International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 3, ISSN No. 2455-2143, Pages 431-438
Published Online July 2020 in IJEAST (http://www.ijeast.com)

A COMPARATIVE STUDY BETWEEN VARIOUS PREPROCESSING TECHNIQUES FOR MACHINE LEARNING

Shruti Rao, Prathiksha Poojary, Jigar Somaiya, Pranita Mahajan
Department of Computer
SIES Graduate School of Technology, Mumbai, India

***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 pre-processing. 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 apart from deficient preprocessing, excessive preprocessing also deteriorate the performance of an algorithm, which proves that preprocessing has a great influence on the performance of a classifier.**

Keywords—***Malignant, Benign, Preprocessing***

I. 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]

II. MACHINE LEARNING ALGORITHMS

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 allowing the computers to learn automatically without human intervention.

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 behaviors and punishing the undesired ones.

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 3, ISSN No. 2455-2143, Pages 431-438
Published Online July 2020 in IJEAST (http://www.ijeast.com)

A. K-Nearest Neighbors (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.

Accuracy of KNN algorithm without performing any preprocessing on the dataset is 91.22%.

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

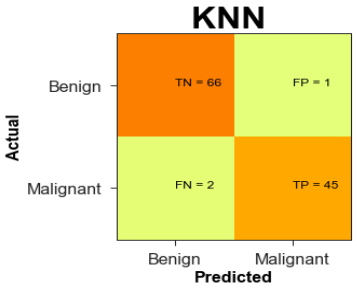

Fig. 1. Confusion Matrix of KNN

B. 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. The accuracy of naive Bayes without performing any preprocessing on Wisconsin dataset is 92.98 %. When the dataset was preprocessed using LDA, the best accuracy of 96.49% was obtained.

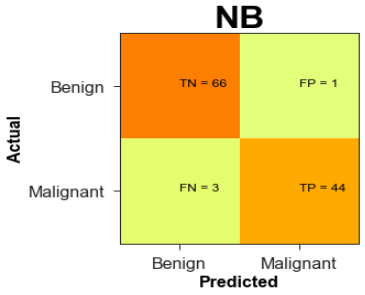

Fig. 2. Confusion Matrix of NB

C. 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. The accuracy of SVM without preprocessing on Wisconsin dataset is 58.77%. When the dataset was preprocessed using standardization, the best accuracy of 98.24% was obtained.

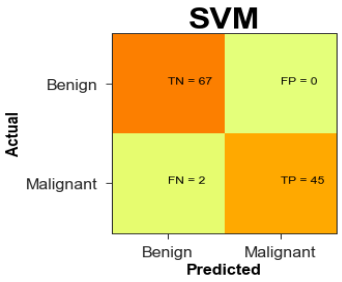

Fig. 3. Confusion Matrix of SVM

D. 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.

The accuracy of DT without any preprocessing on Wisconsin dataset is 91.2%. When the dataset was preprocessed using LDA, the best accuracy of 95.6% was obtained.

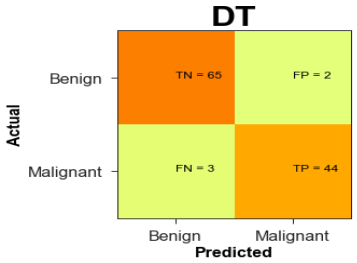

Fig. 4. Confusion Matrix of DT

E. 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. The accuracy of DT without any preprocessing on Wisconsin dataset is 98.24%. When the dataset was preprocessed using standardization or normalization, the best accuracy of 98.24% was obtained.

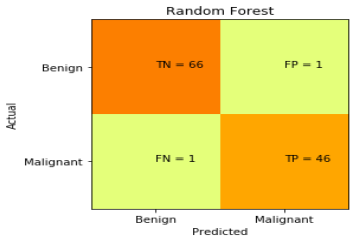

Fig. 5. Confusion Matrix of RF

F. 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

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 3, ISSN No. 2455-2143, Pages 431-438
Published Online July 2020 in IJEAST (http://www.ijeast.com)

important as classification problems depend on it. The value of recall and precision affect the decision for the value of threshold. The accuracy of LR without any preprocessing is 95.6%. When the dataset was preprocessed using LDA, the best accuracy of 97.3% was obtained.

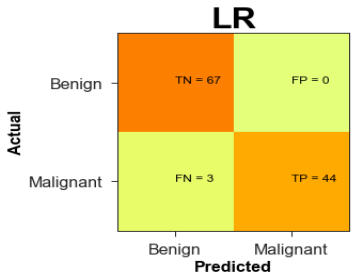

Fig. 6. Confusion Matrix of LR

G. 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. The accuracy of MLP without applying any preprocessing is 90.9% When the dataset was preprocessed using normalization, standardization and PCA, best accuracy of 99.1% was obtained

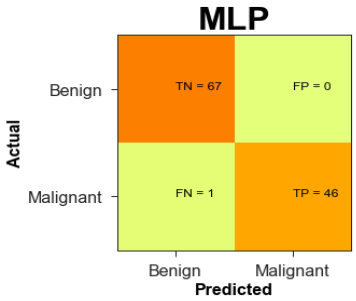

Fig. 7. Confusion Matrix of MLP

H. 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. The accuracy of SGD without any preprocessing is 82.4%. When the dataset was preprocessed using LDA, the best accuracy of 97.36% was obtained.

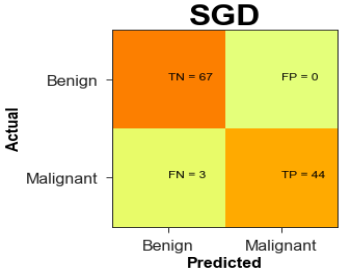

Fig. 8. Confusion Matrix of SGD

I. 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.

The accuracy of adaboost without any preprocessing is 95.61%. When the dataset was preprocessed using PCA with standardization and normalization, the best accuracy of 97.36% was obtained.

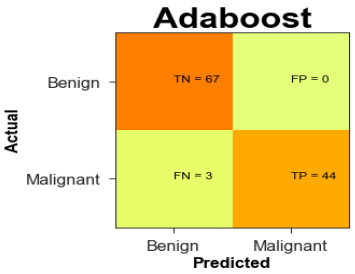

Fig. 9. Confusion Matrix of Adaboost

III. DATA PREPROCESSING

Pre-processing is the transformation that we apply on the data before feeding it to the algorithm. Data preprocessing 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. In this paper we have used the following preprocessing techniques:

A. Label Encoder -

This usually deals with datasets which contain multiple labels in one or more than one column, they can be in numbers or words but to make it human readable they are usually in words. Label encoding refers to the process of converting labels into numeric form to make the machine readable.

Table. 1. Encoding of benign and malignant

Diagnosis	Diagnosis
Malignant	0
Benign	1

B. Normalization -

An attribute in a dataset may contain values with varying scale so ML algorithms always benefit 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 a minmax scaler for this. Formula for min max scaler:

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 3, ISSN No. 2455-2143, Pages 431-438
Published Online July 2020 in IJEAST (http://www.ijeast.com)

$$X - \min / (\max - \min) \quad (1)$$

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. we generally use StandardAero for this and formula for same is:

$$Z = x - \text{mean} / \text{standard deviation} \quad (2)$$

D. Linear Discriminant Analysis (LDA) -

It is the most commonly used dimension reductionality technique in machine learning applications. It decomposes the dataset into lower dimensions, in addition to that it reduces overfitting of data and reduces the computational cost. The approach of LDA is similar to PCA but in addition to finding component axes that maximize the variance, we find axes that maximize 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 the square matrix into eigenvectors and eigenvalues.
4. Selecting linear discriminates for new feature subspace: Sorting of eigenvectors by decreasing order of eigenvalues. Choosing K eigen- vectors with largest eigenvalues.
5. Transforming samples onto the new subspace: Mapping the samples to the new feature subspace.

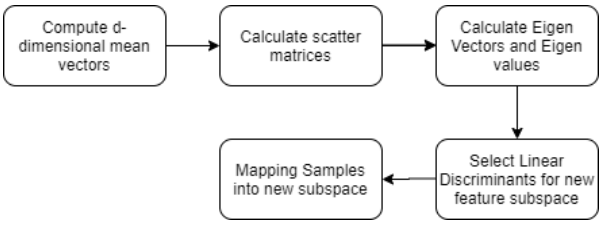

Fig. 10. Steps for LDA

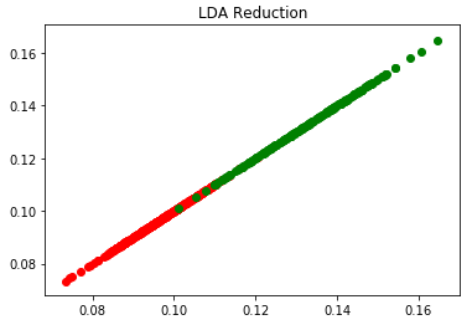

Fig. 11. Separation of classes using LDA

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.
Calculate Covariance: Find covariance matrix for the given data. Covariance is a measure of how two variables move together.
3. Deduce Eigenvectors: Our main principal component becomes the 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 eigenvectors as eigenvectors indicate the directions of new axes.
4. Reorient data: To reorient the data according to new axes we multiply our data with eigenvectors. This reoriented data is scored.
5. Plot the Reoriented data or score.

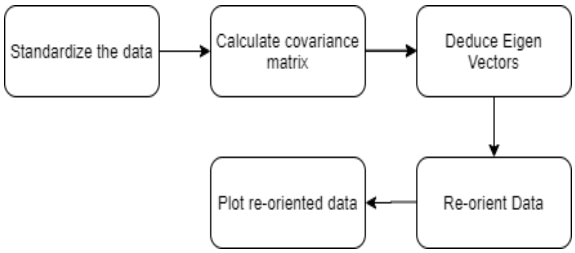

Fig. 12. Steps for PCA

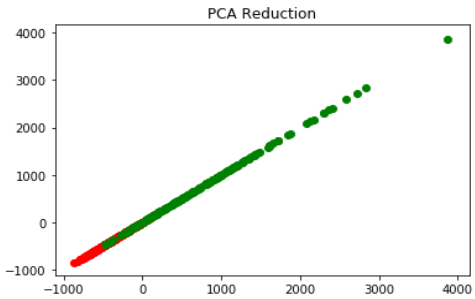

Fig. 13. Separation of classes using PCA

IV. PROPOSED METHODOLOGY

A. 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. There are no missing values. There are 2 classes Malignant and Benign. 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.

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 3, ISSN No. 2455-2143, Pages 431-438
Published Online July 2020 in IJEAST (http://www.ijeast.com)

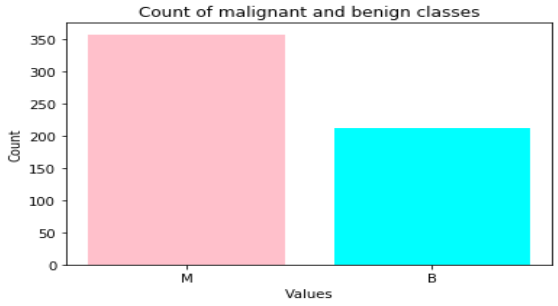

Fig. 14. Count of Benign and Malignant

B. 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:

1. Accuracy-

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

The formula is as follows:

$$\text{Accuracy}=(TP+TN) / (TP+FP+TN+FN) \quad (3)$$

2. Precision-

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

The formula is as follows:

$$\text{Precision} = TP/(TP+FP) \quad (4)$$

where TP is true positive and FP is false negative

3. Recall-

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

The formula is as follows:

$$\text{Recall} = TP/(TP+FN) \quad (5)$$

where TP is true positive and FN is false negative

4. 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 as follows:

$$\text{F1 Score}=2*(P*R)/(P+R) \quad (6)$$

where P is precision and R is recall

C. Implementation and Result Analysis -

A comparative study between various Machine learning algorithms, with and without preprocessing has been proposed. It has 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.

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:

Algorithm without any preprocessing
Algorithm with normalization of the data
Algorithm with standardization of the data
Algorithm with normalization and Standardization of data
Algorithm with PCA:
PCA (components= 8,9,15)
PCA with normalization
PCA with standardization
PCA with normalization & standardization

Algorithm with LDA:
LDA
LDA with normalization
LDA with standardization
LDA with normalization and standardization

Then we selected the preprocessing technique which gave the best results for each individual classifier. The preprocessing techniques that gave the best results for the classifiers are as follows:

Table. 2. Best preprocessing techniques for each algorithm

Sr.no	*Algorithm*	*Preprocessing technique having the best performance*
1.	DT	• LDA • PCA with normalization & standardization
2.	LR	• LDA • PCA with normalization
3.	MLP	• PCA with normalization
4.	SGD	• PCA with normalization & standardization • LDA
5.	SVM	• Standardization

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 3, ISSN No. 2455-2143, Pages 431-438
Published Online July 2020 in IJEAST (http://www.ijeast.com)

6.	RF	• Standardization • Normalization
7.	AdaBoost	• Standardization • PCA with norm and std
8.	kNN	• LDA
9.	NB	• LDA

The comparison of the selected best preprocessing algorithm for each classifier with the classifier without any preprocessing is as follows:

Table.3. Performance comparison with best and without preprocessing.

Sr.no	*Algorithm*	*Performance of classifier with the best preprocessing technique*				*Performance of the classifier without Preprocessing*			
		A	*P*	*R*	*F1*	*A*	*P*	*R*	*F1*
1.	DT	95.6	96	96	96	91.2	91	92	91
2.	LR	97.3	97	97	97	92.9	93	91	92
3.	MLP	99.1	99	99	99	86.8	86	85	85
4.	SGD	96.5	97	97	97	71.9	83	72	71
5.	SVM	98.2	98	98	98	58.7	35	59	44
6.	RF	98.2	98	98	98	98.2	98	98	98
7.	AdaBoost	97.4	97	97	97	95.6	96	95	95
8.	kNN	97.4	97	97	97	91.2	91	91	91
9.	NB	96.5	97	96	96	92.9	93	93	93

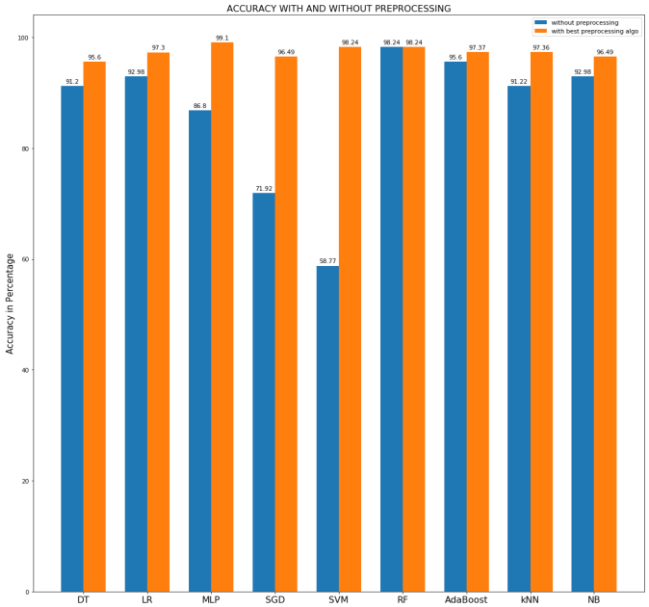

Fig. 15. Accuracy with and without preprocessing

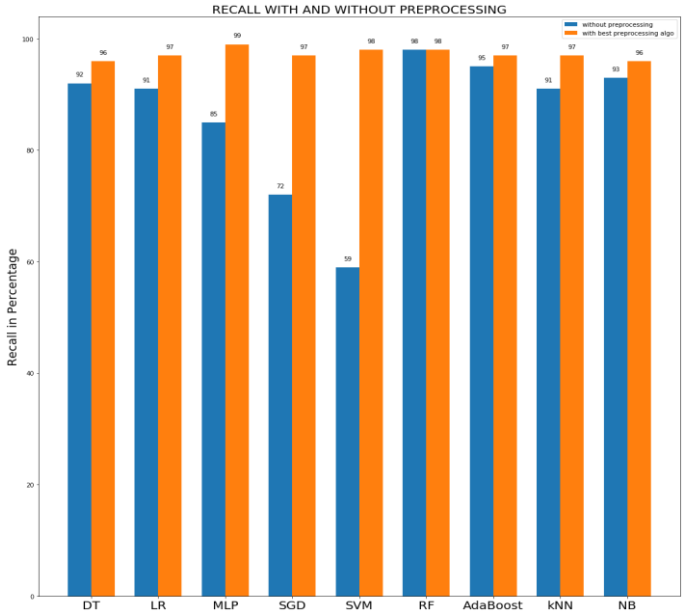

Fig. 16. Recall with and without preprocessing

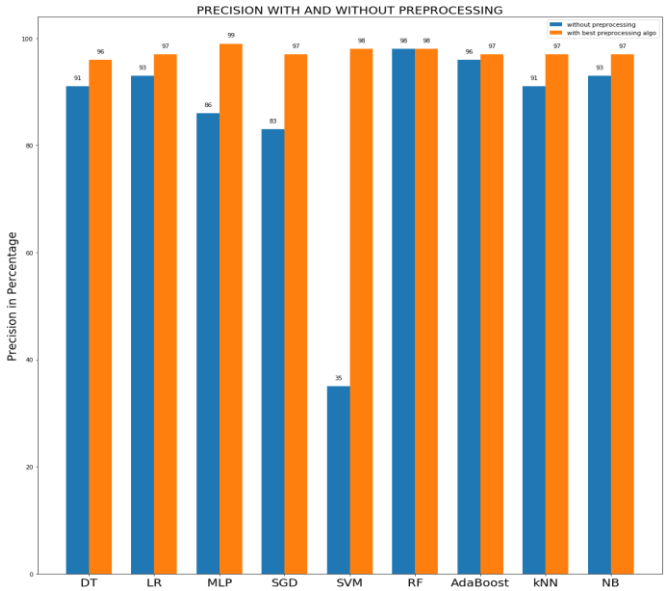

Fig. 17. Precision with and without preprocessing

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 3, ISSN No. 2455-2143, Pages 431-438
Published Online July 2020 in IJEAST (http://www.ijeast.com)

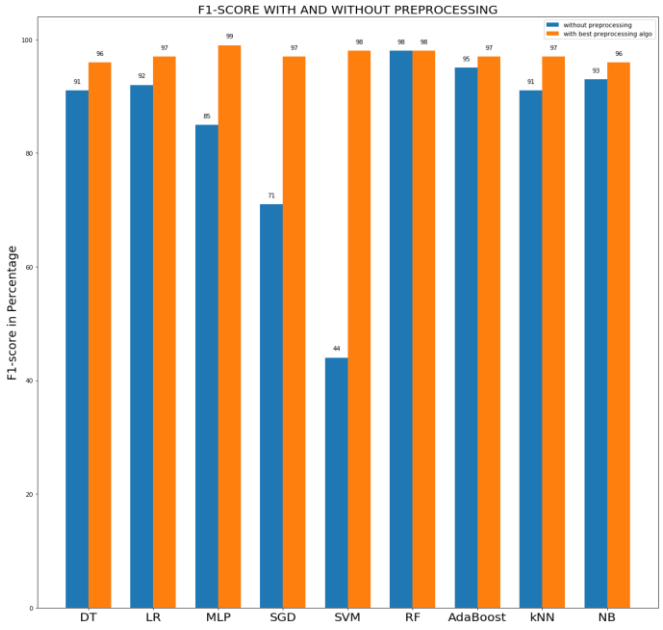

Fig. 18. F1 Score with and without preprocessing

D. **Importance of ideal Preprocessing -**

Data preprocessing is important for any data mining task as the rate of success depends on it. . Preprocessing reduces the complexity of the data. A detailed comparison of all the above-mentioned algorithms were done. They were compared on the basis of accuracy, precision, recall and F1 score. It was observed that the algorithm's performance was poor when the data was not preprocessed or when scanty preprocessing was done. Then as preprocessing steps were performed one by one, the performance started improving. It was also observed that when data was overly preprocessed, the algorithm's performance deteriorated.

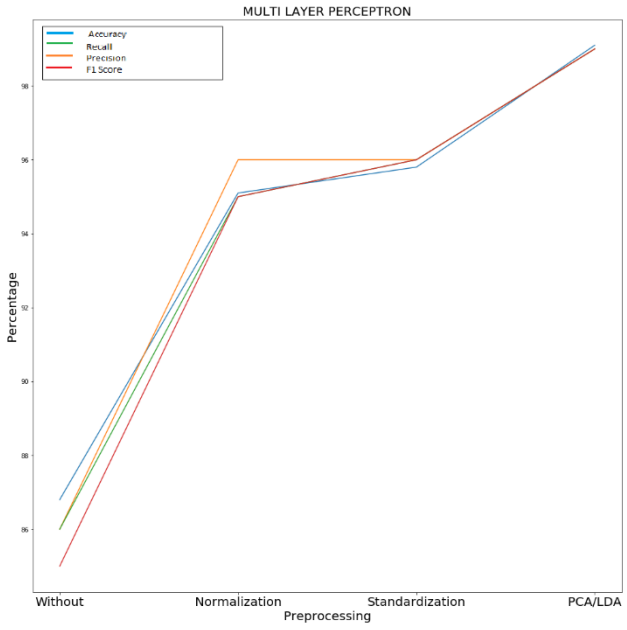

Fig. 19. Preprocessing graph for multilayer perceptron

Here, it can be seen that as we preprocess the data the performance increases and reaches the maximum when PC/LDA along with standardization and normalization is performed.

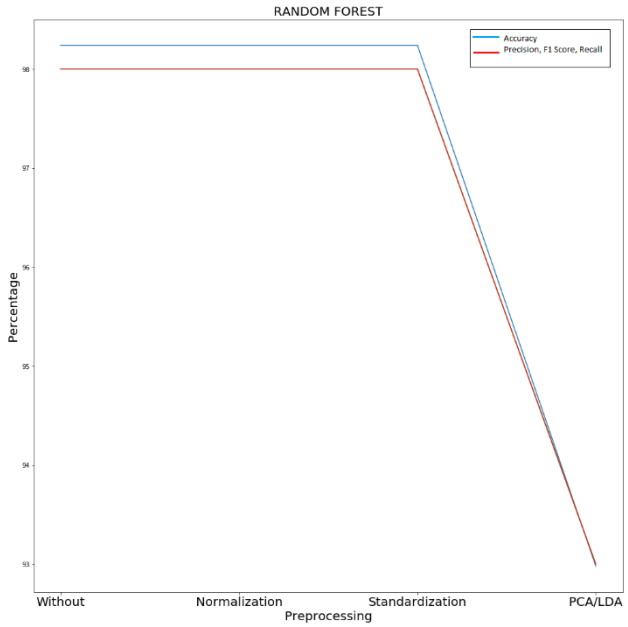

Fig. 20. Preprocessing graph for random forest

Here, it can be observed that as we preprocess the data further after standardization, the performance of the algorithm decreases. This is the effect of over preprocessing.

V. 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 ideally preprocessed. When the data was partially preprocessed the performance stooped down and similar is the case when the data was over preprocessed. Thus, accurate amount of preprocessing is necessary as even excessive preprocessing can hamper the performance of the classifiers as it rounds off or ignores the minute details in the data. Hence, proper preprocessing techniques will be very supportive in raising the accuracy in early diagnosis and prognosis of a cancer type in research

VI. REFERENCE

[1] Bayrak E. , Kırcı P and Ensari T. (2019). "Comparison of Machine LearningMethods for Breast Cancer Diagnosis - IEEE Conference Publication", Iee-explore.ieee.org,.

[2] Sharma S., Aggarwal S. and Choudhury T., 2019 "Breast Cancer DetectionUsing Machine Learning

Algorithms - IEEE Conference Publication", Ieeex-plore.ieee.org,.

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

[4] Bharat A., Pooja N. and Reddy R.A., (2018) "Using Machine Learning algo-rithms for breast cancer risk prediction and diagnosis,"3rd International Con-ference on Circuits, Control, Communication and Computing (I4C), Banga-lore, India,, pp. 1-4.doi: 10.1109/CIMCA.2018.8739696

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

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

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

[8] AttyaLafta H., KdhimAyoob N. and Hussein A. A., (2017),"Breast cancer diag-nosis using genetic algorithm for training feed forward back propagation," Annual Conference on New Trends in Information CommunicationsTechnology Applications (NTICT), Baghdad, 2017, pp. 144-149.16

[9] Alickovic, Subasi Emina, Abdulhamit.",(2015),Breast cancer diagnosis using GAfeature selection and Rotation Fores", Neural Computing and Applications,.10.1007/s00521-015-2103-9.

[10] Khuriwal, N. and Mishra, N., (2019),"Breast Cancer Diagnosis Using Deep Learn-ing Algorithm" - IEEE Conference Publication.

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

[12] 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.

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

[14] (2019)"Breast cancer", World Health Organization, .

[15] (2019)"WHO — WHO position paper on mammography screening", Who.int,.

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

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