Prediction of Diabetes in Females of PIMA Indian Heritage using Machine Learning classification Algorithms

Kritin Pandita
20387234
School of Computer Science
University of Nottingham
Nottingham, United Kingdom
psxkp3@exmail.nottingham.ac.uk

Navya Singh Vasudeva Singh
20351524
School of Computer Science
University of Nottingham
Nottingham, United Kingdom
psxnv1@exmail.nottingham.ac.uk

***Abstract*— Nowadays diabetes is a serious chronic disease that affects millions of people in the world, especially women. According to a lot of factors and variables, the diagnosis of this disease is very complex and there is a lot of potential for possible human error. So, there is a dire need for a tool or technological advancement for prognosis of this disease that can help doctors in early detection of this fatal disease and hence can recommend certain lifestyle changes to stop the progression of diabetes. Recent healthcare research and studies have made a virtuous use of various cutting-edge techniques and advanced scientific methods to accurately diagnose and predict the disease based on the data received from the person. One such technological advancement in this field is Machine Learning which helps us in making smarter and accurate predictions on the data. In this research paper, we have designed a classification model which helps us in the prediction of diabetes in females of Pima Indian Heritage. Since this is a binary classification problem, we have used two popular supervised learning algorithms such as the Naive Bayes Classification and Logistic Regression. The data has been taken from the UCI repository and Kaggle. We have also computed the AUC value and classification accuracies in the case of both types of supervised learning algorithms that would give us a fair idea about which classification algorithm is better and more accurate. We have used R studio and its various libraries to generate results.**

Keywords - Machine Learning, Supervised Learning, Missing Data, Imputation, Classification

I. INTRODUCTION

Diabetes is a condition in which blood glucose, often known as blood sugar, is excessively high. The main source of energy is blood glucose, which comes from the food we eat. Insulin, a hormone produced by the pancreas, aids glucose absorption into our cells for energy usage. Sometimes our body doesn't make enough — or any — insulin or doesn't use insulin well. Glucose remains in the bloodstream and does not reach the cells. Glucose is required for all cells to function. However, if blood glucose levels rise due to a shortage of insulin hormone in the body, blood glucose levels become unbalanced, causing significant damage to other organs such as the eyes, heart, and kidneys, among others. It is stated that there is no permanent treatment for diabetes, but it can be managed by maintaining a healthy lifestyle that includes exercise and a well-balanced diet. Type 1, type 2, and gestational diabetes are the three forms of diabetes. In type-1, the immune system destroys the insulin cells. It generally happens to children and adolescents. In type-2, the pancreas makes very little insulin. It generally happens to adults. Insulin resistance refers to the first type, whilst insulin insufficiency refers to the second. Various technologies have been used in recent healthcare studies to diagnose people and forecast their disease based on clinical data. ML approaches are now being used in the healthcare sector to better forecast diabetes. The major goal of this study is to diagnose whether or not a patient has diabetes using diagnostic parameters from the National Institute of Diabetes and Digestive and Kidney Diseases dataset. Supervised and unsupervised learning are the two primary categories of Machine Learning techniques. We considered this task to be a binary classification problem because we will only be predicting if a female has diabetes or not. We employed supervised learning techniques such as Naive Bayes (NB) and Logistic Regression (LR) and compared the results.

The dataset is taken from UCI Machine Learning — Repository: https://www.kaggle.com/uciml/pima-indians-diabetes-database [7].

Several medical parameters are included in the dataset, with one dependent parameter having a binary value (0 and 1).

The dataset has 768 rows with 8 variables and a target variable output and is primarily for females.

The dataset is described in the table below

TABLE I. SUMMARISING THE PARAMETERS OF PIMA INDIAN DIABETES DATASET

Pregnancies	Number of times pregnant
Glucose Plasma	Glucose concentration 2 hours in an oral glucose tolerance test
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body mass index (weight(kg)/[height(m)]^2)
DiabetesPedigreeFunction	Diabetes pedigree function, provided some data on diabetes mellitus history in relatives and the genetic relationship of those relatives to the patient
Age	Age (years)
Outcome	target variable, whether patient had diabetes, out of 768 - 268 are Positive, 500 are Negative (1 = positive; 0 = negative)

Research Questions

1. What are the factors causing Diabetes?
2. Does number of times a female getting pregnant cause diabetes?
3. Does Blood pressure really have impact on causing diabetes?
4. How BMI is related to diabetes?

II. LITERATURE REVIEW

The analysis of related work gives results on various healthcare datasets, where analysis and predictions were carried out using various algorithms and techniques. Various prediction models have been developed and implemented by various researchers using variants of data mining techniques, machine learning algorithms or also combination of these techniques.

Aishwarya and Vaidehi[1] used various classification algorithms including SVM, Tree and Random Forest classifier, Perceptron, ADA-BOOST algorithm, Linear-Discriminant analysis, Logistic Regression, K-NN, Gaussian Naive Bayes algorithm, Bagging Algorithm and Gradient Boost classifier. They conducted these classification techniques on PIMA Indians and another diabetes dataset for testing the various modelling techniques after compiling the results, they found that logistic regression gave them the maximum accuracy value of 96%.

On the other hand, Tejas and Pramila[2] chose two classification algorithms -Logistic Regression and Support Vector Machine algorithms to build a diabetes prediction model. The pre-processing of the dataset was carried out before hand to get better and accurate results. They found that SVM was a better classification model with an accuracy of 79%.

Olaniyi and Adnan[3] and G. Swapna, K.P. Soman, R. Vinayakumar[4] made use of Deep learning techniques for diabetes prediction. Olaniyi and Adnan made use of Multilayer feed-forward neural network and used a back-propagation algorithm for training the model. They also used the PIMA Indian dataset and normalised it before performing pre-processing techniques on it to increase accuracy of the model. They obtained a classification accuracy of 82%. The latter used a dataset called Electrocardiograms on two models using CNN and CNN-LSTM. The dataset consisted of 142000 Samples and 8 attributes. They obtained an accuracy of 93.6 for the CNN model and an accuracy of 95.1% for the CNN-LSTM model with a five-fold cross validation for both.

Deepti and Dilip [5] used Decision Tree, SVM, and Naive Bayes algorithms. Ten-fold cross validation was used to improve performance. The highest accuracy was obtained by the Naive Bayes, with an accuracy of 76.30%. Both these papers used the Pima Indian Diabetes dataset.

Yuvaraj and Sripreethaa [6] designed a diabetes prediction model using three different Machine Learning algorithms-Random Forest, Decision Tree, and the Naïve Bayes, in Hadoop based clusters. They employed pre-processing techniques on the dataset. The results showed that the highest accuracy rate of 94% was obtained with the Random Forest algorithm.

III. METHODOLOGY

This section focuses on the methods and methodology that was used to predict diabetes in the female population of the Pima Indian heritage. This is a binary classification problem as we have already established earlier since the patient can have diabetes (1) or can't have it (0). This section vividly describes the main dataset and its attributes, inter collinear relationships between various attributes and how we established those and mainly the classification algorithms that we have used to predict diabetes and perform binary classification.

The dataset originally stems from the National Institute of Diabetes and Digestive and Kidney diseases (NIDDK) but is also stored in Kaggle and the UCI data repository where it was more accessibly taken from. The dataset mainly used to predict whether Pima Indian Females suffer from diabetes or not. The ages of the PIMA females in the dataset range from 21 to 81 and the average age is computed to be 33 years.

The following are the main attributes that are present in the dataset that tell us about the present conditional metrics of the PIMA females:

1. AGE: It shows the age of the females in years. The minimum age is seen to be 21 and the maximum is 81. The average age is 33.
2. Pregnancies: It shows how many times that particular female gets pregnant and it ranges from 0 to 17. the average is calculated to be 4.
3. Glucose: Tells us about the plasma glucose concentration levels for 2 hours in an oral glucose tolerance test. Ranges from 0 to 199 and the average is 121.
4. Blood Pressure: It shows the diastolic blood pressure in mm hg. It ranges from 0 to 122 and the average is 69.
5. Skin Thickness: It shows the triceps skin thickness in MM. The range is 0 to 99 and the average is 21.
6. Insulin: It ranges from 0 to 846.the average is 80.
7. BMI: It shows us the body-mass index in kg/m2. The range is calculated to be in between 0 to 67.1 and the average is 32.
8. Diabetes Pedigree Function: It tells us about the likelihood of the particular females having diabetes. The range lies between 0.078 and 2.42 and the average is 0.47.
9. Outcome: Tells us if the female is diabetic (1) or non-diabetic (0).

Outcome is defined as the categorical target variable in this dataset. The dimensions of the dataset come out to be 768 (No of rows-instances) x 9 (No of columns-attributes) (500 non-diabetics ,268 diabetics).

A. Data Analysis

Data Analysis is first step that needs to be done on the given dataset to get the insight of the data and to know how the parameters are affecting the Outcome. We can also extract the information about each parameter.

We have done the Data Analysis mainly by using the Box plot and Bar plot

1. Box plot – In the field of statistics and data analysis, the box plot is a very popular and commonly used plot for visualising data. In comparison to other graphical techniques, the Box Plot not only depicts the data distribution/spread, but also the lowest and maximum values, quartiles, symmetry, and skewness. Outliers are also detected using the box plot.
   As there are no NA values in this dataset using Box plot will provide us more information on outliers.

2. Histogram - Histograms are graphs that show how data is distributed. They are excellent exploratory tools since they disclose aspects about the data that summary statistics cannot.
   The purpose of using histogram is to show the form (distribution) of data for a single quantitative variable like Blood pressure, age, or Glucose.

B. Data Pre-Processing

Data pre-processing is a crucial step in the data modelling process. Here the dataset does not have any N/A values per se but there are major inconsistencies values within the data. There are a lot of values for attributes such as glucose concentration, insulin, blood pressure, body mass index and skin fold thickness that have zero values and are not within the accurate normal ranges.
To deal with the possibility of differences caused by outliers and inconsistent values, we need to apply pre-processing techniques and impute consistent values in the dataset. Since the dataset is not very big it would make little sense to delete instances of useful data unnecessarily. The data is also varied and extreme so using mean, median, mode techniques would also not be a favourable move. Hence, we have two imputation pre-processing algorithms that we have used for fitting missing values in the dataset.

1. KNN Imputation: It is one of the simplest and easiest techniques in machine learning that works on the principle of Euclidean distance between the neighbour coordinates (x, y) to know how similar the data is.
2. MICE Imputation: It is a much more robust and informative method of dealing with missing data that involves an iterative series of predictive models. During each iteration, each specified variable in a dataset is imputed using the other variables in the dataset. The iterations are supposed to run until it appears that convergence has met.

In our project, we have two instances of the same dataset. in one case we have used KNN imputation and then scaled down the data for Logistic Regression classification. In the other instance we have used mice imputation and done Naive bayes classification. Since Naive bayes models are not sensitive to normalisation and scaling, we have not scaled the dataset in this instance.

C. Classification

Algorithms for predictions of the diabetes:

1. Naïve Bayes Classification: As the name would suggest, is based on the bayes theorem and used for solving classification problems. It is a probabilistic classifier, which means it predicts on the basis of the probability of the object. It is a conditional probability model which considers each feature to contribute separately regardless of any correlations between them. The main advantage of this algorithm is that it is usually pretty fast and it requires a small dataset for training to start its classification.
2. Logistic Regression: it is a classification technique used to analyse a dataset in which there are one or more categorical variables that determine an outcome. The basic intention is to find the best fitting model to describe the relationship between the independent and dependent variables. It predicts the probability of an event by using the data in a logistic function.

IV. RESULTS FROM EACH OF THE STAGES

The outcomes of each stage, such as data pre-processing, data analysis, and classification methods, have been described and appraised in this section. This section seeks to present the outcomes of various methodologies used on the dataset at each level.

A. Data Analysis

In the Analysis, we check structure and statistical summary of the data.
The figure below shows the statistical summary of the dataset

FEATURES	PREG	GLUC	BP	SKIN	INSULIN	BMI	DPF	AGE
MIN	0.000	0.0	0.00	0.00	0.00	0.00	0.078	21
1ST QU.	1.000	99.0	62.00	0.00	0.00	27.30	0.2437	24
MEDIAN	3.000	117.0	72.00	23.00	30.5	32.00	0.3725	29
MEAN	3.845	120.9	69.11	20.54	79.8	31.99	0.4719	33.24
3RD QU	6.000	140.2	80.00	32.00	127.2	36.60	0.6262	41
MAX	17.000	199.0	122.00	99.00	846.0	67.10	2.4200	81

Fig. 1. Image of table displaying the statistical summary of data (without outlier removal)

There are no N/A values in the dataset and also there are no duplicates present in the dataset. We then check for the total number of 1's and 0's present in the Outcome of the complete data. It was seen that there were 500 0's and 268 1's without removing the outliers and 481 1's and 248 1's on removing the outliers.
A histogram is plotted to display the total number of 0's and 1's are there in the Outcome.

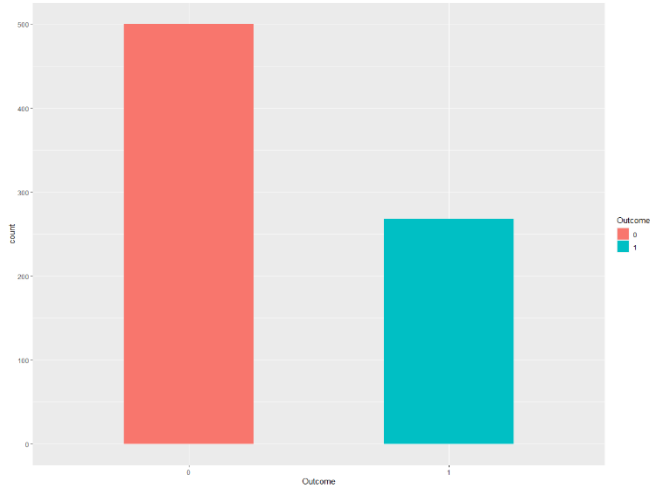

Fig. 2. Histogram showing the Positive (1's) and Negative(0's) ration

We also plot the Correlation plot from which we derive an inference that features such as pregnancy and age, skin thickness and body mass index and glucose and insulin are correlated to each other.

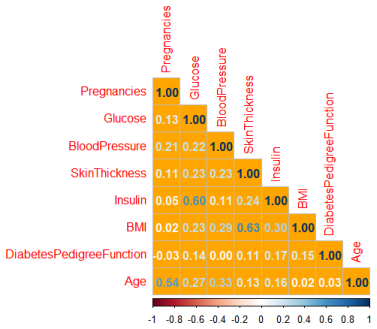

Fig. 3. Correlation Plot showing the correlation between the parameters without outlier removal

We have used Box plot and histogram to analyse the dataset. Each parameter against Outcome is plotted using Box plot which helped in understanding the presence of outlier in DPF (Diabetes Pedigree Function), age, Insulin, glucose, BMI and blood pressure feature which could because of some underlying factors. Hence it would be important to normalize the data to curb the effects of the outliers.

The figure below presents a single image with plots of parameters (BMI, Blood Pressure, Insulin, Glucose, Skin Thickness, Pregnancy) versus Outcome.

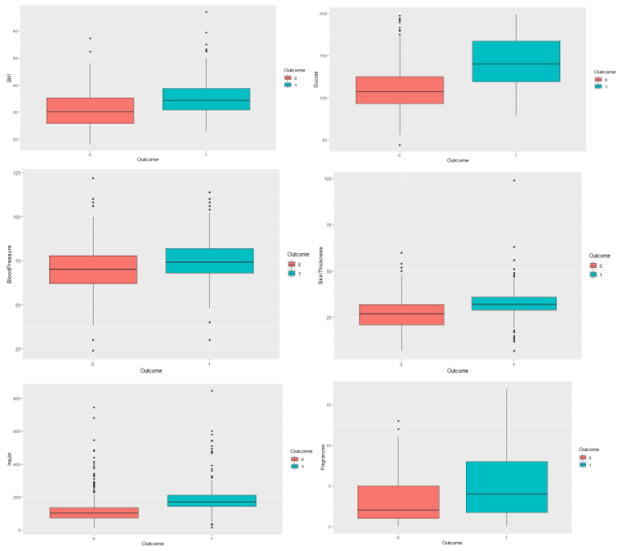

Fig. 4. Single image containing all the Boxplot of each parameter against Outcome

From the histograms plotted for each parameter against the outcome we can infer that Pregnancy does not cause diabetes as we see no trend. The parameter that contributes for diabetes is the Glucose, if the glucose level is high the changes of being a diabetes is high. The blood pressure level cannot be considered to predict diabetes since it is normally distributed.

We can also see that the diabetes increases with increase in BMI.

The figure below shows a single image with all of the parameters' histograms plotted against Outcome.

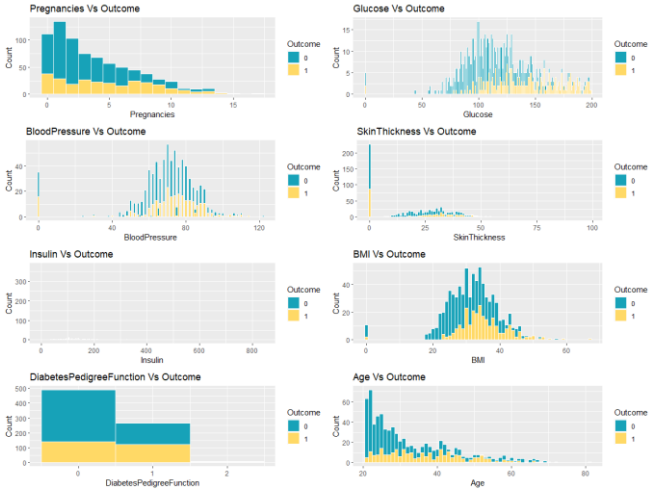

Fig. 5. Single image containing all the Histogram of each parameter against Outcome

B. Data Pre-processing

As the first step we checked for the N/A values and it was found that there were no N/A values but rather There are several zero values in the dataset. These are, in fact, missing data, as variables such as Glucose, Insulin, Blood pressure, Skin Thickness and BMI cannot be zero. Insulin and glucose are the most inadequate variables.

Therefore, to impute these missing values we have used KNN and MICE imputation. The figure below shows the histogram of values missing in each parameter.

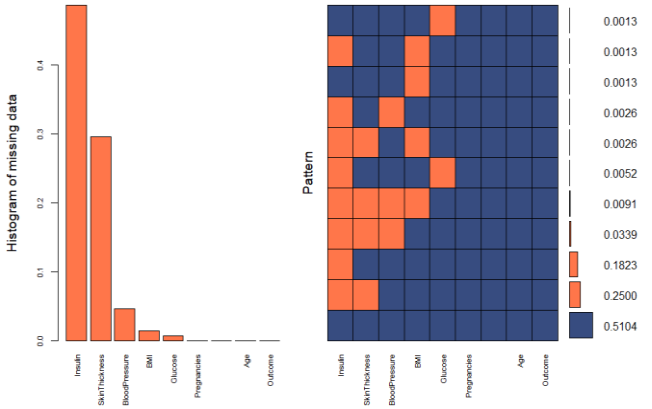

Fig. 6. Histogram and Pattern of Missing data

In the event of KNN imputation, the method uses the remaining variables to estimate the missing value. It is tested in multidimensional space what values of an unknown variable are taken by a set of neighbours. The number of ten nearest neighbours was employed in this situation.

In the event of MICE imputation, the missing values are imputed by using the method pmm(Predictive Mean Matching) and this is performed on every missing value. This type of imputation is referred to as "place holders." Imputations for one variable are set back to missing as a "place holder." In the imputation model, the observed values from the variable are regressed on the other variables. The regression model's predictions (imputations) are used to fill in the missing values for variables. This process is repeated for each missing data point. One iteration consists of cycling over each of the variables. After one iteration, all missing values have been replaced with regression predictions that represent the observed relationships in the data. Five cycles are used to impute all of the missing data. Once the imputation of missing data was done the presence of any outlier is done.

The presence of outliers has an effect on the accuracy of the prediction model, so removing them is essential. To find multivariate outliers, the Mahalanobis Distance will be employed. It's a type of distance calculated using standard deviations between two points in multidimensional space. The distance between each point in the n-dimensional data and the center is calculated to detect outliers using Mahalanobis distance, and outliers are derived from these distances. 5% of data with the largest distance will be eliminated. The box plots are more symmetrical when multivariate outliers are removed. The data in which there were 17 pregnancies has been removed.

We also plot the Correlation plot for the outlier removed data and we noticed that females with high blood glucose levels have a much higher chance of acquiring diabetes. There is also a link between BMI and diabetes in the data. One of the primary causes of diabetes is obesity. Surprisingly, overweight women's have a higher risk of developing diabetes.

The below figure displays the correlation between the parameters.

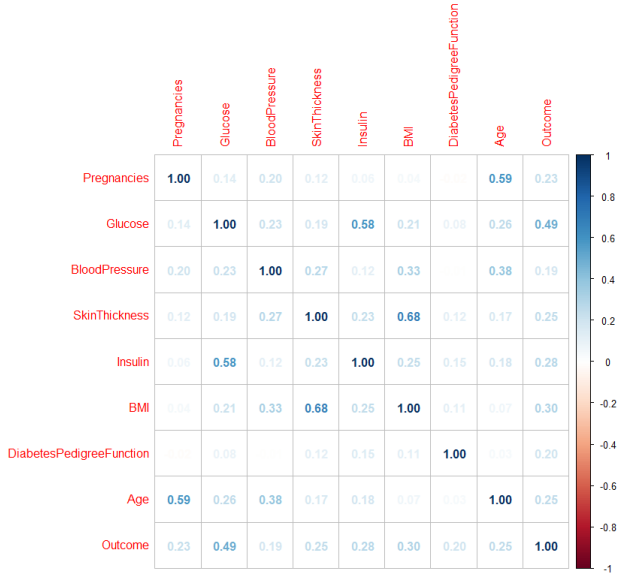

Fig. 7. Corrplot displaying the correlation of Variables after removing Outliers

An important feature selection that we can use is the Boruta which is a wrapper algorithm that is capable of working with any classification algorithm and statistically signifies the importance of certain variables. It performs a top-down search approach by comparing original attributes with importance achievable at random, estimated using their permuted copies and eliminating irrelevant features to make accurate predictions on the test.

When we conducted Boruta test on this dataset we find that blood pressure is confirmed to be the most unimportant attribute of the dataset and insulin and glucose are the most important.

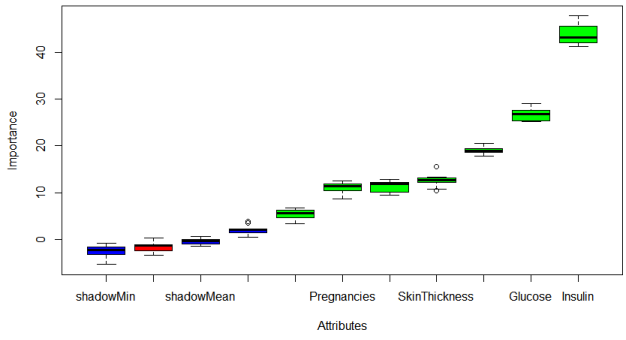

Fig. 8. Box plot showing the Boruta Technique

Apart from the Boruta Technique we have the principal component analysis as well that is a dimensionality reduction technique and a very prominent feature selection method used in pre-processing. With only 8 features in the dataset, it would not be a good idea to reduce the features further but using featural methods such as principal component analysis we can reduce noise in classification and comment about which features are closely related to each other.

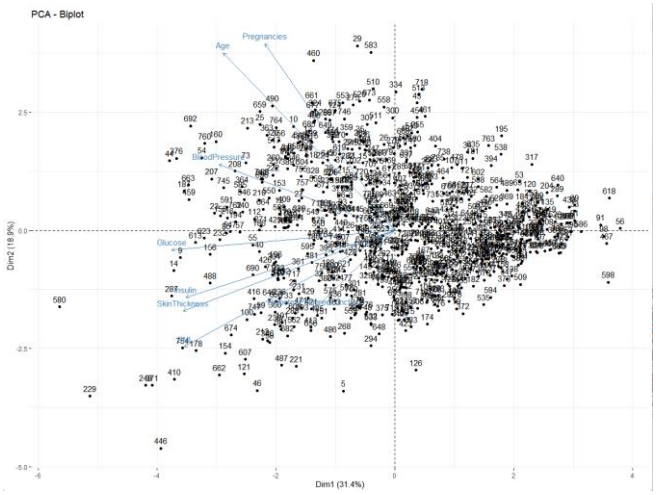

Fig. 9. Scatter plot showing the PCA Analysis

From the figure it is safe to infer that arrows that are adjacent to each other tell us about the attributes that are closely related to each other. The following are pretty closely related to each other: Age , Pregnancies, BMI, DPF, Insulin level, and Skin Thickness , Glucose and Blood pressure. So through pre-processing techniques like the principal component analysis and the Boruta technique we are in a way making it easier for us to reduce the dimension of the data and classify the data even precisely and accurately by focusing on the attributes that rank highly on the importance scale and those attributes that are closely related to the others. Blood pressure does not do too well on the Boruta plot so we can scale down the attributes to everything else but the blood pressure attribute.

C. Classification

In this paper we have used the two machine learning algorithms in the form of Naive Bayes Classifier and the Logistic Regression to analyze the Pima Indian heritage dataset. The data was manually divided into training and testing dataset respectively with respect to 70/30 split.

Metrics such as Classification Accuracy, Precision, Sensitivity, Specificity, F-score and Area under ROC Curve. These metrics are eventually calculated using the confusion matrix which is a special kind of contingency table that is useful for summarizing the performance of any particular classification algorithm. They give us a concise summary about the True Positive, False Positive, True Negative and False Negative Values and basically talks about how effective a particular classification algorithm is when it comes to actually classifying an instance correctly.

The classification accuracy refers to the ratio of the sum of true positives and true negatives to the sum of all the predictions combined. It gives us a fair idea about how effectively accurate our classification algorithm is.

The specificity refers to the ratio of the samples that were actually correctly predicted to be false to the sum of the samples that were correctly predicted to be false along with those that were false even if predicted incorrectly.

The sensitivity refers to the ratio of the samples that were actually correctly predicted to be true to the sum of the samples that were predicted to be true but were actually false along with the ones which were correctly predicted to be true

The Precision refers to the ratio of the sample that were correctly predicted to be true among all those that were predicted to be true even though they were false.

F-score is a standard indicator about the classification efficiency of the algorithm. In mathematical terms it refers to the ratio of the product of the precision and sensitivity to the sum of the precision and sensitivity and multiplying the result by 2.

Mathematically all the above metrics can be written as follows:

Accuracy = (TP+TN)/(TP+TN+FP+FN)
Sensitivity =TP/(TP+FN)
Specificity =TN/(TN+FP)
Precision =TP/(TP+FP)
f-score = 2*((Precision * sensitivity)/(Precision + sensitivity))

AUC-ROC Curves:

The area under receiver operating characteristic curve or the AUC-ROC curve is a graphical plot that shows the comparison between sensitivity and specificity and signifies the diagnostic ability of a binary classification algorithm. Classifiers that give curves closer to the top-left corner indicate a better performance. To get started with graphically plotting the AUC curve, it is important to create a baseline, any random classifier is expected to get points plotted diagonally to the sensitivity and specificity curve. A significantly good classifier will deviate from this baseline diagonal plot and will be more accurate.

The main outline of the confusion matrix can be summarized as follows:

TABLE II. OUTLINE OF CONFUSION MATRIX

	ACTUAL NEGATIVE	ACTUAL POSITIVE
PREDICTED NEGATIVE	TRUE NEGATIVE	FALSE POSITIVE
PREDICTED POSITIVE	FALSE NEGATIVE	TRUE POSITIVE

After performing Logistic Regression, the confusion matrix for the testing dataset that has 230 instances in it we get:

TABLE III. CONFUSION MATRIX FOR LOGISTIC REGRESSION

	ACTUAL POSITIVE	ACTUAL NEGATIVE
PREDICTED POSITIVE	128	22
PREDICTED NEGATIVE	25	55

After performing Naive Bayes Classification, the confusion matrix we get:

TABLE IV. CONFUSION MATRIX FOR NAIVE BAYES CLASSIFICATION

	ACTUAL NEGATIVE	ACTUAL POSITIVE
PREDICTED NEGATIVE	121	31
PREDICTED POSITIVE	23	42

TABLE V. COMPARISON OF VALUES BETWEEN LOGISTIC REGRESSION AND NAÏVE BAYES

MODEL TYPE	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)	PRECISION (%)	F-SCORE	AUC SCORE
LOGISTIC REGRESSION	79.6	71.4	83.7	68.7	0.70	0.86
NAIVE-BAYES CLASSIFIER	75	57.53	84.02	64.61	0.60	0.82

The important performance measures for classification models are receiver operating characteristics (ROC) curve and the resulting area under the curve (AUC) as they show the degree of class separation. The likelihood that the described model ranks a random positive example higher than a random negative example is indicated by AUC. AUC is a two-dimensional area beneath the ROC curve that ranges from (0.0) to (1.1), with a maximum value of 1.

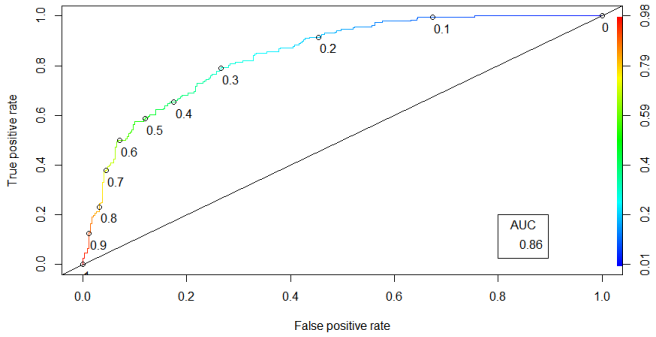

Fig. 10. ROC of Logistic Regression

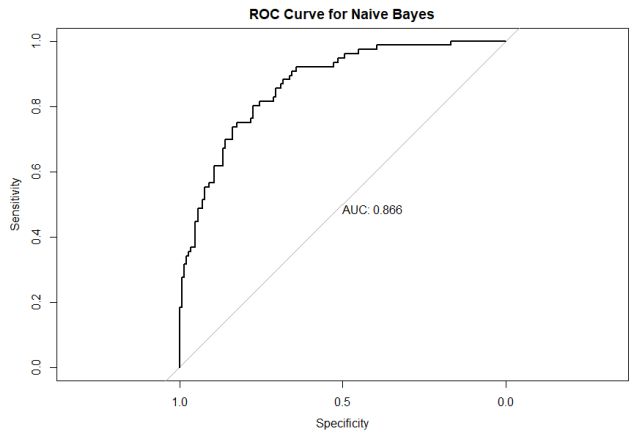

Fig. 11. ROC of Naïve Bayes

V. DISCUSSION

This research paper presented classification models suitable for predicting Diabetes in females of PIMA Indian heritage. We used two instances of the same dataset to perform two different classification algorithms and compare and contrast the metrics associated with how accurate the classifier is.

The experimental result table on the PIMA dataset shows that Logistic Regression Classifier outperforms the Naive Bayes Classifier when it comes to metrics such as classification Accuracy (79.6%), Precision (68.7%), Sensitivity (71.4%), AUC-value (0.86) and F-score (0.70). Naive Bayes Classifier performs minutely better on the specificity parameter(84.02%). It would not be wrong to assume that the major difference between the sensitivity and specificity is due to the imbalance caused by different amounts of diabetic and non-diabetic samples. The major differences in the results are mainly caused by the different operations that were conducted on the dataset before actually conducting the classification and fitting the model.

We performed different sets of pre-processing and feature selection algorithms on the two instances of the dataset. For the dataset that underwent classification using logistic regression, we dealt with minor inconsistencies in the data like zero values for insulin, skin thickness and glucose by performing KNN imputation on the data which made the dataset more cohesive and consistent. We also conducted feature selection analysis tests on the data such as Boruta and Principal Component Analysis that gave us a brief idea about the important features of the dataset which was further used to improve the classifier by checking for statistically insignificant features which were later found out to be Blood Pressure, Age and Skin Thickness(since they had p-values that were more than 0.01) as well as reducing the dimensionality of the data. Before doing PCA, the data was also standardised by scaling it.

For the other dataset instance, that underwent classification using Naïve Bayes Classification, we dealt with the same inconsistencies using a different imputation algorithm called Multiple Imputation by Chained Equations which is helpful in imputing mixes of binary, continuous and categorical data. After the data was consistent enough, we removed the outliers in the data to make the data more statistically powerful and significant.

Although the models used in this experiment have classification accuracy less than 80 % and can be improved, our research outputs are in line with similar work presented by Aishwarya and Vaidehi [1] that got classification accuracy of 96% with Logistic Regression and 93% with Gaussian Naive Bayes Algorithm. In our case, there was no overfitting done and results are genuine and in line with their approach. In both cases Blood Pressure was the most statistically insignificant feature and hence was not used in prediction purposes.

VI. CONCLUSION AND FUTURE RECOMMENDATIONS

In this paper, we have predicted diabetes on the females of PIMA Indian Heritage dataset. In this dataset, we have dealt with attributes such as Glucose, Age, Pregnancies, Insulin, Diabetes Pedigree Function, Skin Thickness, BMI and the target outcome variable. We have conducted data analysis and pre-processing to refine the data and make it more consistent and cohesive. We have used supervised machine learning classification algorithms such as Logistic Regression and Naive Bayes Classifier on the data and computed training and testing datasets. Metrics such as Accuracy, Precision, Specificity and Sensitivity are defined and computed for the

dataset that underwent Logistic Regression and Naive Bayes Classification. Upon comparison, we conclude that Logistic Regression was the better classification algorithm with a better accuracy and AUC score but Naive bayes Classification had a better specificity score. Naïve Bayes did a bit better when it came to ruling those people out that did not have diabetes.

The dataset used was fairly small and contained limited attributes to describe the patients. Our Future work will include developing and using other innovative classification algorithms on more complex medical datasets. For example, if the quality of the data received can't be improved upon, our accuracy can be by inserting more layers of pre-processing techniques for data analysis.

REFERENCES

[1] A. Mujumdar, V. Vaidehi, "Diabetes prediction using machine learning algorithms International Conference on Recent Trends in Advanced Computing", 2019, ICRTAC (2019)

[2] T.N. Joshi, P.M. Chawan, "Logistic regression and SVM based diabetes prediction system Int. J. Technol. Res. Eng., 11 (5) (2018) July-.

[3] E.O. Olaniyi, K. Adnan Onset diabetes diagnosis using artificial neural network Int. J. Sci. Eng. Res., 5 (2014), pp. 754-759.

[4] G. Swapna, K.P. Soman, R. Vinayakumar Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals Procedia Comput. Sci., 132 (2018), pp. 1253-1262.

[5] D. Sisodia, D.S. Sisodia, "Prediction of diabetes using classification algorithms" Procedia Comput. Sci., 132 (2018), pp. 1578-1585

[6] N. Yuvaraj, K.R. SriPreethaa, "Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster Cluster Comput.", 22 (2017), pp. 1-9

[7] UCI Machine Learning — Repository: https://www.kaggle.com/uciml/pima-indians-diabetes-database

[8] Pouya Saeedi, Inga Petersohn, Paraskevi Salpea, Belma Malanda, Suvi Karuranga, Nigel Unwin, et al., Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas 9th edition Diabetes Research and Clinical Practice, vol. 157, pp. 107843, 2019,[online]

Available: https://doi.org/10.1016/i.diabres.2019.107843

[9] Victor Chang, Jozeene Bailey, Qianwen Ariel Xu, Zhili Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms".

[10] M.S. Barale, D.T. Shirke, "Cascaded Modeling for PIMA Indian Diabetes Data", International Journal of Computer Applications (0975 - 8887)

[11] The MICE Algorithm - Sam Wilson [Online]. www.cran.r-project.org. Available from - https://cran.r-project.org/web/packages/miceRanger/vignettes/miceAlgorithm.html.

[12] kNN Imputation for Missing Values in Machine Learning - Jason Brownlee [Online]. www. machinelearningmastery.com. Available from – https://machinelearningmastery.com/knn-imputation-for-missing-values-in-machine-learning

[13] Outlier Detection with Mahalanobis Distance [Online]. www.r-bloggers.com. Available from – https://www.r-bloggers.com/2016/12/outlier-detection-with-mahalanobis-distance/

[14] ggplot2 from r-statistics.co by Selva Prabhakaran [Online]. Available from http://r-statistics.co/ggplot2-Tutorial-With-R.html