1 ARCHITECTURE

The Architecture of the proposed system is as shown in the Figure1.

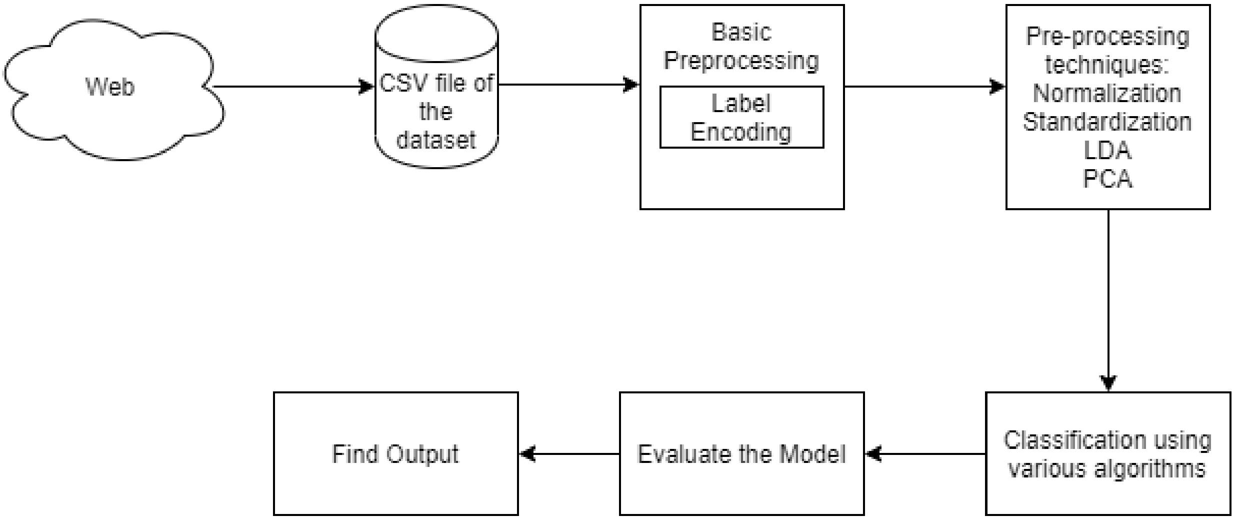
Web
CSV file of the dataset
Basic Preprocessing
Label Encoding
Pre-processing techniques: Normalization Standardization LDA PCA
Classification using various algorithms
Evaluate the Model
Find Output

Figure 3: Proposed System Architecture

The main methodology consists of two parts:
1. Pre-processing.
2. Classification Algorithms.

Data Preprocessing:

It is a technique to transform data into understandable format. Real world data is often inconsistent, incomplete and contains many errors. So data preprocessing resolves all these issues and prepares data for further processing.

In the proposed system, we are going to use the following pre-processing techniques:
1)Label Encoding
2)Standardization
3)Normalization
4)Linear Discriminant Analysis(LDA)
5)Principal Component Analysis(PCA)

Classification:
Classification is a technique where we categorize data into a given number of classes. The main goal of a classification problem is to identify the category/class to which a new data will fall under.

4.2 DATASET:

The Dataset used for the project is the Diagnostic Wisconsin Breast Cancer Dataset.The dataset as been obtained from the UCI Machine Learning Repository.It has 569 instances and 32 attributes[17].There are no missing values.There are 2 classes Malignant and Benign.[17]

Train-test split has been used on the data where a constant test size of 0.2 has been used across the various algorithms.The positive class is Malignant and the negative class is Benign.

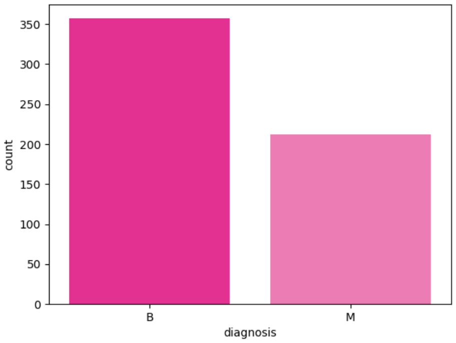

Figure 4: Count of Malignant and Benign cases in the dataset

Attribute Information:[17]

1) ID number
2) Diagnosis (M = malignant, B = benign)
3-32)Ten real-valued features are computed for each cell nucleus. They are:
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 (”coastline approximation” - 1)

Resulting in 30 features,the mean, standard error and ”worst” or largest (mean of the three largest values) of these features were computed for each image.[12][13]

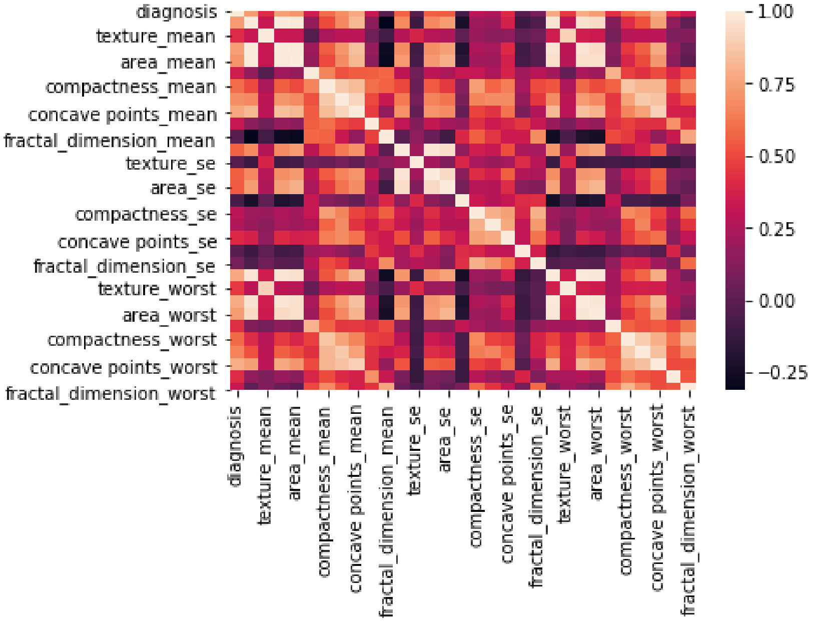

Figure 5: Count of Malignant and Benign cases in the dataset

4.3 PERFORMANCE METRICS:

The parameters that are used for determining the performance of the various Machine learning algorithms are described in this section.A confusion matrix is derived of the actual and predicted values.It comprises of four values namely: True Positive(TP),False Positive(FP),False Negative(FN),True Negative(TN).

Using these four values,the following performance metrics are derived:

Accuracy:

It can be described as the ratio of correctly predicted observations to the total number of observations.

The formula is as follows:

Accuracy=TP+TN/TP+TN+FP+FN

Precision:

It can be described as the ratio of correctly predicted positive observations to the total predicted positive observations.

The formula is a follows:

Percision=TP/TP+FP

Recall:

It can be described as the ratio of correctly predicted positive observations to all observations in actual class.

The formula is a follows:

Recall=TP/TP+FN

F1 score:

It can be described as the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

The formula is a follows:

F1 score=2*(precision*recall)/Precision+recall

4.4 DETAILS OF HARDWARE AND SOFTWARE:

All the computations have been implemented on a computer having the specifications of Intel Core i5 with 8GB RAM.An open source web application called Jupyter Notebook has been used for running the programs.In the programs numpy,pandas,Scikit-learn libraries have been used.These are open source machine learning libraries in python.Matplotlib library has also been used for visualization.

Chapter 5

RESULTS AND DISCUSSIONS

A comparative study between various Machine learning algorithms,with and without preprocessing has been proposed. The various classifiers were tested by performing train-test split on the dataset.The dataset was divided in a ratio of 80-20.Out of the 569 observations,455 were used for training the classifiers and the remaining 114 were used for testing.

For each algorithm,we computed the results for the following:

1.Algorithm without any preprocessing

2.Algorithm with normalization of the data

3.Algorithm with standardization of the data

4.Algorithm with normalization and Standardization of data

5.Algorithm with PCA:

a.PCA (number of components= 8,9,15)
b.PCA with normalization
c.PCA with standardization
d.PCA with normalization standardization
6.Algorithm with LDA:
a.LDA

b.LDA with normalization
c.LDA with standardization
d.LDA with normalization and standardization

5.1 PERFORMANCE OF CLASSIFICATION ALGORITHMS WITH AND WITHOUT PREPROCESSING TECHNIQUES

1. k-Nearest Neighbours(KNN): Accuracy of KNN algorithm without performing any preprocessing on the dataset is 91.22 percentage.

When the dataset was preprocessed using LDA, the best accuracy of 97.36 percentage was obtained.

Following values of Precision,Recall and F1 Score was observe:

Precision-97 percentage.
Recall-97percentage.
F1 score-97 percentage

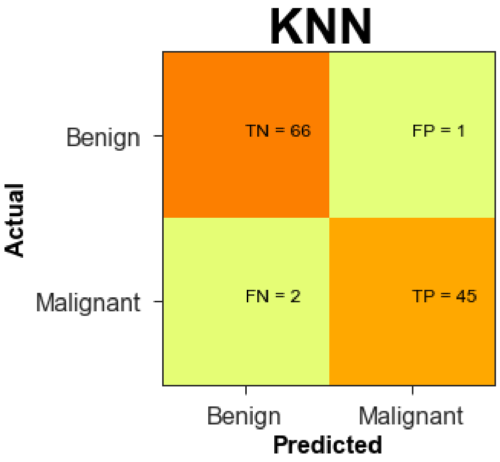

Figure 6: Confusion Matrix of KNN

2. Naive Bayes(NB): The accuracy of naive bayes without performing any preprocessing on Wisconsin dataset is 92.98 percentage.

When the dataset was preprocessed using LDA, the best accuracy of 96.49 percentage was obtained.

Following values of Precision,Recall and F1 Score was observe:

Precision-97 percentage
Recall-96 percentage.
F1 score-96 percentage.

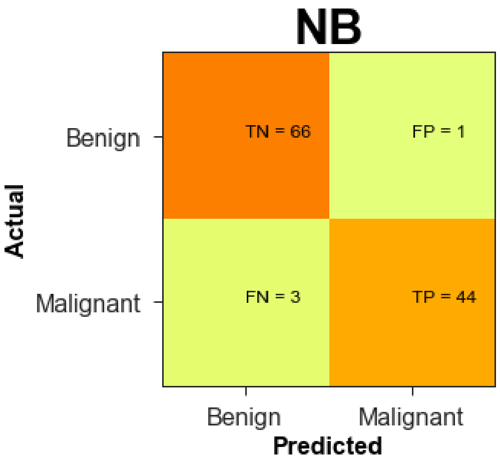

Figure 7: Confusion Matrix of Naive Bayes

3. Support Vector Machine(SVM): The accuracy of SVM without preprocessing on Wisconsin dataset is 58.77 percentage.

When the dataset was preprocessed using standardization, the best accuracy of 98.24 percentage was obtained.

Following values of Precision,Recall and F1 Score was observe:

Precision-98 percentage.
Recall-98 percentage.
F1 score-98 percentage.

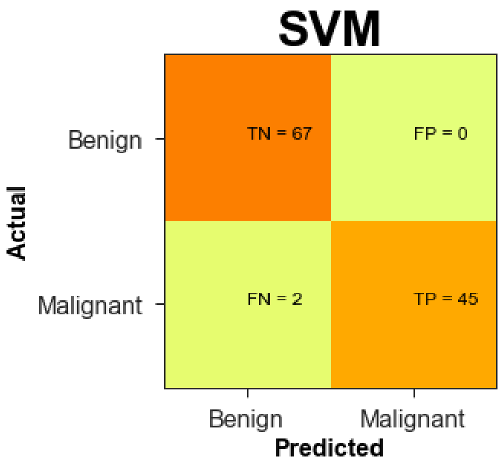

Figure 8: Confusion Matrix of SVM

4. **Decision Tree(DT):** The accuracy of DT without any preprocessing on wisconsin dataset is 91.2 Percentage.

When the dataset was preprocessed using LDA, the best accuracy of 95.6 percent was obtained.

Following values of Precision,Recall and F1 Score was observe:

Precision-96 percentage.
Recall-96 percentage.
F1 score-96 percentage.

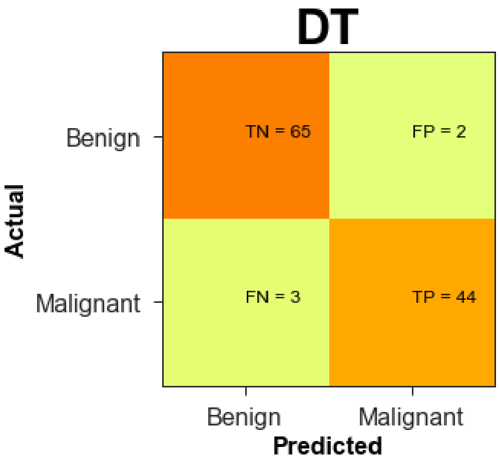

Figure 9: Confusion Matrix of DT

5. Random Forest(RF): The accuracy of RF without any preprocessing on wisconsin dataset is 98.24 percent.

When the dataset was preprocessed using standardization or normalization, the best accuracy of 98.24 percent was obtained.

Following values of Precision,Recall and F1 Score was observe:

Precision-98 percentage.
Recall-98 percentage.
F1 score-98 percentage.

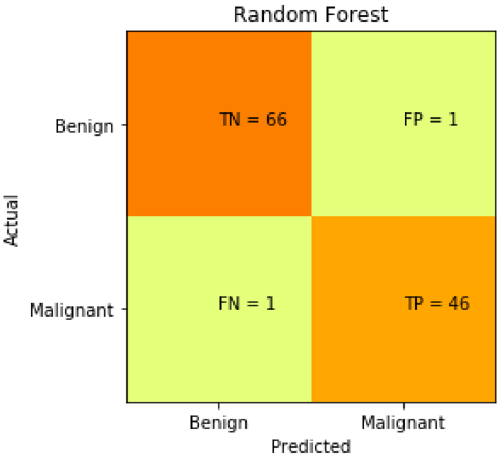

Figure 10: Confusion Matrix of RF

6. Logistic Regression (LR): The accuracy of LR without any preprocessing is 95.6 percent.

When the dataset was preprocessed using LDA, the best accuracy of 97.3 percent was obtained.

Following values of Precision,Recall and F1 Score was observe:

Precision-97 percentage.
Recall-97percentage.
F1 score-97 percentage.

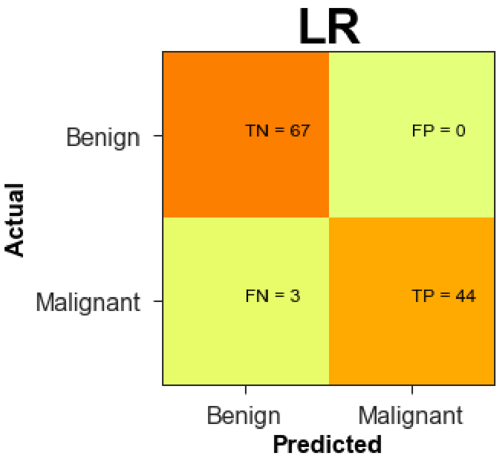

Figure 11: Confusion Matrix of LR

7. Multilayer Perceptron (MLP): The accuracy of MLP without applying any preprocessing is 90.9 percent.

When the dataset was preprocessed using normalization, standardization and PCA, best accuracy of 99.1 percent was obtained.

Following values of Precision,Recall and F1 Score was observe:

Precision-99 percentage.
Recall-99 percentage.
F1 score-99 percentage.

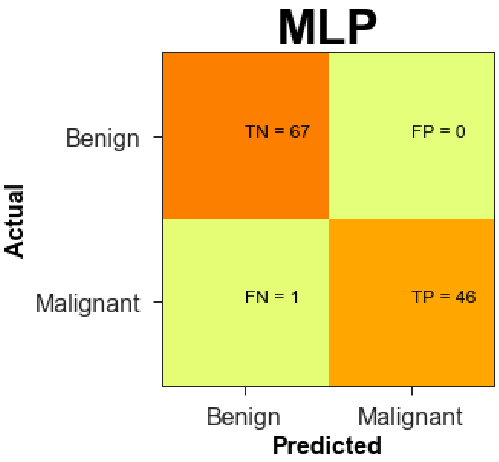

Figure 12: Confusion Matrix of MLP

8. Stochastic Gradient Descent (SGD): The accuracy of SGD without any preprocessing is 82.4 percent.

When the dataset was preprocessed using LDA, the best accuracy of 97.36 percent was obtained.

Following values of Precision,Recall and F1 Score was observe:

Precision-97 percentage.
Recall-97 percentage.
F1 score-97 percentage.

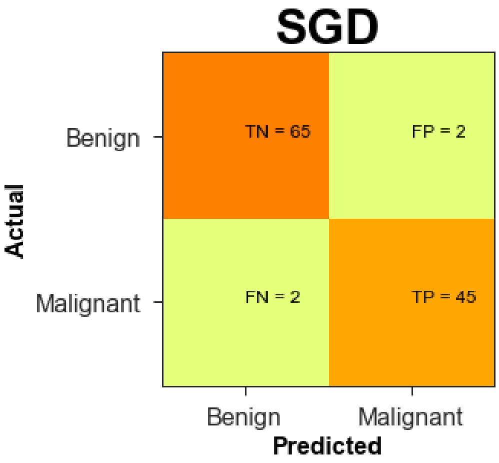

Figure 13: Confusion Matrix of SGD

9. Adaboost: The accuracy of Adaboost without any preprocessing is 95.61 percent.

When the dataset was preprocessed using PCA with standardization and normalization, the best accuracy of 97.36 percent was obtained.

Following values of Precision,Recall and F1 Score was observe:

Precision-97 percentage.
Recall-97 percentage.
F1 score-97 percentage.

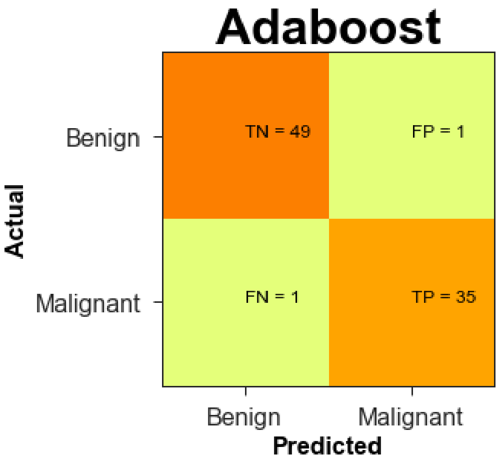

Figure 14: Confusion Matrix of Adaboost

5.2 RESULTS:

The best performing preprocessing technique for each classification algorithm respectively is as follows:

Algorithm	**Preprocessing techniques having best performance**
DT	LDA PCA with normalization and Standardization
LR	LDA PCA with normalization
MLP	PCA with normalization
SGD	LDA PCA with normalization and Standardization
SVM	Standardization
RF	Standardization Normalization
Adaboost	Standardization PCA with normalization and Standardization
KNN	LDA
NB	LDA

Table 1: The best preprocessing technique for each classification algorithm respectively

The performance of the Classification Algorithms without preprocessing and with the best performing preprocessing technique respectively is as follows:

1.DECISION TREE

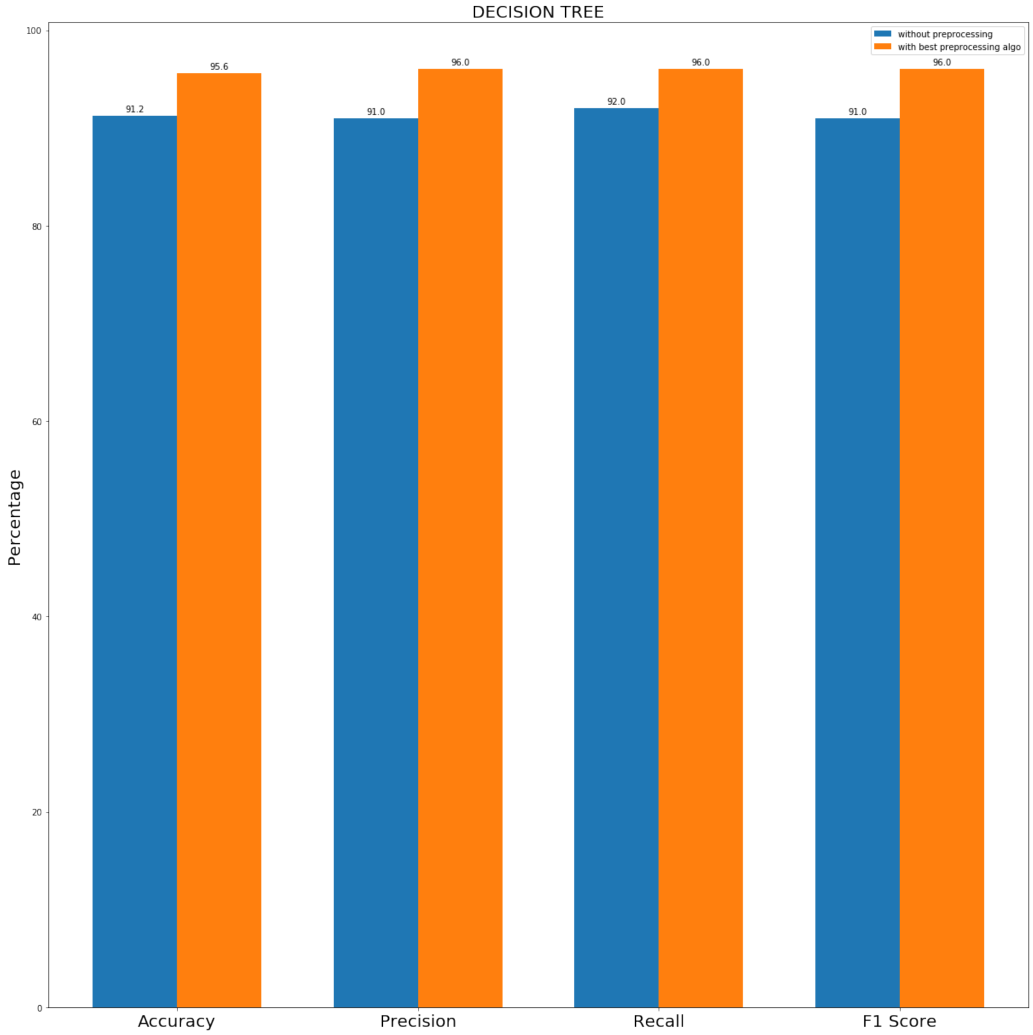

Figure 15: Performance metrics of Decision Tree

2.LOGISTIC REGRESSION

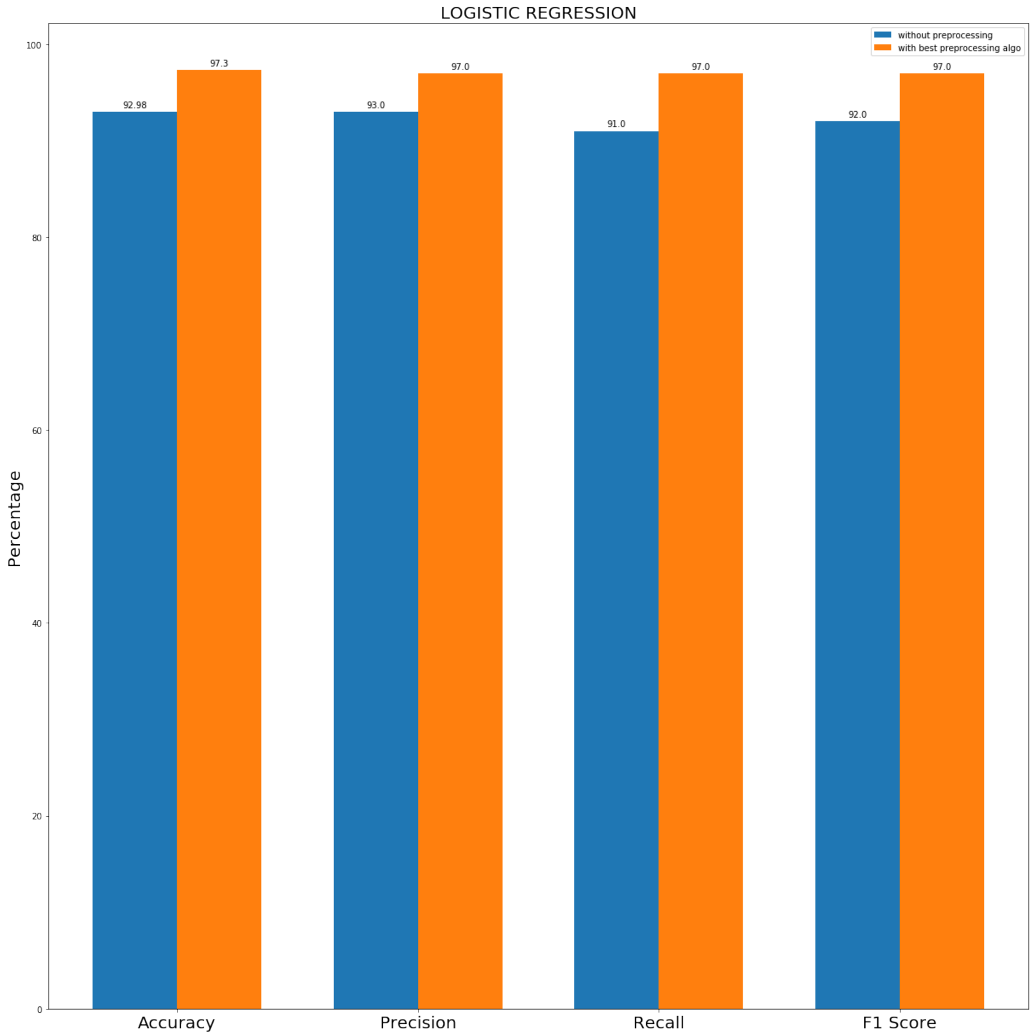

Figure 16: Performance metrics of Logistic Regression

3.MULTI-LAYER PERCEPTRON

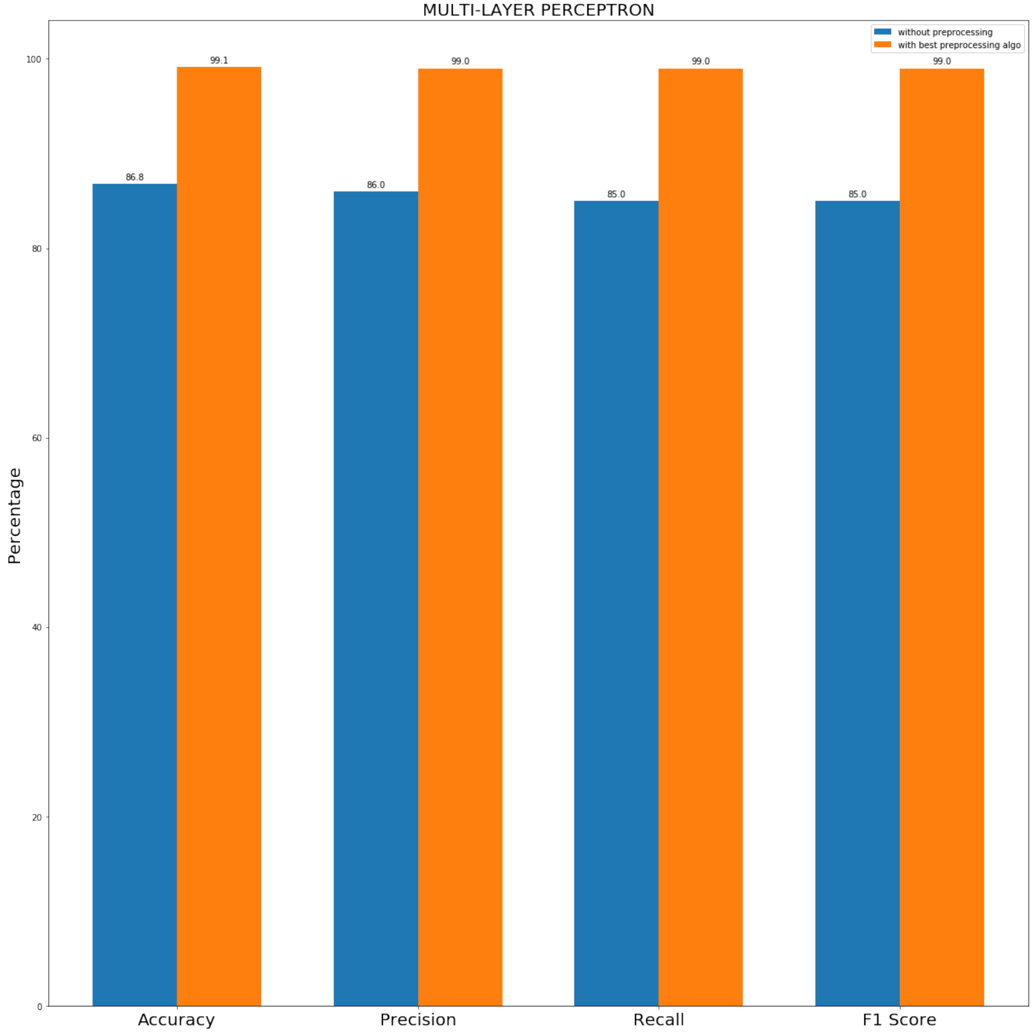

Figure 17: Performance metrics of Multi-layer Perceptron

4.STOCHASTIC GRADIENT DESCENT

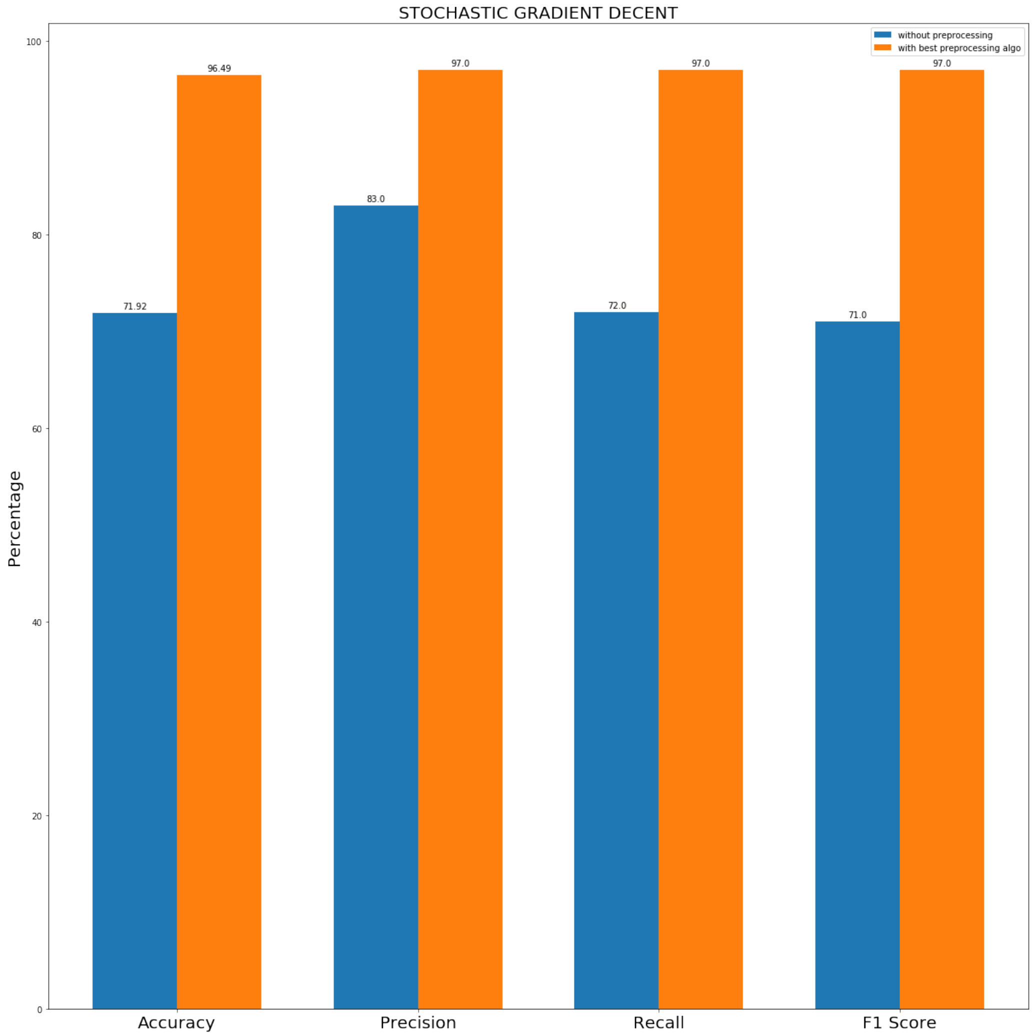

Figure 18: Performance metrics of Stochastic Gradient Descent

5.SUPPORT VECTOR MACHINE

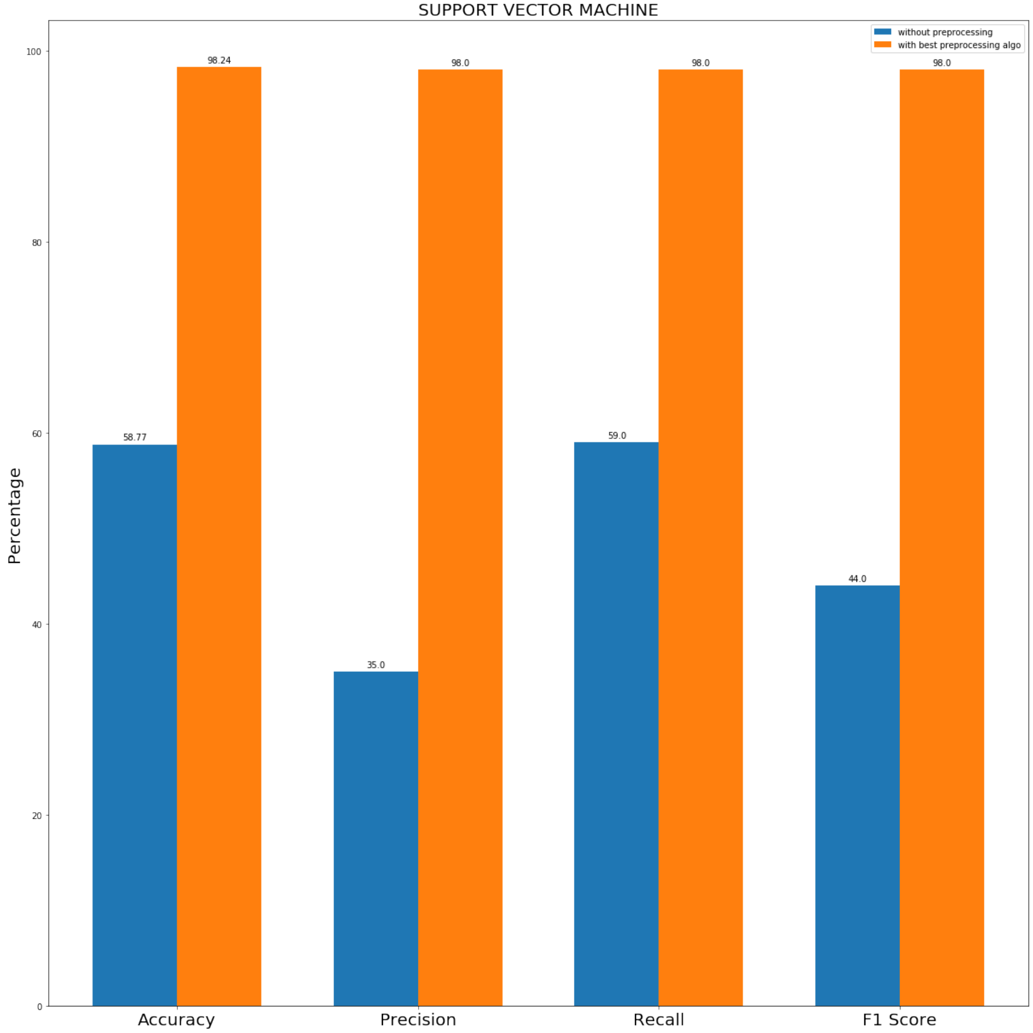

Figure 19: Performance metrics of Support Vector Machine

6.RANDOM FOREST

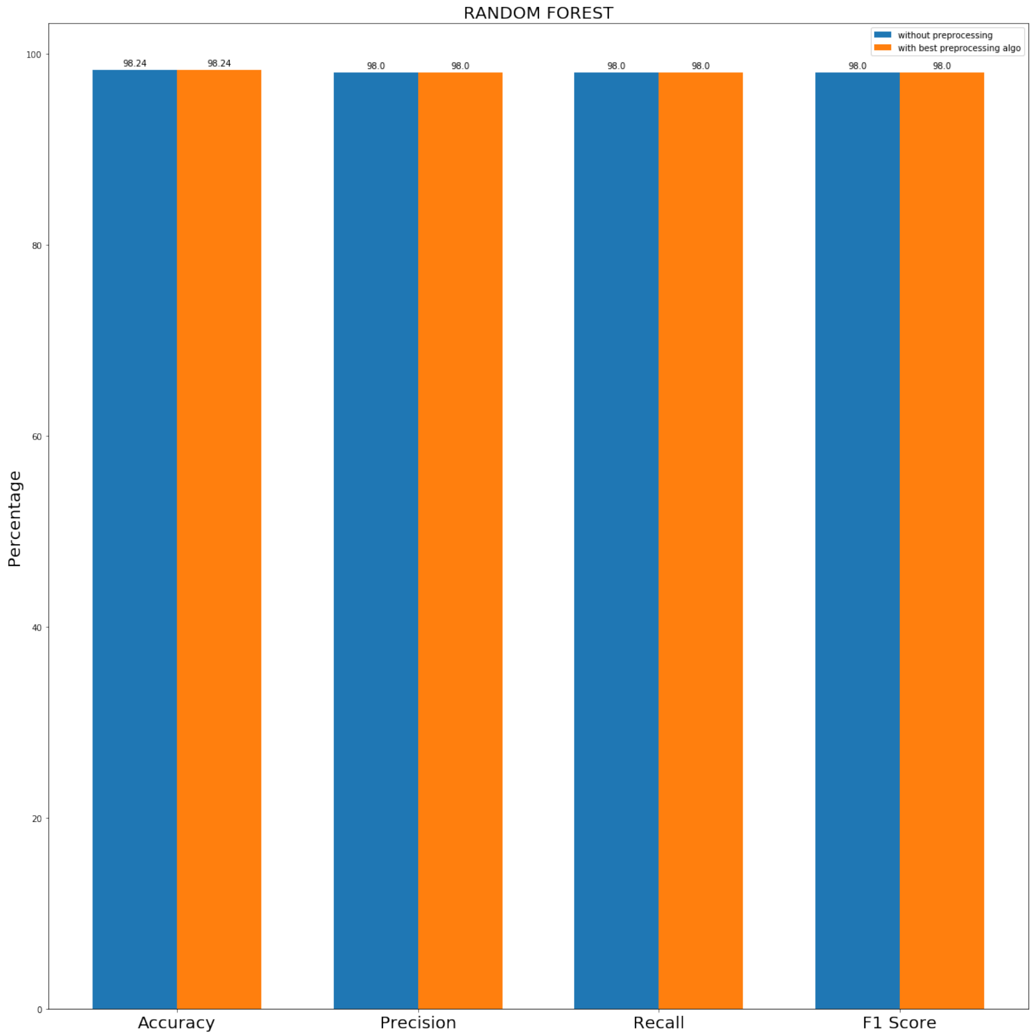

Figure 20: Performance metrics of Random Forest

7.ADABOOST

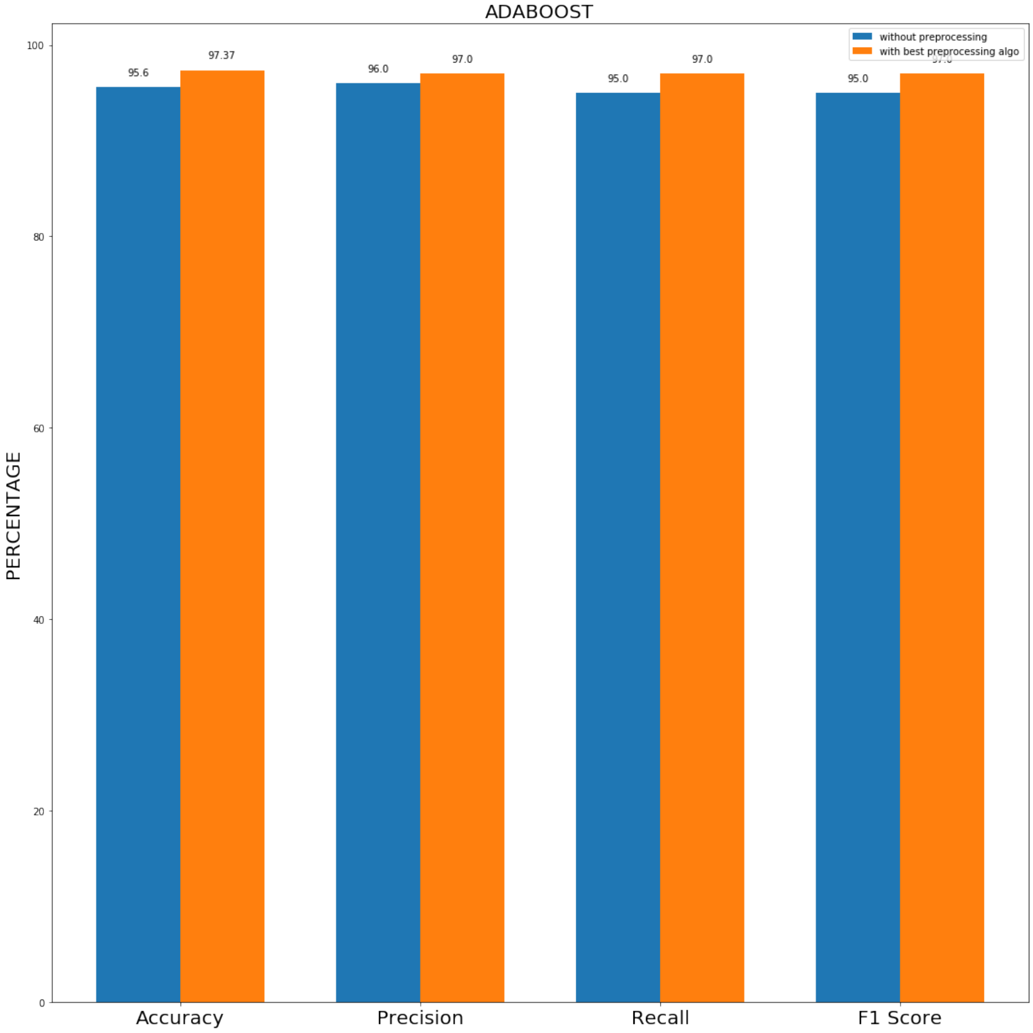

Figure 21: Performance metrics of Adaboost

8.k-NEAREST NEIGHBOURS

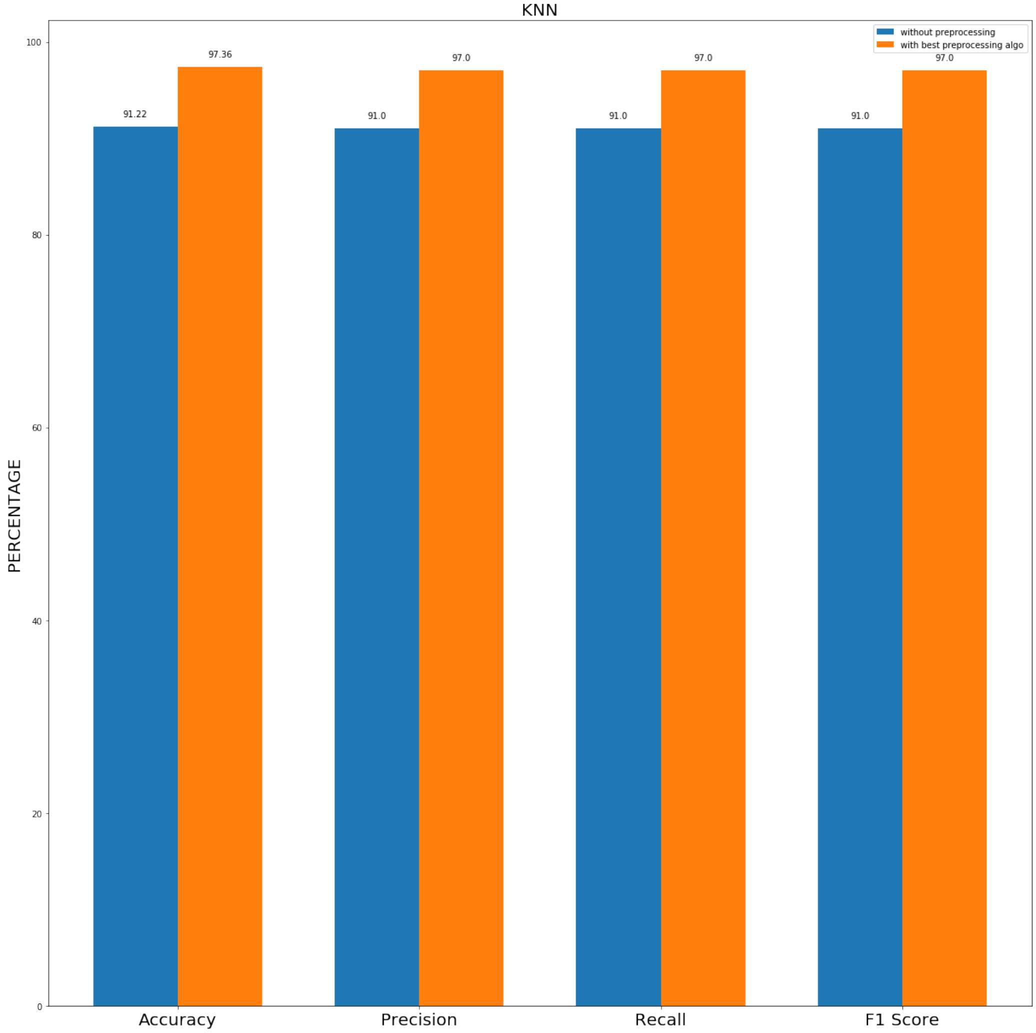

Figure 22: Performance metrics of k-nearest neighbours

9.NAIVE BAYES

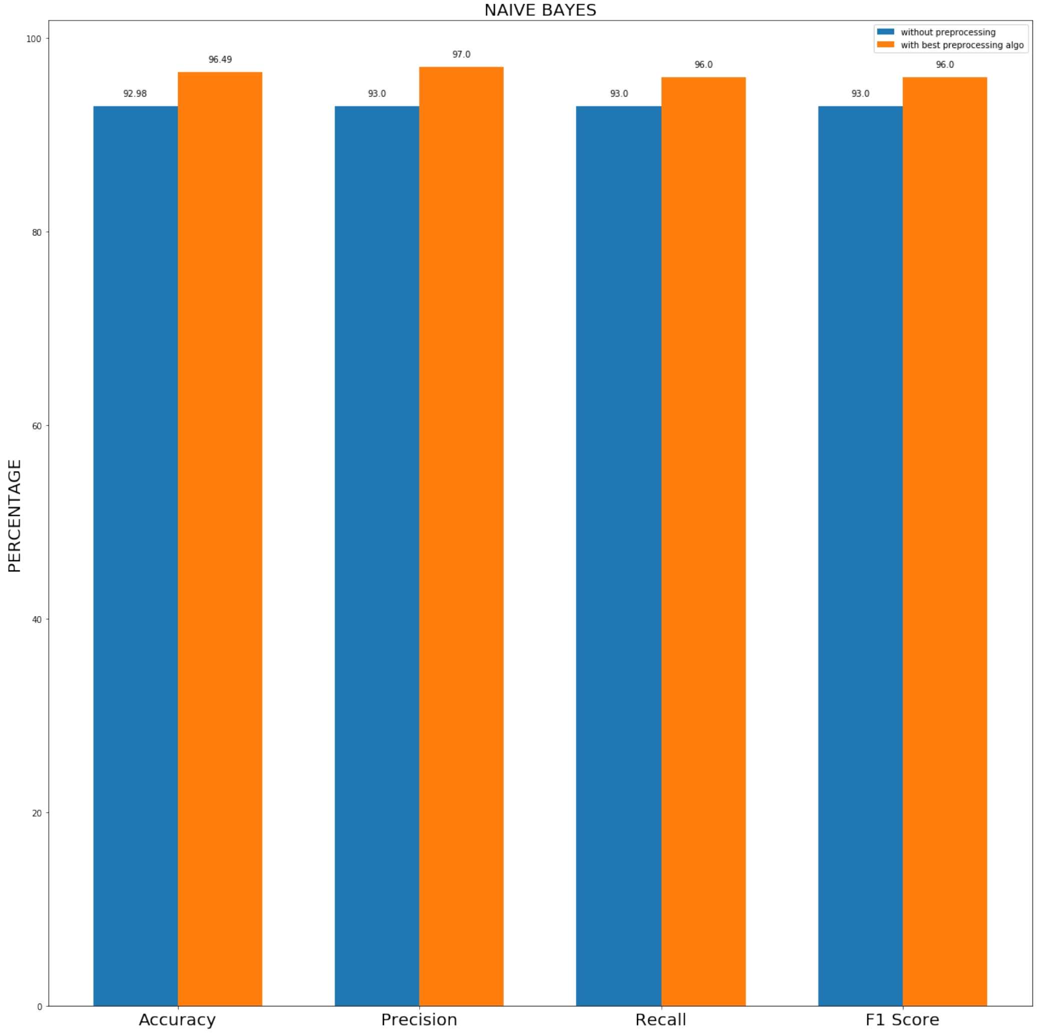

Figure 23: Performance metrics of Naive Bayes

Chapter 6

CONCLUSION AND FUTURE SCOPE

6.1 CONCLUSION

The proposed model in this paper presents a comparative study of different preprocessing algorithms and focuses on the importance of preprocessing the dataset. Using the Wisconsin Diagnosis Breast Cancer Dataset, performance comparison of various machine learning algorithms techniques with and without preprocessing techniques has been carried out. It has been observed that each of the algorithms had an accuracy of more than 95%,when the data was preprocessed. Thus preprocessing techniques will be very supportive in raising the accuracy in early diagnosis and prognosis of a cancer type in research.

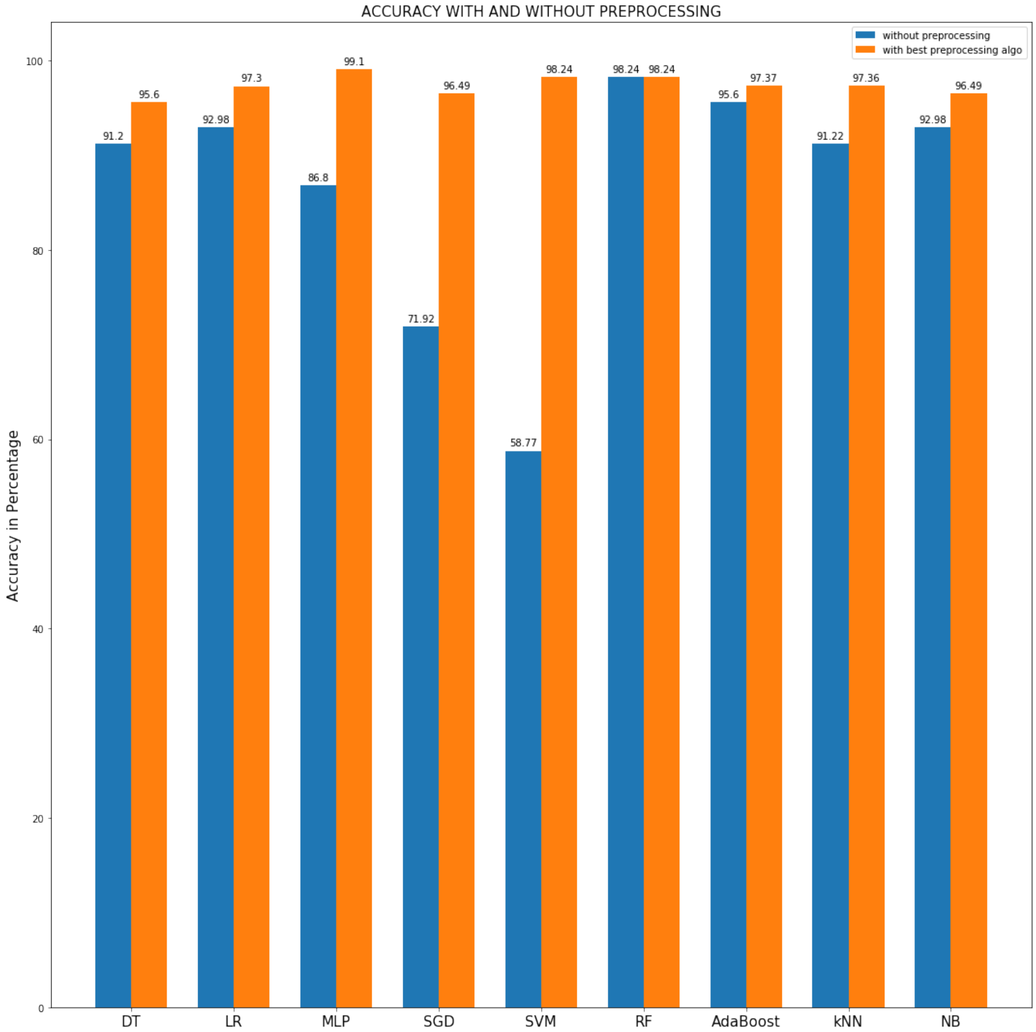

Figure 24: Comparative Analysis of classification algorithms in terms of Accuracy

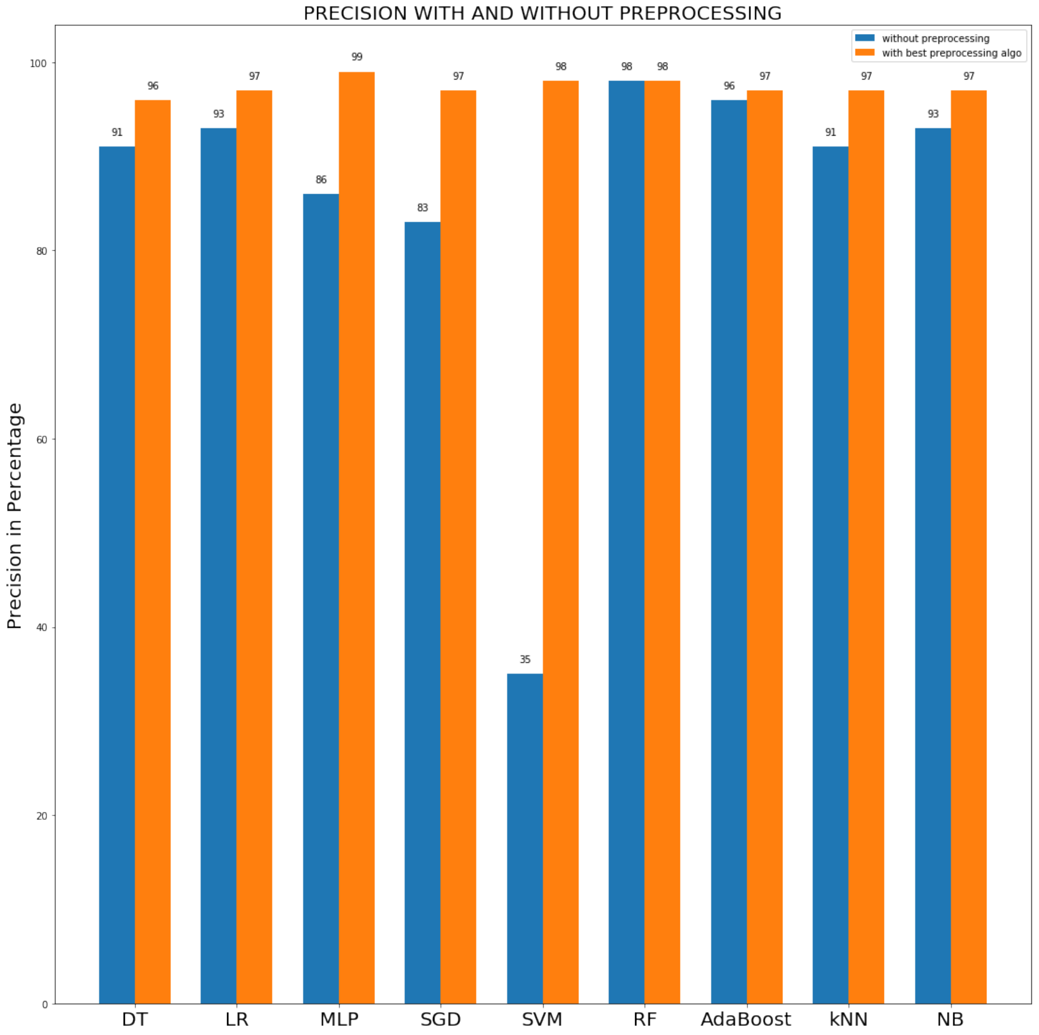

Figure 25: Comparative Analysis of classification algorithms in terms of Precision

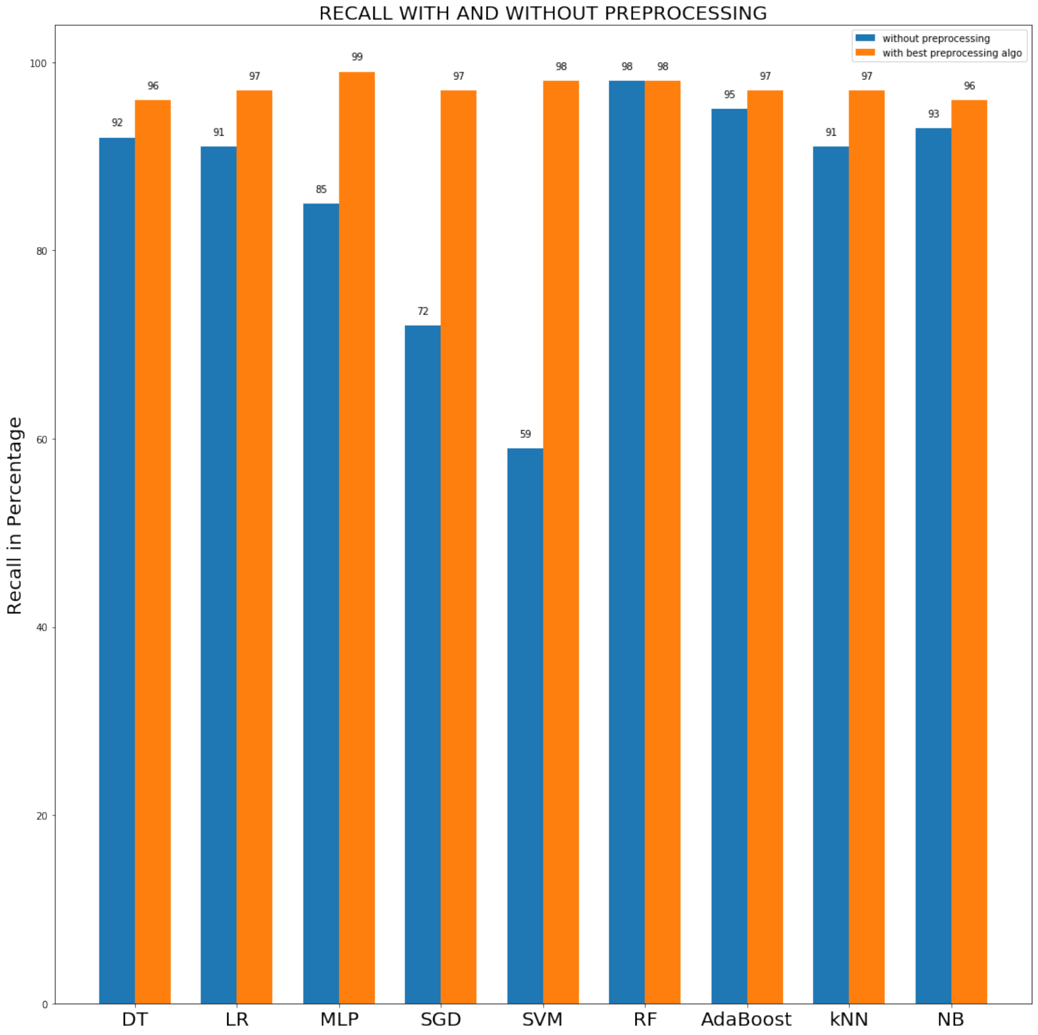

Figure 26: Comparative Analysis of classification algorithms in terms of Recall

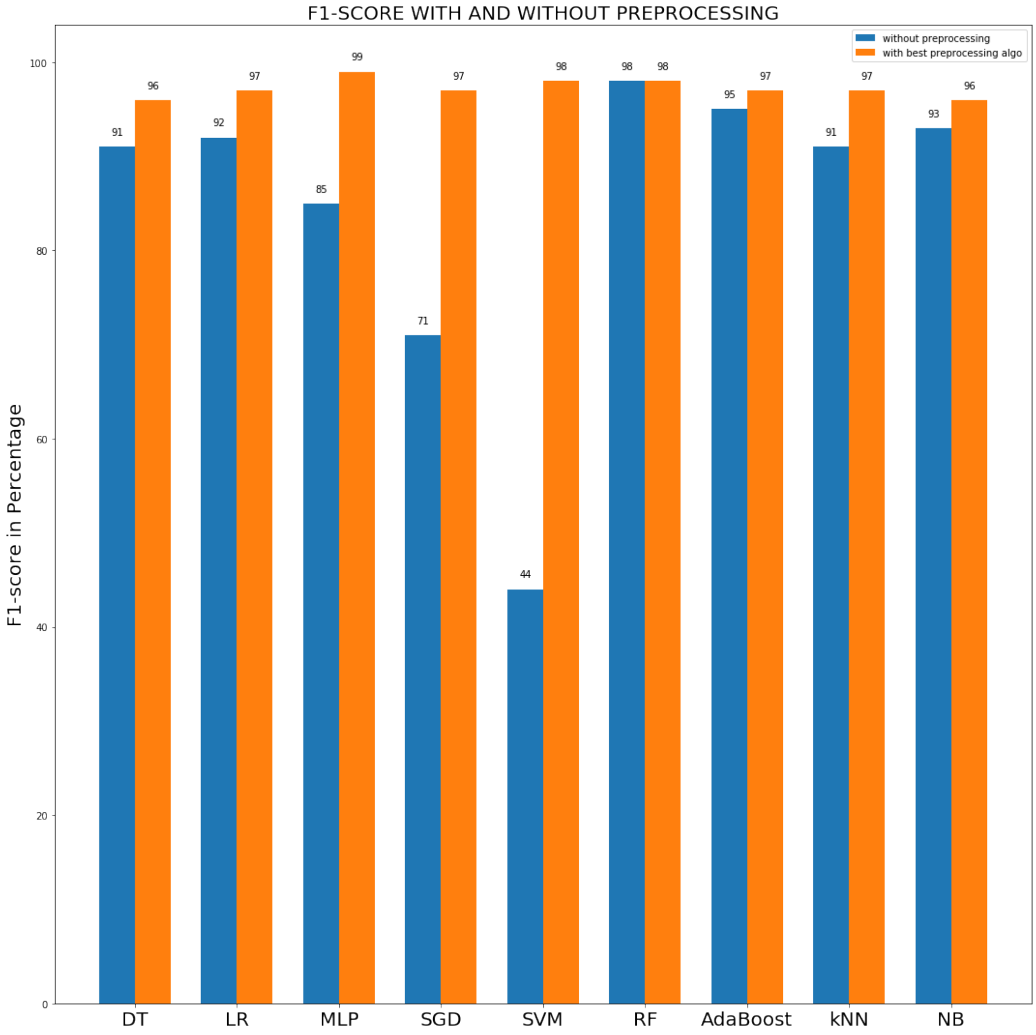

Figure 27: Comparative Analysis of classification algorithms in terms of F1-score

6.2 FUTURE SCOPE

With rise in the breast cancer cases,there is a need to detect the presence of the tumor as early as possible.For this reason,Machine learning algorithms can be of great benefit.The future scope of this project can be expanded for the same.The research can be expanded by including more Machine Learning Algorithms that are present currently or which may be introduced in the future.The research can also be expanded by including more preprocessing techniques that may be introduced in the future.Therefore,this topic has immense future scope.

REFERENCES

[1]E. Bayrak, P. Kırcı and T. Ensari, ”Comparison of Machine Learning Methods for Breast Cancer Diagnosis - IEEE Conference Publication”, Ieeexplore.ieee.org, 2019.

[2]S. Sharma, A. Aggarwal and T. Choudhury, ”Breast Cancer Detection Using Machine Learning Algorithms - IEEE Conference Publication”, Ieeexplore.ieee.org, 2019.

[3]M. Amrane, S. Oukid, I. Gagaoua and T. Ensarİ, ”Breast cancer classification using machine learning,” 2018 Electric Electronics, Computer Science, Biomedical Engineerings’ Meeting (EBBT), Istanbul, 2018, pp. 1-4. doi: 10.1109/EBBT.2018.8391453

[4]A. Bharat, N. Pooja and R. A. Reddy, ”Using Machine Learning algorithms for breast cancer risk prediction and diagnosis,”3rd International Conference on Circuits, Control, Communication and Computing (I4C), Bangalore, India, 2018, pp. 1-4.doi: 10.1109/CIMCA.2018.8739696

[5]Singh, S. Thakral, Shivani, ”Using Data Mining Tools for Breast Cancer Prediction and Analysis.”,2018,CCAA,1-4 10.1109/.2018.8777713.

[6]P. Mekha and N. Teeyasuksaet, ”Deep Learning Algorithms for Predicting Breast Cancer Based on Tumor Cells,” ,Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON), Nan, Thailand, 2019, pp. 343-346. doi: 10.1109/ECTI-NCON.2019.8692297

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

[8]H. AttyaLafta, N. KdhimAyoob and A. A. Hussein, ”Breast cancer diagnosis using genetic algorithm for training feed forward back propagation,” 2017 Annual Conference on New Trends in Information Communications Technology Applications (NTICT), Baghdad, 2017, pp. 144-149.

[9]Alickovic, Emina Subasi, Abdulhamit.”Breast cancer diagnosis using GA feature selection and Rotation Fores”, Neural Computing and Applications,2015. 10.1007/s00521-015-2103-9.

[10]Khuriwal, N. and Mishra, N.,”Breast Cancer Diagnosis Using Deep Learning Algorithm” - IEEE Conference Publication,2019.

[11]Agarap, Abien Fred,” On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset”. International Conference on Machine Learning and Soft Computing (ICMLSC) 2018, At Phu Quoc Island, Viet Nam 10.1145/3184066.3184080.

[12]Wolberg, W.H., Mangasarian, O.L. ”Multisurface method of pattern separation for medical diagnosis applied to breast cytology.”, In Proceedings of the National Academy of Sciences,1990, 87, 9193–9196.

[13]Zhang, J.”Selecting typical instances in instance-based learning.”, In Proceedings of the Ninth International Machine Learning Conference ,1992,(pp. 470–479).

[14]”Breast cancer”, World Health Organization, 2019.

[15]”WHO — WHO position paper on mammography screening”, Who.int, 2019.

[16]”Breast cancer - Diagnosis and treatment - Mayo Clinic”, Mayoclinic.org, 2019.

[17]Dua, D. and Graff, C.,UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science.

urkund

Document Information

Analyzed document	BreastCancer-blackbook-33-59-66.pdf (D75845051)
Submitted	6/30/2020 7:41:00 PM
Submitted by	Ujwala
Submitter email	ujwala.ravale@siesgst.ac.in
Similarity	8%
Analysis address	ujwala.ravale.sies@analysis.urkund.com

Sources included in the report

SA	URL: CBProjectReport.pdf Fetched: 11/4/2019 10:30:00 AM	⊞	**2**
SA	URL: report.docx Fetched: 5/7/2018 6:39:00 PM	⊞	**4**
W	URL: https://www.ijcseonline.org/pub_paper/37-IJCSE-05639.pdf Fetched: 10/8/2019 9:54:17 PM	⊞	**2**
W	URL: https://www.researchgate.net/publication/292225458_Breast_Cancer_Diagnosis_on_Thre ... Fetched: 11/25/2019 6:58:52 PM	⊞	**2**
SA	URL: DG_BC.docx Fetched: 10/8/2019 9:54:00 PM	⊞	**2**
SA	URL: USING PREDICTIVE ANALYSIS FOR BREAST CANCER DETECTION.docx Fetched: 1/5/2020 7:31:00 PM	⊞	**3**
SA	URL: journal main (4).docx Fetched: 6/8/2019 9:57:00 PM	⊞	**1**
W	URL: https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast- ... Fetched: 6/30/2020 7:43:00 PM	⊞	**2**
W	URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4452509/ Fetched: 6/30/2020 7:43:00 PM	⊞	**1**
W	URL: https://repository.aust.edu.ng/xmlui/bitstream/handle/123456789/4907/Ude%20Anthony ... Fetched: 4/5/2020 2:37:23 PM	⊞	**1**
W	URL: https://dergipark.org.tr/en/download/article-file/911317 Fetched: 4/17/2020 8:17:54 AM	⊞	**1**
W	URL: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original) Fetched: 6/30/2020 7:43:00 PM	⊞	**1**