Diabetes Prediction Using Artificial Intelligence (AI) Algorithms

Proposed by:
Ummer Shakeel (20-MS-DS-18) & Noman Khan (20-MS-DS-01)
Supervised by:
Dr. Munawar Iqbal
University: UET Taxila

ABSTRACT

Diabetes is a significant disease that affects a large number of individuals. Diabetes mellitus may be caused by various factors such as age, obesity, genetic diabetes, a lack of physical activity, lifestyle, lousy nutrition, high blood pressure, and so on. Diabetes patients are at an increased risk of developing heart disease, renal disease, stroke, vision difficulties, and nerve damage. The current standard of care in the hospital is to gather the necessary information for the diagnosis of diabetes via a variety of tests and then to offer suitable therapy depending on the results of the tests. Several machine learning algorithms are implemented in this proposal to determine which is the most effective algorithm for predicting diabetes in humans. For this purpose, we implement various types of machine learning and deep learning algorithms to evaluate using the cross-validation method, which classifiers include Linear Discriminant Analysis, Classification Regression, Generalized linear model, KNN, SVM, Random Forest, Naïve Bayes, Adaboost and Artificial Neural Network . In this example, we will utilize R-programming and python google-colab to predict diabetes using various algorithms. Accuracy may help identify whether or not a class has the same number of samples as the other classes; however, accuracy is rendered worthless when we have an unbalanced collection of data to deal with. Then we looked into additional performance measures like as recall, precision, and F1-Score to see what we might learn. ANN has an outstanding F-measure (F-1 score) of 80.28% that sets her above all other models in the competition.

INTRODUCTION

According to recent research, the number of diabetics patients increased from 108 million in 1980 to 422 million in 2000, indicating that 8.5 percent of the world's population now has diabetes, up from 4.7 percent in 1980. We may learn from the fact that it is essential to diagnose and adjust the sugar level.

Diabetes is a condition in which the amount of glucose in the blood increases. Blood glucose levels tend to rise as a result since sugar is required for the primary energy source. However, both low and high blood glucose levels, referred to as blood sugars, may result in various diabetes-related health complications. The amount of glucose in the blood increases insulin production. Whenever it is high, blood sugar levels increase, and when it is low, blood sugar levels decrease. Insulin can activate the muscle, fat cells, and red blood cells, causing them to take glucose from the bloodstream and use it for essential energy, returning blood glucose levels to normal. Diabetes may be classified into three categories Type-1, Type-2 and Gestational diabetes are the most common.

In Type-1, Insulin production may be limited or absent altogether in diabetics, resulting in cells unable to absorb glucose. This causes high blood glucose levels and various health complications such as heart attack, renal failure, and kidney disease.

In Type-2, The insulin generated may not be adequate (i.e., minimum level of insulin required), or the body may resist accepting insulin from the lead, resulting in elevated blood glucose levels. As a consequence, a large number of individuals die.

Gestational diabetes is a condition that develops during pregnancy. When a hormone generated by the placenta interferes with the body's capacity to utilize insulin effectively during pregnancy, it is referred to as Gestational Diabetes Mellitus (GDM). The glucose in the circulation, instead of the glucose taken by cells, accumulates in the bloodstream.

Diabetes diagnosis is the most important research direction in the medical sciences. In Predictive Analysis, various Machine Learning (ML) algorithms and statistical techniques are used to discover information and anticipate future occurrences. This methodology is based on current and historical data to discover knowledge and anticipate future events. When predictive analytics is applied, it is possible to make meaningful choices and predictions based on healthcare data. In addition to the regression method, machine learning may be used to do predictive analytics. Predictive modeling improves clinical outcomes by detecting diseases with the highest possible accuracy, improving patient care, maximizing resources, and increasing patient safety. A significant amount of research has been carried out to get more knowledge about the risk factors for diabetes and or before and how to diagnose them. Despite this, only a few studies have been conducted to determine the prevalence of diabetes-related complications, particularly those associated with health risk factors. Diabetic consequences diseases, as a result, continue to be underused in disease preventive efforts, and they are often identified only after the disease has shown itself in a potentially dangerous condition.

Insulin resistance is regarded to be one of the most serious illnesses impacting modern society. A growing number of patients are being admitted daily. The illness results in a wide range of different types of health issues. Heart attack, renal failure, excessive blood pressure, and diabetic neuropathy are just a few examples. A fasting and random blood sugar test are performed to diagnose diabetes, and thorough consultation with the patient is needed. The sample facility carries out the test. An abnormally high glucose level indicates the presence of diabetes in a patient in the blood. Diabetes mellitus (DM) is a metabolic disorder characterized by an abnormally high insulin level in the bloodstream. An insufficient quantity of insulin causes high blood glucose levels. Insulin resistance and high blood sugar levels are two characteristics of diabetes mellitus, the most common type of the disease. The disease will grow more common in the future years as a result of the way that the younger generation lives their life.

Artificial Intelligence Algorithms

The rapidly developing area of machine learning and deep learning are being used in a variety of medical-related endeavors. All machine learning models learn from their predecessors and make predictions based on a given set of training data. It is a branch of artificial intelligence that studies machine learning and deep learning. The diagnosis of diabetes will become easier and less expensive due to recent advancements in machine learning. A large number of diabetes-related datasets are accessible. As a result, machine learning is needed for applications in medical diagnostics. Our goal is to forecast the likelihood of developing diabetes in a patient based on their medical history using machine learning algorithms.

Specifically, for the sake of our research, there are two kinds of learning.

a. Supervised learning
b. Unsupervised learning
c. Reinforcement Learning

Based on labeled data, an algorithm for supervised learning is used to make predictions. In supervised learning, the data is labeled before being used. It is a simulation of learning from a teacher. Unsupervised learning, on the other hand, does not need any labeling of the data. It's more like self-learning based on previous experience rather than formal education. The goal is to make an educated guess about the value of a variable. Each characteristic and feature of the data is represented as a separate object in the database. In supervised learning, the outcome is predetermined before the process begins. Techniques, including Decision Trees (DT), Rule Learning, and Instance-Based Learning (IBL), as well ask Nearest Neighbors (k-NN), Naive Bayes, Genetic Algorithms (GA), Artificial Neural Networks (ANN), and Support Vector Machines (SVM), are the most often used (SVM) etc.

Unlike supervised learning, unsupervised learning uses values as input data rather than labels, and the outcome is not predefined. The model makes predictions based on the fact that it is always learning. The primary goal of these models is to forecast, classify, detect, segment, and categorize information. The most frequent applications of machine learning are in analysis, recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics, among others.

It is important to use a penalty method in reinforcement learning because if a detector does not detect properly, a penalty is imposed for each incorrect detection. There is no previous information given; only testing is carried out to learn the model via trial and error and, if feasible, through formal learning. Because of its independence from its surroundings, reinforcement learning is most often used in autonomous systems.

Machine Learning Process

The use of machine learning techniques for predicting diabetes is becoming more popular, and the outcomes are favorable. Machine learning (ML) models are created via a highly iterative process that involves learning from past events and evaluating historic data. AI algorithms are also capable of identifying trends in a way to produce predictions of a given sample dataset.

Machine learning methods are utilized in this proposed study to predict if a person has diabetes or not, using patient data such as blood pressure, BMI, age, and glucose depending on the degree of obesity, etc.

When it comes to artificial intelligence, machine learning is regarded as one of the most significant characteristics. It helps to facilitate the creation of computer systems that can learn from their previous experiences without programming in every situation. It is widely acknowledged that machine learning is a crucial need in today's environment. It allows for eliminating the need for human involvement by allowing automation with the lowest possible amount of errors. For the time being, laboratory procedures such as fasting blood glucose and oral glucose tolerance testing are used to identify people who have diabetes. This technique, on the other hand, is time-consuming. This article focuses on developing a diabetes prediction model utilizing AI algorithms and predictive analytics methods to predict the disease.

Collects data, does classification or prediction, and concluded by fitting the proper classifier to accurately predict values. It has developed its own set of algorithms to enhance scores and predictions.

We use a step-by-step method; in this case, we have an ML-pipe line.

A Standard Machine Learning Pipeline:

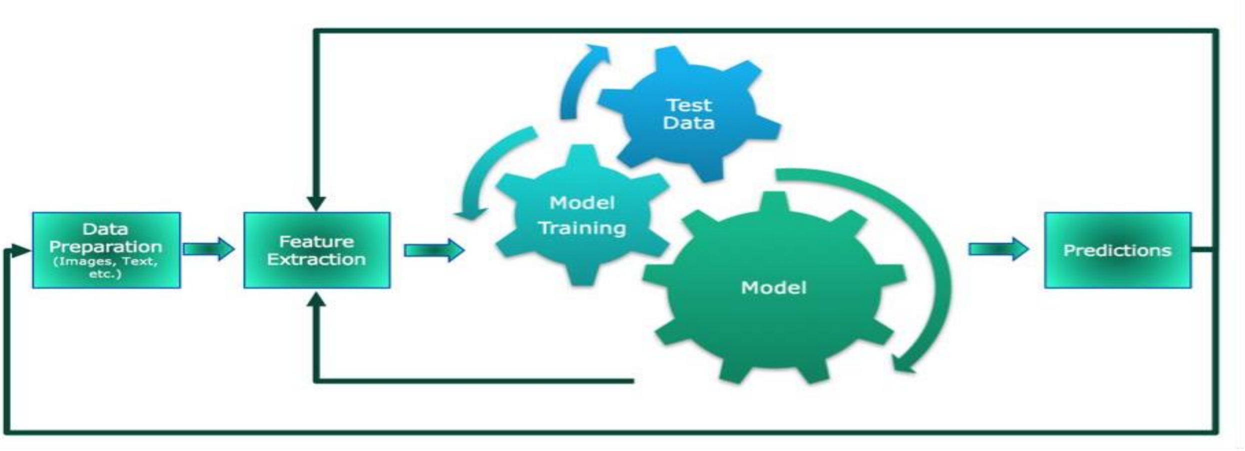


Figure Source: https://bit.ly/3BQ7Mmz

Machine Learning Steps

Data Acquisition

Data acquisition is the process of retrieving or locating a dataset from a readily accessible resource. We will be using the Kaggle website. Kaggle is a website that enables people to discover and post data sets. We obtained the diabetes dataset from the following website; link:*https://www.kaggle.com/mathchi/diabetes-data-set* . This dataset is initially from the National Institute of Diabetes and Digestive and Kidney Diseases.

Data Exploration and Analysis

Data exploration is the first step of data analysis, during which users seek patterns, characteristics, and points of interest in a massive array of randomly organized data. In Data Exploration, human techniques and automated tools such as data visualization, charts, and first reports may be used in conjunction with one another to produce results.

Data Cleaning and Preprocessing

Data cleaning is the process of checking for missing data, null values, repeated values, smoothing or eliminating noisy data and outliers, as well as addressing anomalies and inconsistency correction. Data preprocessing is a data mining technique that is used to convert raw data into forms that are useful and effective. Data preprocessing is a data mining technique that involves transforming raw data into a format that can be interpreted by other applications. In real-world situations, data is often insufficient, inconsistent, and/or missing in particular behaviors or patterns, and it is likely to include a significant number of mistakes. Preprocessing has been shown to be an effective way of addressing such problems.

Model Evaluation

The model evaluation aims to assess the correctness of a model's generalization of future (often missing/out-of-sample) data based on experience. The methods for assessing the performance of a model are split into two categories: holdout and cross-verification. For both approaches, the model's performance is evaluated using a test set (i.e., a set that is not observed by the data model). Model evaluation is a critical step in the model creation process and should not be neglected. It aids in selecting the best model to determine our data and the prediction of how well the selected model will perform in the future. Both approaches utilize a test set (not visible to the model) to assess model performance to prevent overfitting.

Prediction

Predictive modeling is the process of predicting outcomes based on data. In most cases, the event you wish to forecast will take place in the future. Nevertheless, predictive modeling may be used for any unknown event, regardless of whether it has already occurred.

Problem Statement

Since the turn of the century, there has been a dramatic rise in many individuals suffering from diabetes. The modern human lifestyle is the primary cause of diabetes's rise. There are three categories of errors in the current medical diagnosing system.

1. When a patient is diagnosed as having diabetes, the false-negative type occurs. Despite the fact that the patient is already diabetic, the results of the tests show that they do not have diabetes.
2. The second kind is the false-positive type. Patients in this category do not have diabetes in reality, but their test results indicate that they have diabetes.
3. The third kind is the undefinable type, which occurs when a system cannot diagnose a certain situation. This occurs due to inadequate information extraction from previous data, and a particular patient may be projected as belonging to an undefined category.

In reality, the patient must anticipate whether they will fall into the diabetic or non-diabetic group. Such mistakes in diagnosis may result in the giving of inappropriate therapies or the failure to provide any medication at all when essential. To prevent or decrease the severity of such an effect, it is necessary to develop a system that uses machine learning methods and data mining approaches to provide efficient estimation results while also reducing the time and effort required by humans.

RELATED WORKS

Machine learning and data mining techniques are used to create different classifications based on the resources that are made accessible to them. A 2011 study by Lawrence Fisher, Joseph T.Mullan, Russell E. Glasgow, and Umesh Masharani mentioned the neglect factor, saying that age and lifestyle variables, in particular, are essential for diabetes in the human body and its treatment. They also cited the neglect factor, adding that it was aimed at adults and the senior population. According to research, training based on poor automation systems may be produced automatically by utilizing a variety of learning algorithms, such as random forest algorithms based on specific categorization methods, which may be beneficial in the process.[1]

In the year 2014, Mira KaniaSabariah, AiniHanifa, and SitiSa'adah worked together on several projects. This paper describes an investigation into predicting type 2 diabetes in people before blood sugar levels get dangerously high. Many methods and approaches are used to implement this system to demonstrate that the first prognosis for categorizing a diabetic patient is extremely clear. His study has previously been carried out on Chinese individuals to evaluate the characteristics of datasets about complicated variables to determine whether or not a patient has diabetes. This study is based on clinical data that is publicly accessible. They have incorporated lifestyle characteristics into diabetes prediction and created a model for predicting with logistical regression to make these predictions more accurate.[2]

Janice Lopez, Marcia and Kathy Annunziata, and Robert Bailey are among those who have contributed to this work. Some risk factors, such as age, family history, physical inactivity, and obesity, are believed to significantly impact the occurrence of diabetes. It was completely dependent on the patient's viewpoint on the characteristics that were impacted. As a result, the finding indicates that characteristics are essential for prediction. Algorithms for machine learning are essential for the correct categorization of classes. It is suggested that suitable methods be identified to make accurate predictions. It should be kept in mind that the experience and observations described in this proposed study pertain only to the Pima Indian Diabetes Dataset. Following rigorous testing, our suggested approach has been determined to be the most effective in certain parameters.[3]

Durga Toshniwal and R.C. Joshi (2010), When the results of the clustering method are achieved, the suggested hybrid prediction model, which comprises a straightforward K-means clustering algorithm, then apply the classification model to the results of the proposed technique.[4]

Shraddha Kumar and Mani Butwall (2015) propose a system to diagnose diabetes using a Random Forest Classifier.[5]

Nawaz Mohamudally and Dost Muhammad created a diabetes prediction system based on the C4.5 decision tree method, Neural Network, K-Means clustering algorithm, and proposed an idea in 2011.[6]

Aiswarya Iyer (2015) used classification methods were used to uncover hidden patterns in diabetes datasets, according to the researchers. In this model, twenty-two trees were utilized, all of which were significant. The performance of both algorithms was compared, and the efficacy of both algorithms was shown as a consequence of the comparison.[7]

K. Rajesh and V. Sangeetha (2012) used classification model to predict diabetes.[8]

In this study, Sneha et al. utilized a modified method to extract important features from the Pima Indians Diabetes dataset. Those attributes were used for early diabetes prediction using SVM, decision tree, naive Bayes, and KNN classifiers, among other techniques.[9]

Researcher Yukai et al. developed diabetes prediction techniques based on SVM, decision tree, and supervised learning algorithms. The categorization tree category had less accuracy than the other two.[10]

There isn't a single disadvantage. Another well-known machine learning technique known as a recurrent neural network (RNN) is extensively utilized in various prediction tasks and is readily available online. The LSTM is a RNN unit that addresses long-term dependence among data and implements well in diabetes prediction since it has a fast learning time. Qingnan et al. presented a deep neural network-based model for diabetes prediction that used LSTM and Bi-LSTM. Comparing the proposed model's performance with SVR and ARIMA classifiers revealed that it overtook the other two models. A recurrent neural network (RNN), another popular machine learning method, is widely used in various prediction applications, including weather forecasting and conflict of interest prediction, among other things. A recurrent neural unit known as the LSTM addresses long-term dependence across data. Also, it operates well in the prediction of diabetic complications. Qingnan et al. presented an artificial learning network-based model for glucose prediction that used LSTM and Bi-LSTM. Comparing the proposed model's efficiency with ARIMA and SVR algorithms revealed that it exceeded some other two models.[11]

To predict diabetes disease, researchers used the Pima Indians Diabetes dataset to proposed research. They used three machine-learning techniques to predict diabetes: the Decision Tree, the Support Vector Machine, and the Bayesian Classifier (Naive Bayes). They discovered that the Naive Bayes classifier exceeded the other two algorithms with an accuracy of 76.30%.[12]

The Publication the pre-proof neural network contributes much to the prediction of diabetes since it is strongly predictive. Quan et al. examined various algorithms to prevent diabetes, including the decision tree, random forest, and neural networks.[13]

Apoorva et al. used the decision tree method to predict type-2 diabetes using the PID dataset. comparison of the results with the SVM algorithm supported the predicted type-2 diabetes in the decision tree.[14]

Swapna et al. have produced a diabetic and normal HRV signal categorization technique using long-term memory (LSTM), convolutional neural networks (CNN) and their variants.[15]

The RF, DT, SVM, CNN, ANN, RNN, LSTM, ARIMA, and SVR, work well for predicting diabetes. The random forest technique, on the other hand, performs well in classification problems, but it does not predict regression concerns further than range of training data. Because a slight change in the data causes a large change in the decision structure, the decision tree approach is unstable.[16]

Convolutional Neural Network (CNN) methods are similar to neural networks in that each neuron receives specific inputs. These neurons learn from weighted and biassed data via processes like the dot product study, which suggests based on CNN diabetes prediction.[17]

PROPOSED WORK

The goal of the performed study is to categorize the data provided into two categories: diabetic and non-diabetic. This will be accomplished by utilizing supervised learning algorithms. The dataset will be split into two parts: training sets and testing sets. To improve accuracy, we must train on a larger amount of data. A comparison study of the findings obtained from the algorithms for early diagnosis of diabetes will be conducted after that.

Machine learning techniques include Linear Discriminant Analysis, Generalized linear model, Classification and Regression Trees, K-Nearest Neighbors, Support Vector Machine, Random Forest, Naive Bayes, AdaBoost, and ANN are included in the suggested work. As previously stated in the introduction, the National Institute of Diabetes and Digestive and Kidney Diseases dataset has been used for sampling to put the proposed work into practice so that we can determine whether a person has been tested for diabetes and has returned a positive (0) or negative result (1). It has been put through its paces. Datasets were originally pre-implemented to replace lost values with resources rather than the median to improve the core trend, as is the case in this investigation for model training, pre-processed data is supplied then different models are trained to predict diabetes.

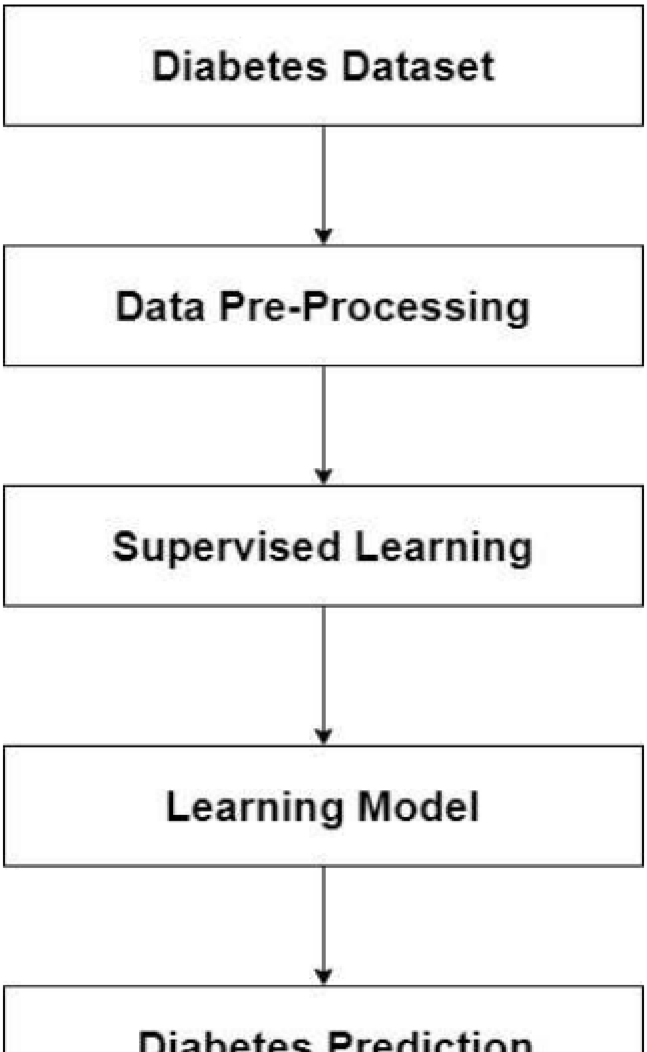
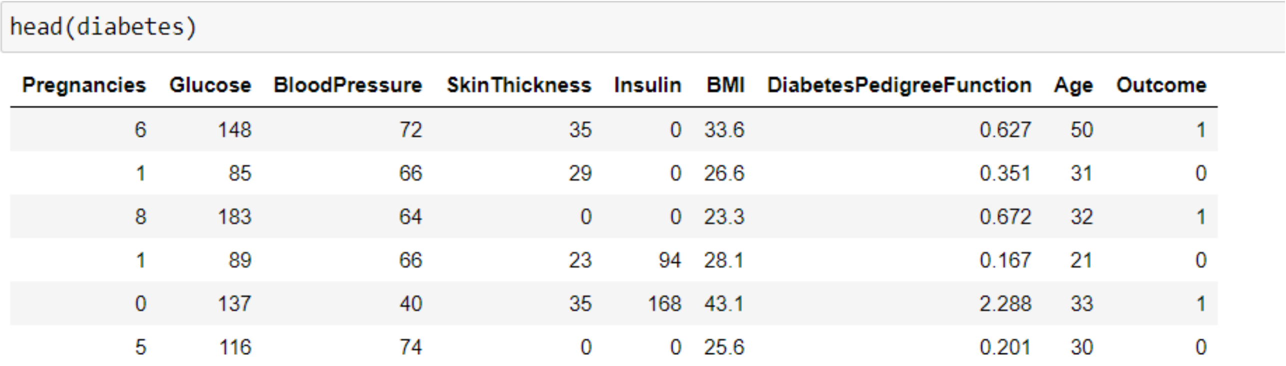


Figure: Proposed Methodology

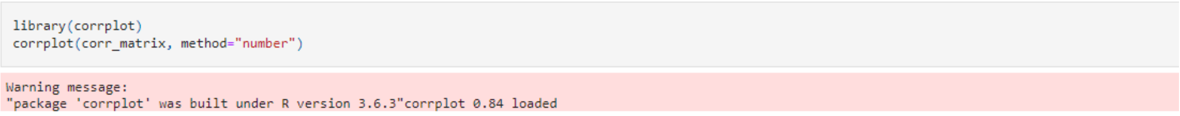
Different machine learning classifications employed in this research are presented individually for final verification concerning the accuracy of the findings to choose which model is most appropriate for the particular prediction job. The dataset used for the study may be seen in the following image, which highlights the most important characteristics of the dataset, which includes all of its attributes.



```
head(diabetes)
```

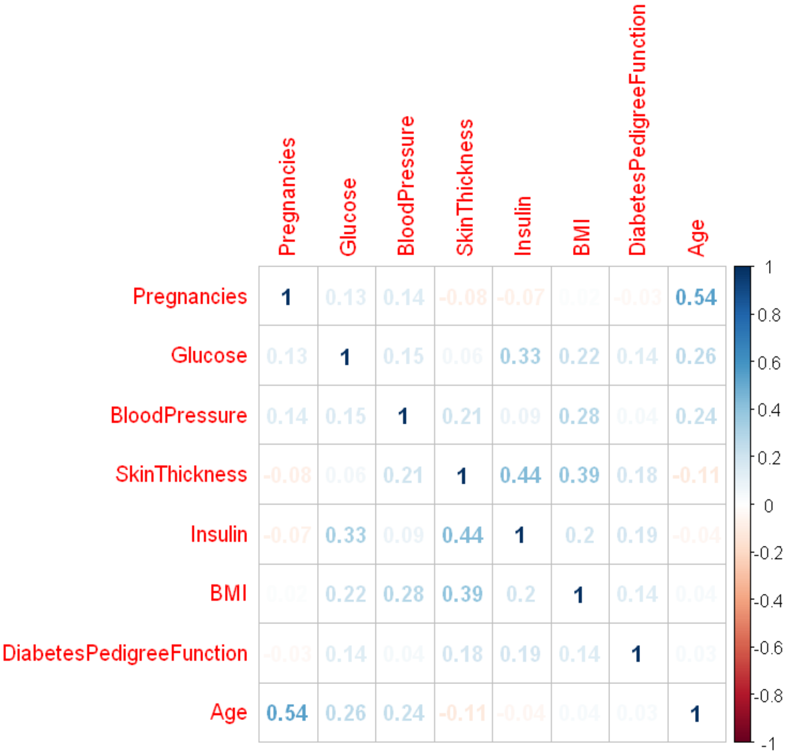
Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0

Several features are used to establish whether or not a person has diabetes. The relationship between the variables is shown in the correlogram below. Because the most important variables are shown in vivid colors, the number of pregnancies, body mass index and glucose levels all have significant relationships with the class variables. There are 768 rows and 9 columns in this dataset. The last column property is Outcome with 0/1 and 0:500/1:268 as its value. Below show the variables linear relationship between each attributes of diabetes dataset.



```
library(corrplot)
corrplot(corr_matrix, method="number")
```

```
Warning message:
"package 'corrplot' was built under R version 3.6.3"corrplot 0.84 loaded
```

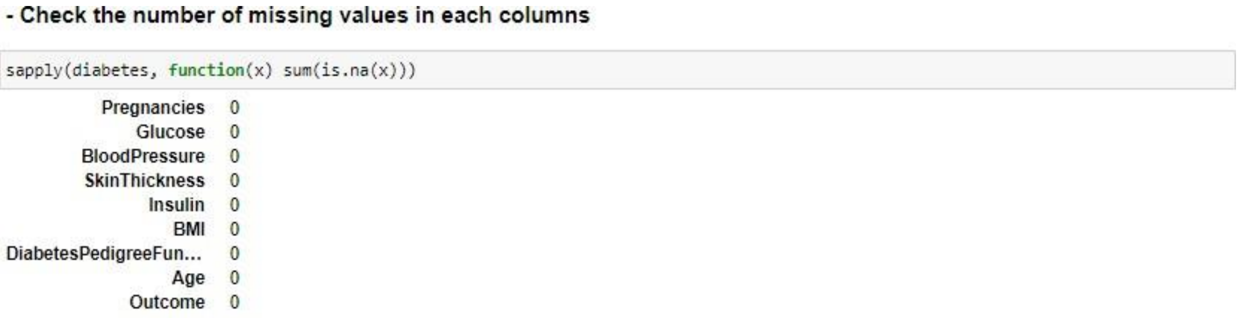


	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
Pregnancies	1	0.13	0.14	-0.08	-0.07	0.02	-0.03	0.54
Glucose	0.13	1	0.15	0.06	0.33	0.22	0.14	0.26
BloodPressure	0.14	0.15	1	0.21	0.09	0.28	0.04	0.24
SkinThickness	-0.08	0.06	0.21	1	0.44	0.39	0.18	-0.11
Insulin	-0.07	0.33	0.09	0.44	1	0.2	0.19	-0.04
BMI	0.02	0.22	0.28	0.39	0.2	1	0.14	0.04
DiabetesPedigreeFunction	-0.03	0.14	0.04	0.18	0.19	0.14	1	0.03
Age	0.54	0.26	0.24	-0.11	-0.04	0.04	0.03	1

Data Preprocessing

The preprocessing of data is a critical stage in the data mining process and should not be ignored. It is especially relevant to data mining and machine learning initiatives when the term "garbage in, trash out" describes the process. Data-gathering techniques are often not tightly regulated, resulting in out-of-range numbers, impossible data combinations, and missing information, among other things, as a consequence.

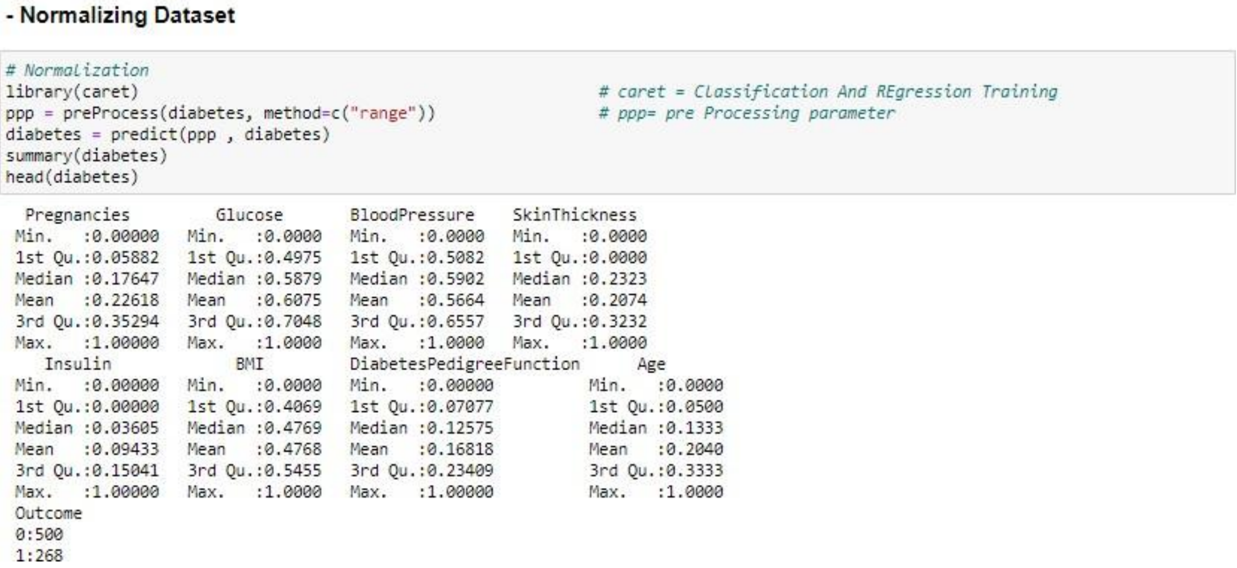
This phase of the model deals with data in dispute with one another, resulting in more accurate and exact outputs. To check some missing values in this dataset. Because the values of certain chosen variables, such as glucose level, blood pressure, skin thickness, body mass index (BMI), and age, can not be zero, we eliminated the missing values for those attributes from the model. Afterward, we scale the dataset to normalize all of the values, and for this purpose, the statistical Range technique is used.



- Check the number of missing values in each columns

```
sapply(diabetes, function(x) sum(is.na(x)))
```

```
         Pregnancies  0
             Glucose  0
       BloodPressure  0
       SkinThickness  0
             Insulin  0
                 BMI  0
DiabetesPedigreeFun...  0
                 Age  0
             Outcome  0
```



- Normalizing Dataset

```
# Normalization
library(caret)                                          # caret = Classification And REgression Training
ppp = preProcess(diabetes, method=c("range"))           # ppp= pre Processing parameter
diabetes = predict(ppp , diabetes)
summary(diabetes)
head(diabetes)
```

```
  Pregnancies        Glucose       BloodPressure    SkinThickness
 Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.05882   1st Qu.:0.4975   1st Qu.:0.5082   1st Qu.:0.0000
 Median :0.17647   Median :0.5879   Median :0.5902   Median :0.2323
 Mean   :0.22618   Mean   :0.6075   Mean   :0.5664   Mean   :0.2074
 3rd Qu.:0.35294   3rd Qu.:0.7048   3rd Qu.:0.6557   3rd Qu.:0.3232
 Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
    Insulin            BMI         DiabetesPedigreeFunction      Age
 Min.   :0.00000   Min.   :0.0000   Min.   :0.00000          Min.   :0.0000
 1st Qu.:0.00000   1st Qu.:0.4069   1st Qu.:0.07077          1st Qu.:0.0500
 Median :0.03605   Median :0.4769   Median :0.12575          Median :0.1333
 Mean   :0.09433   Mean   :0.4768   Mean   :0.16818          Mean   :0.2040
 3rd Qu.:0.15041   3rd Qu.:0.5455   3rd Qu.:0.23409          3rd Qu.:0.3333
 Max.   :1.00000   Max.   :1.0000   Max.   :1.00000          Max.   :1.0000
 Outcome
 0:500
 1:268
```

Model Evaluation

To assess the generalization accuracy on future (unknown/out-of-sample) data, it is necessary to conduct a model evaluation. Methods for assessing the performance of a model are classified into two categories: holdout and cross-validation. Both techniques rely on a test set (i.e. data that has not been viewed by the model) to assess the model's performance.

Data partitioning (70 % for training, 30 % for testing), model tuning, and model fitting are completed before the model is evaluated. Using different diagnostic measures, such as classification accuracy, confusion matrix recall, and sensitivity and F-measure (F-1 Score), we assess the prediction outcomes at this stage of the prediction model's development.

Confusion Matrix

A confusion matrix is a tabular or matrix representation of the number of accurate and erroneous predictions produced by a classifier during its training. To assess the effectiveness of a classification algorithm, it is employed. It may be used to assess the evaluation of a classification model by calculating performance measures such as accuracy, precision, recall, and the F1-score, among others.

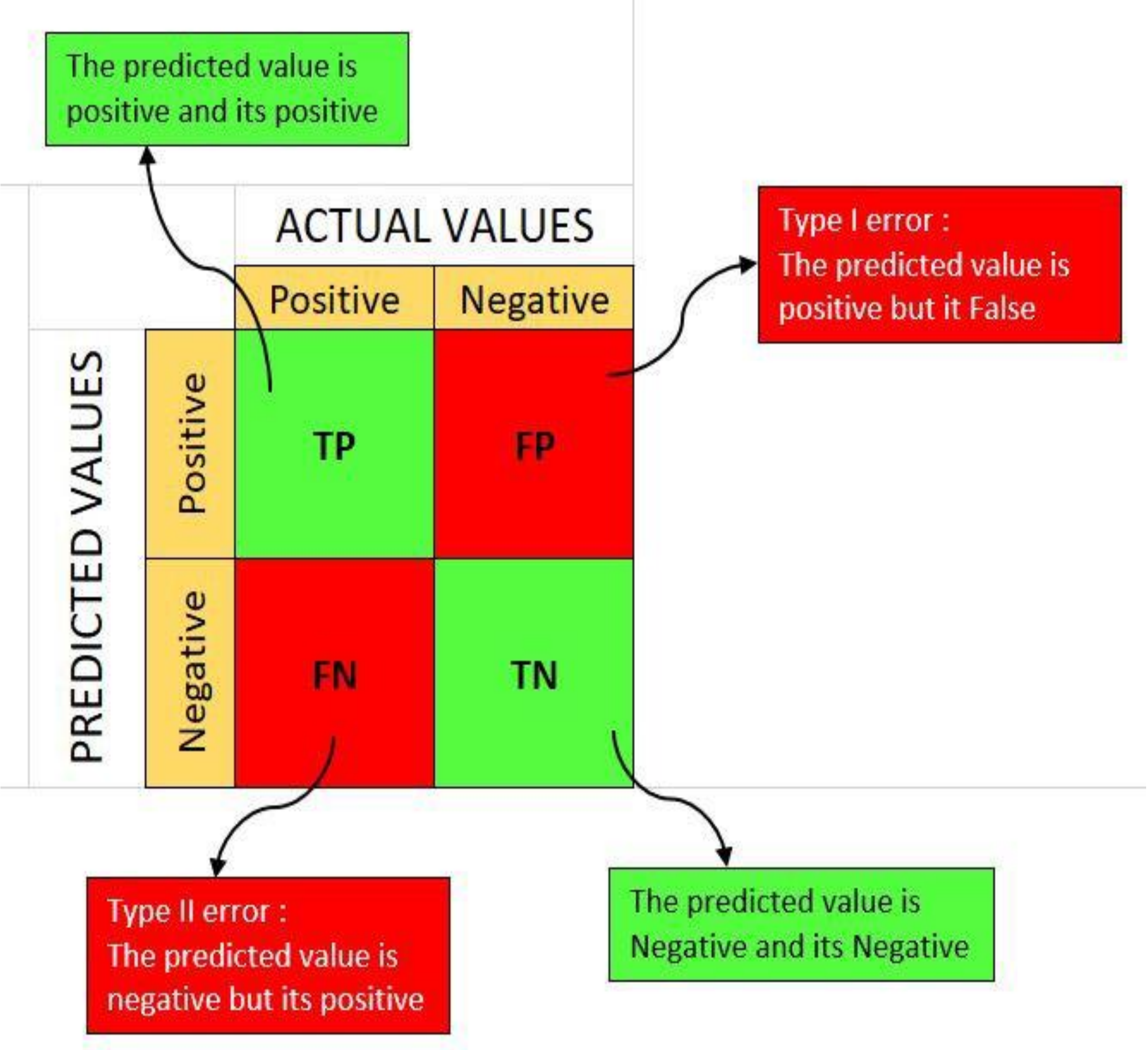

The predicted value is positive and its positive
ACTUAL VALUES
Positive
Negative
PREDICTED VALUES
Positive
Negative
TP
FP
FN
TN
Type I error :
The predicted value is positive but it False
Type II error :
The predicted value is negative but its positive
The predicted value is Negative and its Negative

Figure source : https://bit.ly/38ah7sq

Accuracy

Accuracy = Number of Correct Predictions / Total No. of Predictions made.

OR

Accuracy = TP + TN / (TP + TN + FP + FN)……………………………. (1)

Recall or Sensitivity

Recall = TP/ (TP+FN)……………………………………………………….(2)

Precision or Specificity

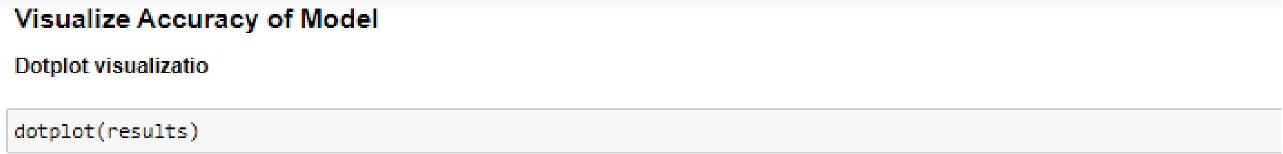
Precision = TN/ (TN+FN)……………………………………………….... (3)

F-1 Score

F-1 score = 2(Precision * Recall)/ (Precision + Recall)……………………(4)

Model Evaluation Base on Accuracy

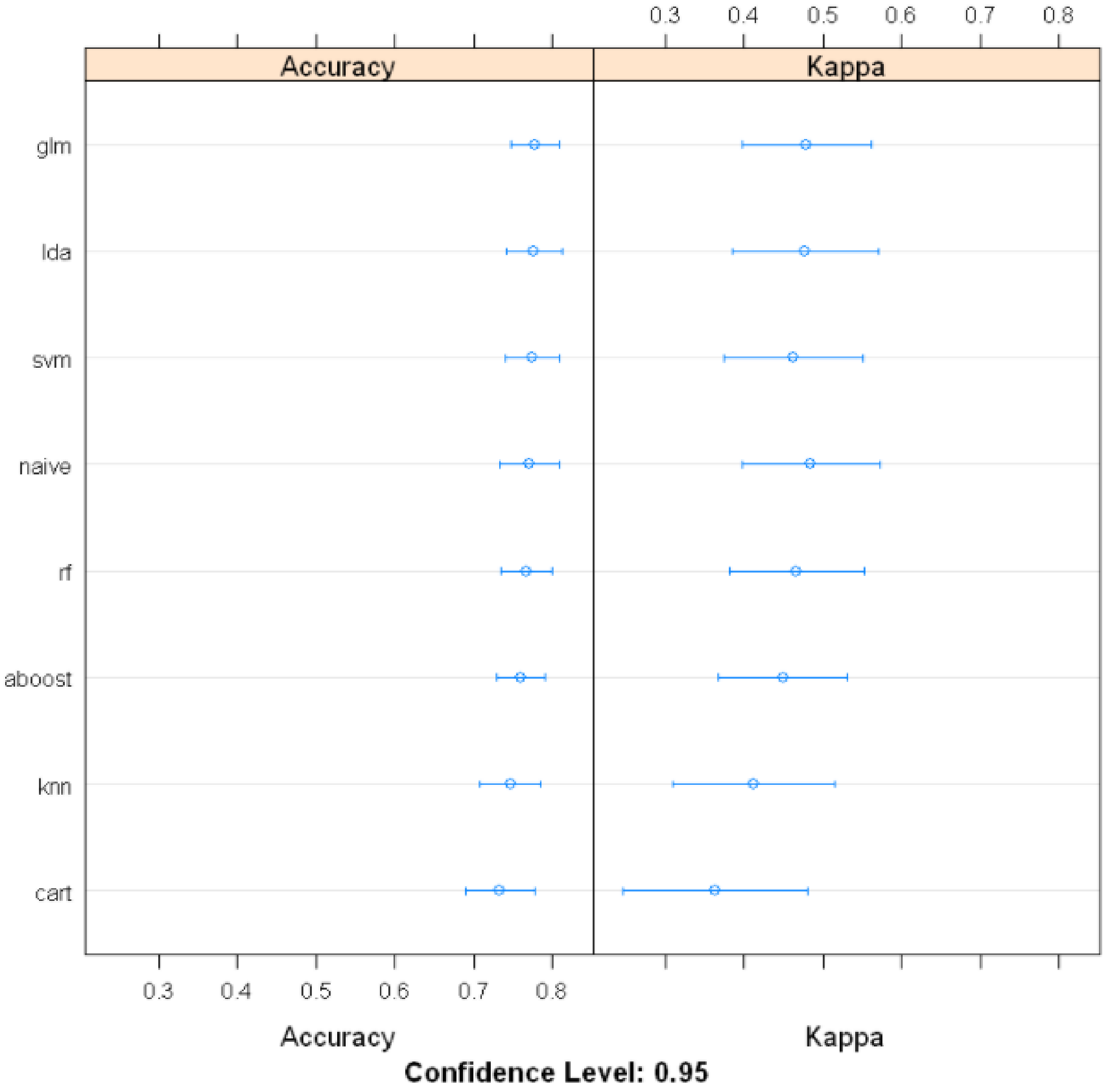
The accuracy of a prediction refers to the number of data points that are properly predicted. It is one of the most specific types of assessment measures. When assessing classification models, accuracy is one of the metrics to consider. Informally, accuracy may be defined as the percentage of correct predictions made by our model. Here below GLM and LDA models predicts a high accuracy rate with all others.



Visualize Accuracy of Model

Dotplot visualizatio

```
dotplot(results)
```



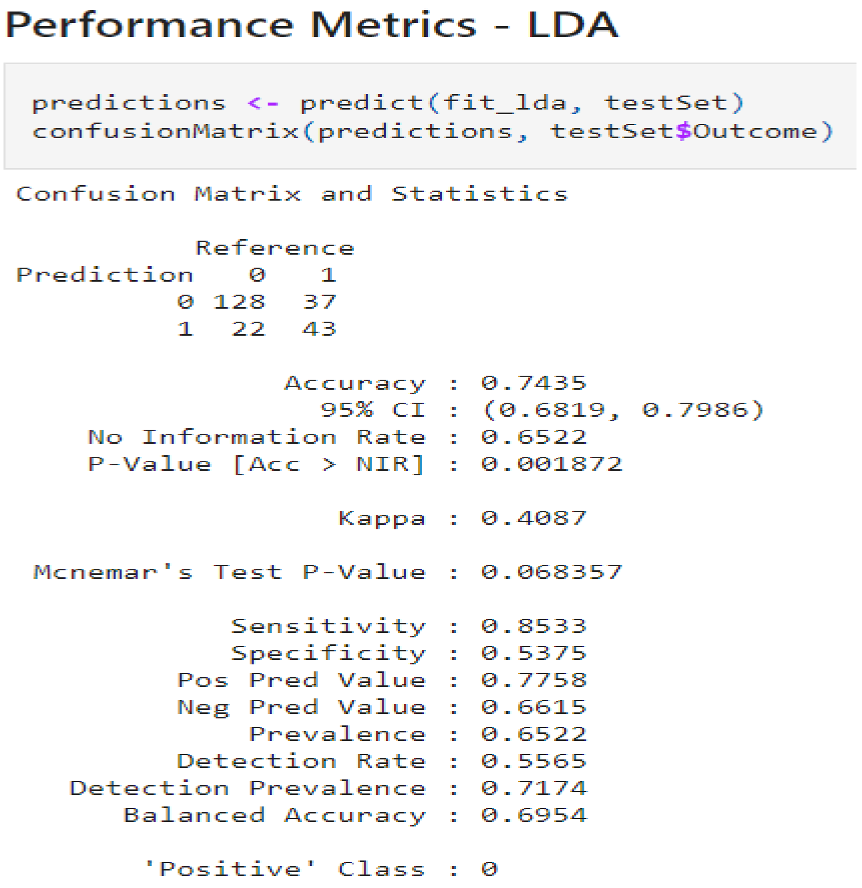
Prediction on Models

When we have an equal number of examples per class, accuracy may be a helpful metric but, when we have an unbalanced collection of data, accuracy isn't useful at all. To make matters worse, a test may have a high accuracy rating but perform much worse than another test with a lower accuracy rating. Here in our dataset is unbalanced, so accuracy is not best in this scenario.

Linear Discriminant Analysis :

In the beginning, the Linear Discriminant Analysis algorithm looked for directions that would maximize the separation between classes and then used these directions to forecast the class of people. In this case, the directions are referred to as linear discriminants, and they are made up of linear function of predictors. LDA is predicated on the assumption that predictors are normally distributed (Gaussian distribution). The various classes have mean values that are unique to each class and equal variance and covariance. To predict the likelihood of belonging to a particular class (or category), linear discriminant analysis is used in conjunction with one or more predictor variables. It may be used with predictor variables that are either continuous or categorical. Group means are determined using LDA, which then computes the likelihood of each person being assigned to a certain group for each individual. The person is then assigned to the group having the greatest chance of being assigned to the group.

Below are the Linear Discriminant Analysis model's prediction performance metrics: accuracy is 74.35%, sensitivity is 85.33%, and specificity value is 53.75%. More detail is shown in the figure.



Performance Metrics - LDA

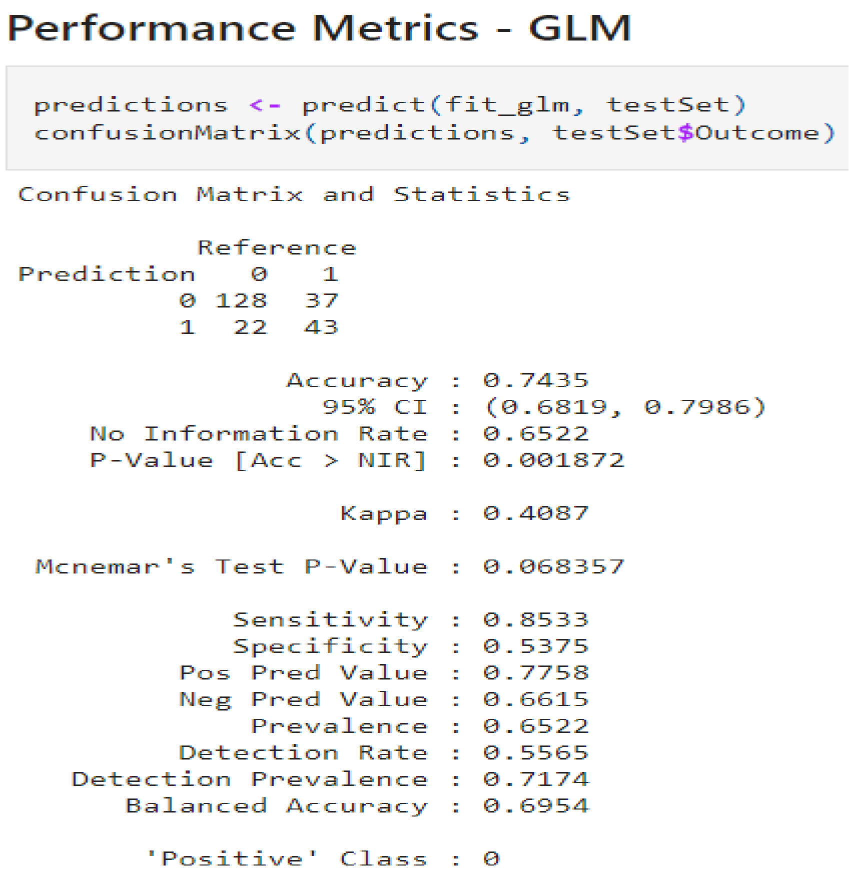
```
predictions <- predict(fit_lda, testSet)
confusionMatrix(predictions, testSet$Outcome)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 128  37
         1  22  43

               Accuracy : 0.7435
                 95% CI : (0.6819, 0.7986)
    No Information Rate : 0.6522
    P-Value [Acc > NIR] : 0.001872

                  Kappa : 0.4087

 Mcnemar's Test P-Value : 0.068357

            Sensitivity : 0.8533
            Specificity : 0.5375
         Pos Pred Value : 0.7758
         Neg Pred Value : 0.6615
             Prevalence : 0.6522
         Detection Rate : 0.5565
   Detection Prevalence : 0.7174
      Balanced Accuracy : 0.6954

       'Positive' Class : 0
```

Generalized Linear Model:

GLM (generalized linear models) are used to fit generalized linear models, defined by giving a symbolic description of the linear predictor and a symbolic description of the error distribution. A generalized linear model glm() is the function that instructs R to execute a generalized linear model. R is provided with critical information about the model via the information included inside the parenthesis. The dependent variable, which is a success, is located to the left of the symbol. It must be coded with the numbers 0 and 1 in order for GLM to recognize it as binary. Following the apostrophe, we provide the two predictor variables.

Below are the Generalized Linear Model's prediction performance metrics: accuracy is 74.35%, sensitivity is 85.33%, and specificity value is 53.75%. More detail is shown in the figure.



Performance Metrics - GLM

```
predictions <- predict(fit_glm, testSet)
confusionMatrix(predictions, testSet$Outcome)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 128  37
         1  22  43

               Accuracy : 0.7435
                 95% CI : (0.6819, 0.7986)
    No Information Rate : 0.6522
    P-Value [Acc > NIR] : 0.001872

                  Kappa : 0.4087

 Mcnemar's Test P-Value : 0.068357

            Sensitivity : 0.8533
            Specificity : 0.5375
         Pos Pred Value : 0.7758
         Neg Pred Value : 0.6615
             Prevalence : 0.6522
         Detection Rate : 0.5565
   Detection Prevalence : 0.7174
      Balanced Accuracy : 0.6954

       'Positive' Class : 0
```

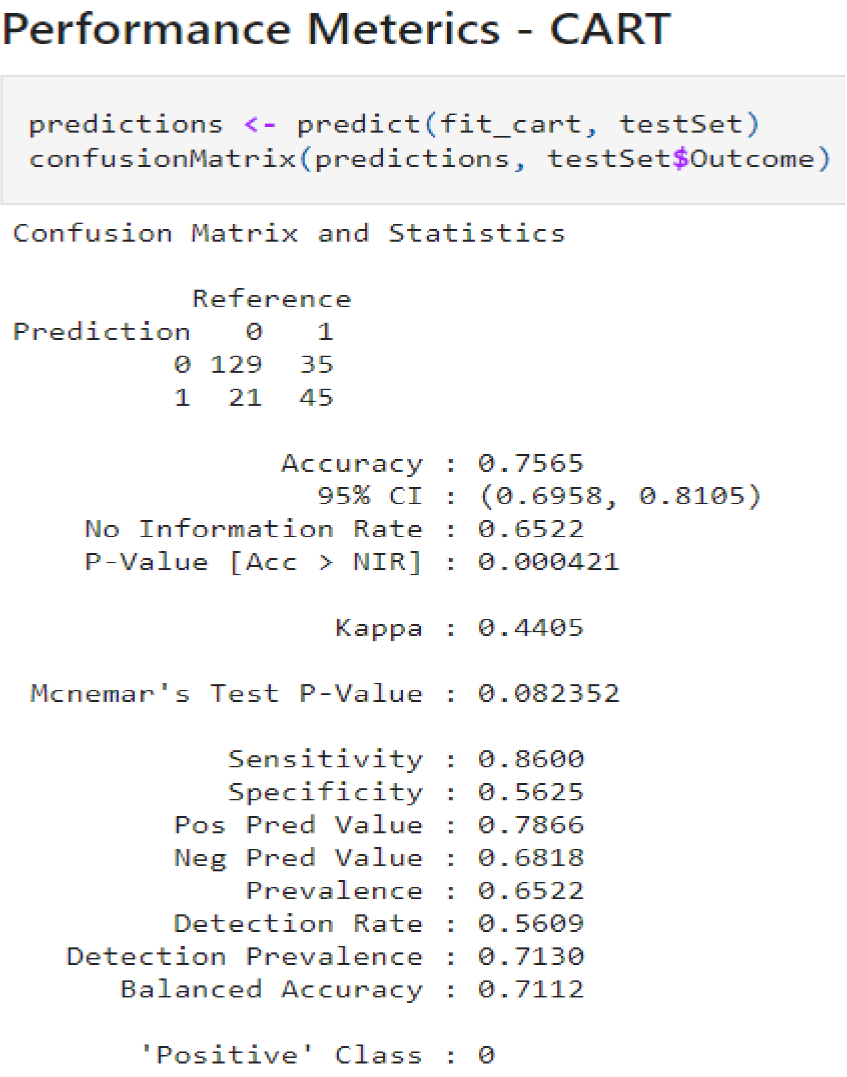
Classification and Regression Trees:

The rpart package may be used to create classification and regression trees (as defined by Brieman, Freidman, Olshen, and Stone) and other data structures. An Introduction to Recursive Partitioning Using the RPART Routines contains in-depth information on the rpart routines and how to use them.

When it comes to predictive machine learning techniques, the decision tree approach is one of the most powerful and widely utilized. It is used for both classification and regression. As a result, it is referred to as Classification and Regression Trees (CART). It should be noted that the R version of the CART method is referred to as RPART [Recursive Partitioning and Regression Trees], and it is accessible in a package with the same name as the algorithm.

The method of decision tree models operates by continually dividing the data into numerous sub-spaces to make the outcomes in each final sub-space as homogenous as possible in each sub-space. Technically speaking, this technique is referred to as recursive partitioning.

Below are the Classification and Regression Tree Model’s prediction performance metrics: accuracy is 75.65%, sensitivity is 86%, and specificity value is 56.25%. More detail is shown in the figure.



Performance Meterics - CART

```
predictions <- predict(fit_cart, testSet)
confusionMatrix(predictions, testSet$Outcome)
```

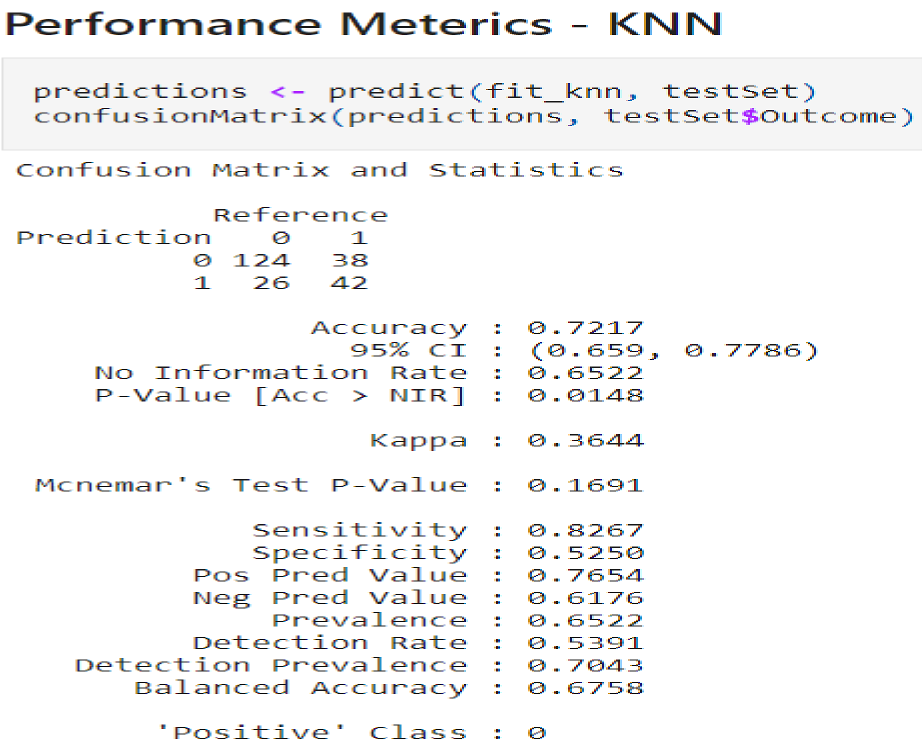
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 129  35
         1  21  45

               Accuracy : 0.7565
                 95% CI : (0.6958, 0.8105)
    No Information Rate : 0.6522
    P-Value [Acc > NIR] : 0.000421

                  Kappa : 0.4405

 Mcnemar's Test P-Value : 0.082352

            Sensitivity : 0.8600
            Specificity : 0.5625
         Pos Pred Value : 0.7866
         Neg Pred Value : 0.6818
             Prevalence : 0.6522
         Detection Rate : 0.5609
   Detection Prevalence : 0.7130
      Balanced Accuracy : 0.7112

       'Positive' Class : 0
```

K- Nearest Neighbor:

The k-nearest neighbors KNN method is a supervised machine learning technique that is simple and straightforward to implement. It may be used to solve classification and regression issues. Because it produces very precise predictions, the KNN algorithm can compete with the most accurate models. To take advantage of the KNN algorithm, you must have an application that requires great accuracy but does not need a model that humans can read. The distance metric has an impact on the quality of the predictions made.

In the context of machine learning, the k-nearest-neighbors method is an example of a 'lazy learner', which means that it does not construct a model using the training set until a query of the data set is executed. All computations are performed when the data point's neighbors are requested to be polled by the algorithm. As a result, KNN is very simple to apply in data mining applications.

Below are the K-nearest neighbor Model’s prediction performance metrics: accuracy is 72.17%, sensitivity is 82.67%, and specificity value is 52.50%. More detail is shown in the figure.



Performance Meterics - KNN

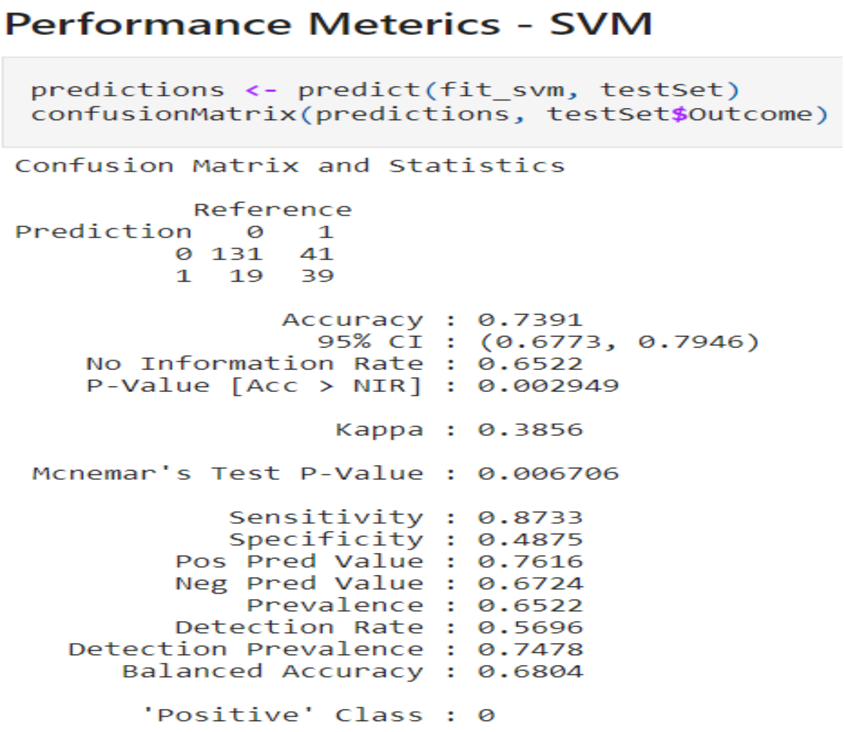
```
predictions <- predict(fit_knn, testSet)
confusionMatrix(predictions, testSet$Outcome)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 124  38
         1  26  42
                                          
               Accuracy : 0.7217          
                 95% CI : (0.659, 0.7786)
    No Information Rate : 0.6522          
    P-Value [Acc > NIR] : 0.0148          
                                          
                  Kappa : 0.3644          
                                          
 Mcnemar's Test P-Value : 0.1691          
                                          
            Sensitivity : 0.8267          
            Specificity : 0.5250          
         Pos Pred Value : 0.7654          
         Neg Pred Value : 0.6176          
             Prevalence : 0.6522          
         Detection Rate : 0.5391          
   Detection Prevalence : 0.7043          
      Balanced Accuracy : 0.6758          
                                          
       'Positive' Class : 0               
```

Support Vector Machine:

The Support Vector Machine, often known as SVM, is a linear model that may be used to solve issues in classification and regression. It is capable of solving both linear and nonlinear problems, and it is effective in a wide range of real situations. The concept of SVM is straightforward: The method generates a line or a hyperplane that divides the data into groups of similarity.

Below are the SVM Model’s prediction performance metrics: accuracy is 73.91%, sensitivity is 87.33%, and specificity value is 48.75%. More detail is shown in the figure.



Performance Meterics - SVM

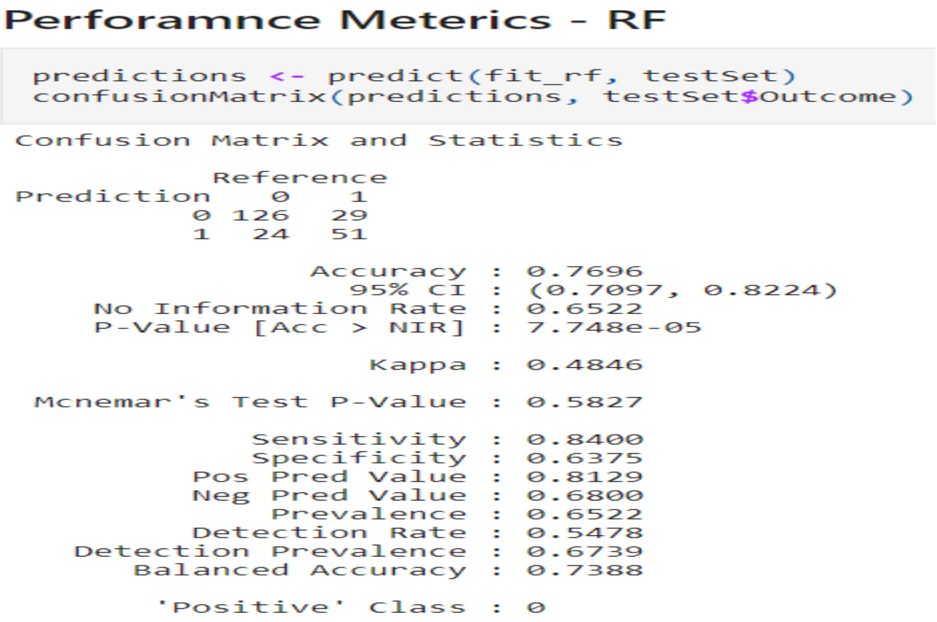
```
predictions <- predict(fit_svm, testSet)
confusionMatrix(predictions, testSet$Outcome)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 131  41
         1  19  39
                                          
               Accuracy : 0.7391          
                 95% CI : (0.6773, 0.7946)
    No Information Rate : 0.6522          
    P-Value [Acc > NIR] : 0.002949        
                                          
                  Kappa : 0.3856          
                                          
 Mcnemar's Test P-Value : 0.006706        
                                          
            Sensitivity : 0.8733          
            Specificity : 0.4875          
         Pos Pred Value : 0.7616          
         Neg Pred Value : 0.6724          
             Prevalence : 0.6522          
         Detection Rate : 0.5696          
   Detection Prevalence : 0.7478          
      Balanced Accuracy : 0.6804          
                                          
       'Positive' Class : 0               
```

Random forest:

The random forest approach, a machine learning methodology, may be used to deal with problems involving regression and classification data. An example of ensemble learning is found in machine learning, which offers solutions to complex problems by combining the efforts of many classifiers. A random forest technique is comprised of a massive number of decision trees that are randomly selected.

Below are the RF Model’s prediction performance metrics: accuracy is 76.96%, sensitivity is 84%, and specificity value is 63.75%. More detail is shown in the figure.



Perforamnce Meterics - RF

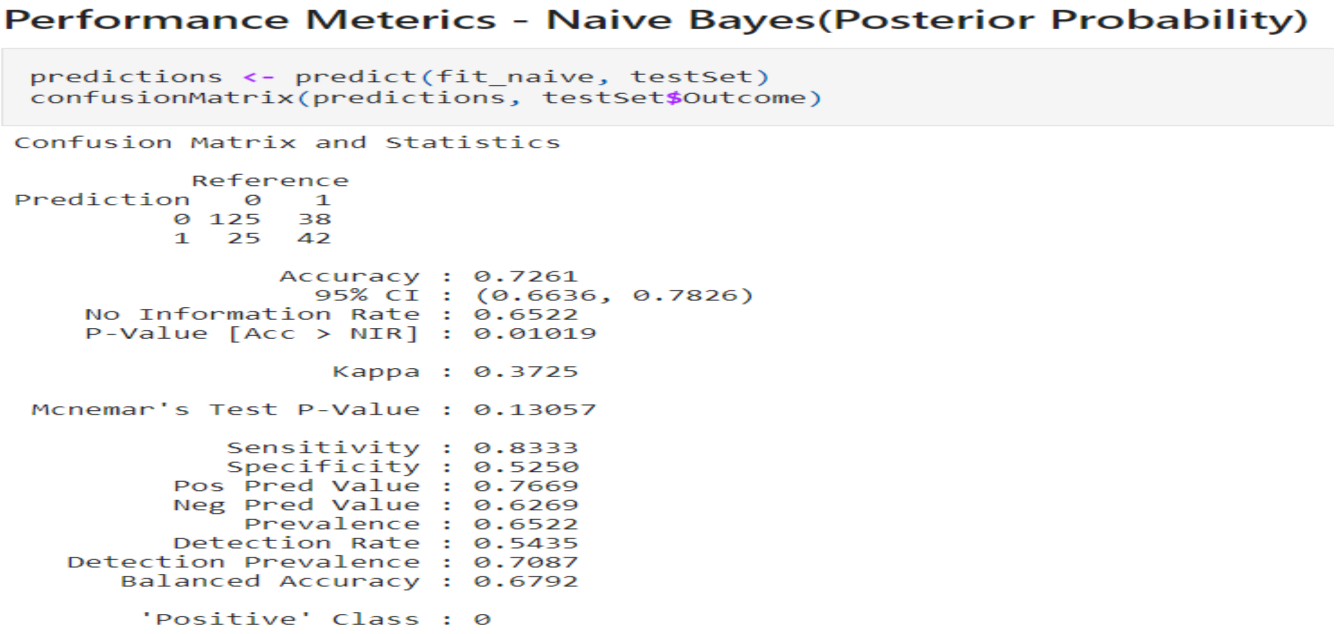
```
predictions <- predict(fit_rf, testSet)
confusionMatrix(predictions, testSet$Outcome)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 126  29
         1  24  51

               Accuracy : 0.7696
                 95% CI : (0.7097, 0.8224)
    No Information Rate : 0.6522
    P-Value [Acc > NIR] : 7.748e-05

                  Kappa : 0.4846

 Mcnemar's Test P-Value : 0.5827

            Sensitivity : 0.8400
            Specificity : 0.6375
         Pos Pred Value : 0.8129
         Neg Pred Value : 0.6800
             Prevalence : 0.6522
         Detection Rate : 0.5478
   Detection Prevalence : 0.6739
      Balanced Accuracy : 0.7388

       'Positive' Class : 0
```

Naive Bayes:

It is a classification method that is based on the Bayes' Theorem and the assumption that predictors are independent of one another, among other things. The majority of the time, a Naive Bayes classifier works on the premise that the existence of a specific feature in one class has no relationship to the presence of any other feature in another class in the same class.

Below are the Naïve Bayes Model’s prediction performance metrics: accuracy is 72.61%, sensitivity is 83.33%, and specificity value is 52.50%. More detail is shown in the figure.



Performance Meterics - Naive Bayes(Posterior Probability)

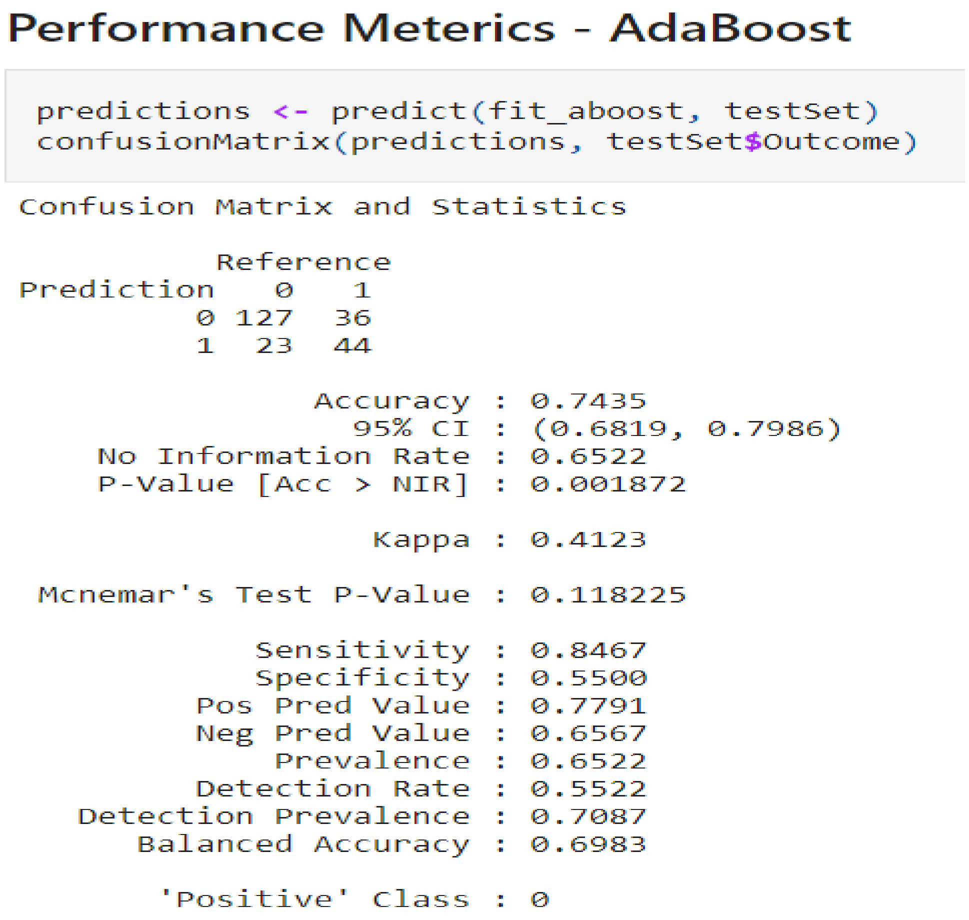
```
predictions <- predict(fit_naive, testSet)
confusionMatrix(predictions, testSet$Outcome)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 125  38
         1  25  42

               Accuracy : 0.7261
                 95% CI : (0.6636, 0.7826)
    No Information Rate : 0.6522
    P-Value [Acc > NIR] : 0.01019

                  Kappa : 0.3725

 Mcnemar's Test P-Value : 0.13057

            Sensitivity : 0.8333
            Specificity : 0.5250
         Pos Pred Value : 0.7669
         Neg Pred Value : 0.6269
             Prevalence : 0.6522
         Detection Rate : 0.5435
   Detection Prevalence : 0.7087
      Balanced Accuracy : 0.6792

       'Positive' Class : 0
```

Adaboost:

AdaBoost (Adaptive Boosting) is a machine learning boosting technique that is used to improve performance. Boosting algorithms are based on improving week learners and building an aggregated model to increase model accuracy. This is a fundamental concept in machine learning. A weak learner is characterized as performing poorly or just slightly better than a random guess classifier in terms of classification accuracy. Classifiers are combined by raising their weights and soliciting votes from users to produce the final combined model.

Below are the AdaBoost Model's prediction performance metrics: accuracy is 74.35%, sensitivity is 84.67%, and specificity value is 55%. More detail is shown in the figure.



Performance Meterics - AdaBoost

```
predictions <- predict(fit_aboost, testSet)
confusionMatrix(predictions, testSet$Outcome)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 127  36
         1  23  44

               Accuracy : 0.7435
                 95% CI : (0.6819, 0.7986)
    No Information Rate : 0.6522
    P-Value [Acc > NIR] : 0.001872

                  Kappa : 0.4123

 Mcnemar's Test P-Value : 0.118225

            Sensitivity : 0.8467
            Specificity : 0.5500
         Pos Pred Value : 0.7791
         Neg Pred Value : 0.6567
             Prevalence : 0.6522
         Detection Rate : 0.5522
   Detection Prevalence : 0.7087
      Balanced Accuracy : 0.6983

       'Positive' Class : 0
```

Artificial Neural Network ANN:

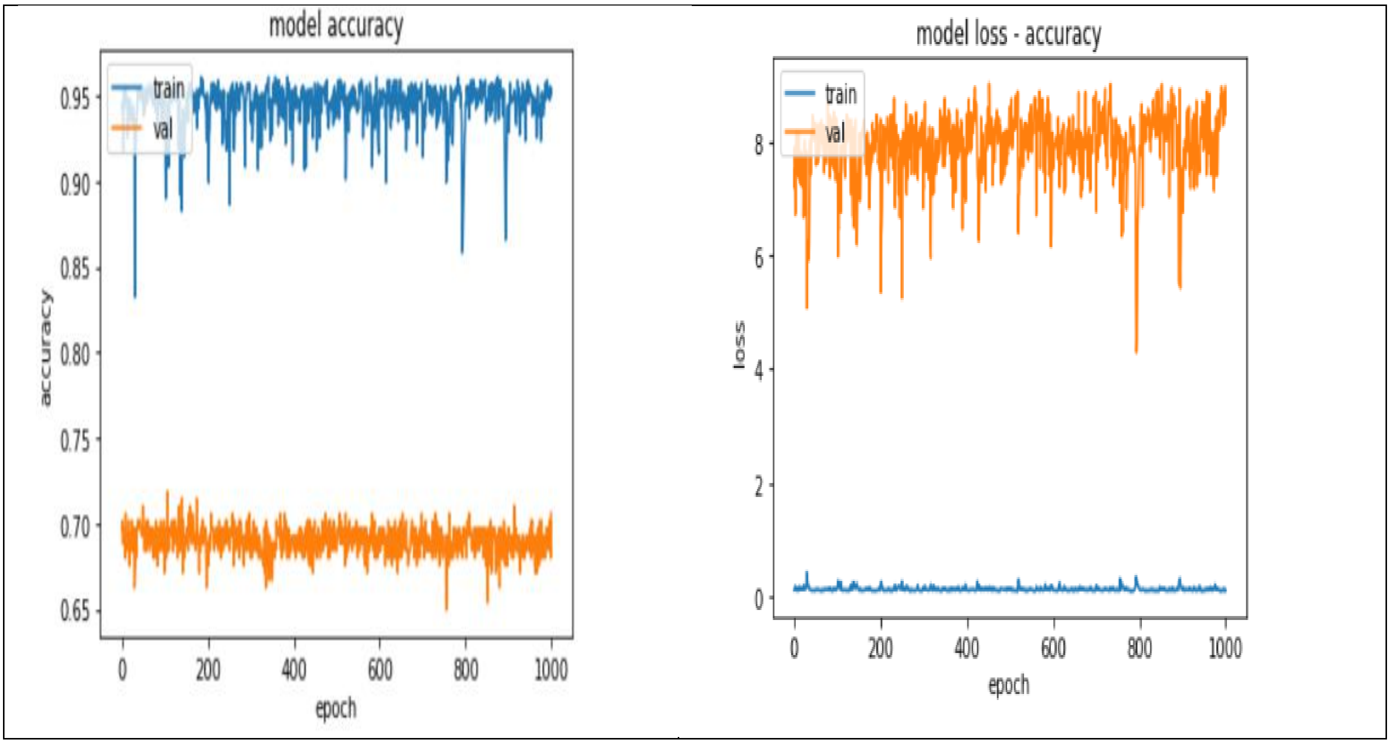
It is possible to mimic how information from the human brain is evaluated and processed using artificial neural networks, which may be implemented as a computer system. Artificial intelligence is the building block upon which issues that are impossible or difficult to solve using human or statistical variables are addressed.

Keras is a lightweight Python deep learning framework that may be used on top of Theano and Tensorflow to do deep learning tasks. It was created to enable the implementation of deep learning models for research and development as quickly and simply as feasible.

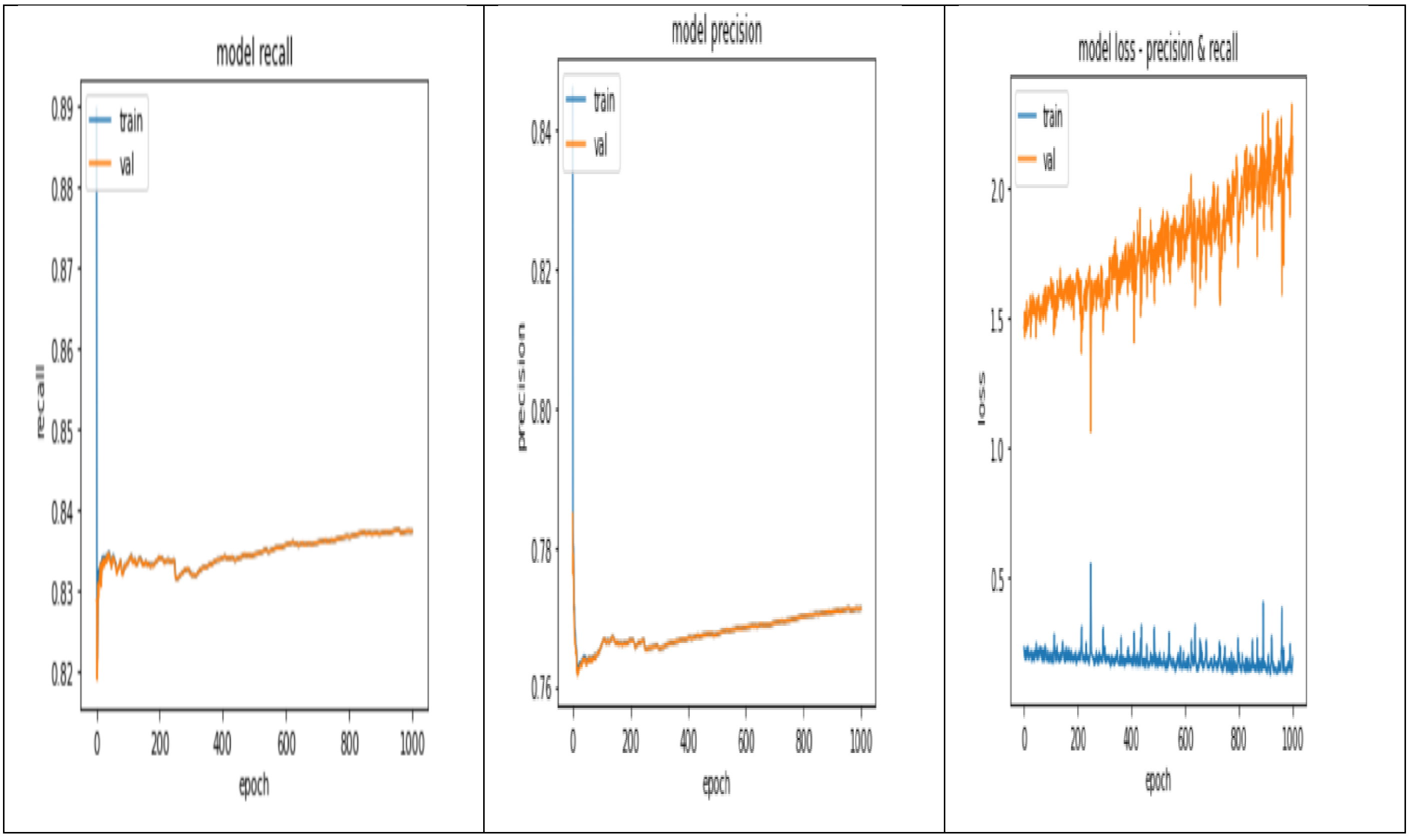
Google Colaboratory is a web-based platform that enables anybody to create and run Python code via the internet. It is particularly well suited for machine learning, data scientists, and educational applications.

Specifically, in this proposed study for deep learning approach to predict diabetes disease, we use the sequential model in the Keras package with five Dense layers, using relu activation function in the first four Dense layers and sigmoid activation function used instead of relu in the Dense final layer to predict probability between 0 and 1, compile our model with ADAM Gradient Descent optimizer with loss-binary-crossentropy.

In model fitting on training-validation dataset with 1000 number of epochs and 20 batch size we calculate accuracy performance metric in this scenario model compute Accuracy 70.56% with loss score of 8.95% . More clarity is shown in the following figures.



The training-validation dataset has 1000 epochs and 20 batches, and the model fitting is done on that dataset we also calculate sensitivity (recall) and specificity (precision) in terms to best performance metrics validation here recall value is 83.73% , precision is 77.12% and their loss score value is 2.19% .



Accuracy Table:

Algorithms	Accuracy
LDA	74.35%
GLM	74.35%
CART	75.65%
KNN	72.17%
SVM	73.91%
RF	76.96%
Naïve Bayes	72.61%
AdaBoost	74.35%
ANN	70.56%

(Table: 1)

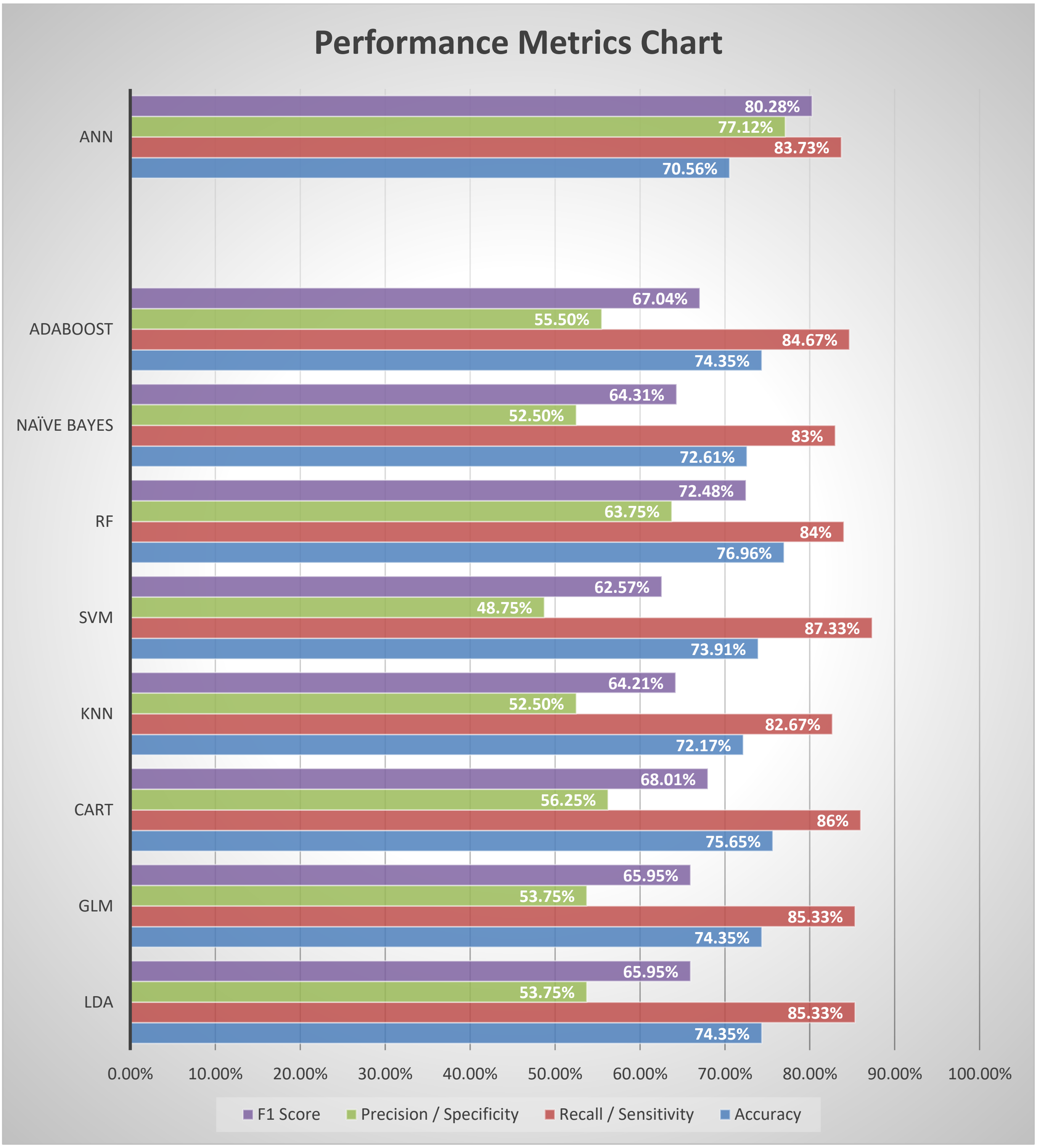
ROC Table:

Algorithms	Recall / Sensitivity	Precision / Specificity
LDA	85.33%	53.75%
GLM	85.33%	53.75%
CART	86%	56.25%
KNN	82.67%	52.50%
SVM	87.33%	48.75%
RF	84%	63.75%
Naïve Bayes	83%	52.50%
AdaBoost	84.67%	55.50%
ANN	83.73%	77.12%

(Table: 2)

Table 1 and Table 2 include performance measures for the machine learning methods that we used in this proposed research, which are described later. The Random Forest model achieves a high Accuracy value of 76.96 %, while the SVM model achieves a high Recall value of 87.33 %, and the ANN model achieves a very high Precision value of 77.12 %. The SVM model achieves a high Recall value of 87.33 %, and the ANN model achieves a very high Precision value of 77.12 %.

Result Chart:
The performance metrics of our all machine learning models covered in this suggested research are clearly shown and visualized in the chart below. We can see the Accuracy, Recall, Precision, and F-measure (F1-Score) and comparison with all models. Although accuracy may be useful in determining whether or not a class has the same number of examples as the other classes, accuracy is useless when we have an imbalanced collection of data to work with. Following this, we examined into other performance metrics like as recall, precision and F1-Score. ANN has an excellent F-measeure (F-1 score) of 80.28 %, above all moldels.



CONCLUSIONS

The model proposed in this article can predict in diabetics that satisfactory sensitivity is higher using some commonly used lab results. These models can be implemented and developed as an application used by doctors' computers and mobile or tablet gadgets to help predict the presence of the disease in precautionary patients and the future. Provides appropriate medications and health indicators to prevent such diseases. This proposed model is designed, developed and validated on real-time datasets designed for future growth; various data mining techniques and their application were studied or reviewed, we use different machine learning and Deep Learning algorithms to perform prediction in this way, Artificial Neural Network ANN has high F-measure (F-1 score) value i.e. 80.28% this algorithm is best than all others in this scenario.

Conflict of Interest:

There is no conflict of interest in this situation.

REFERENCES

[1] Hessler, D.M., Fisher, L., Mullan, J.T., Glasgow, R.E., & Masharani, U.(2011). Patient age: a neglected factor when considering disease management in adults with type 2 diabetes. *Patient education and counseling, 85*(2), 154-159.

[2] Lopez, J.M., Bailey, R.A., Rupnow, M.F., & Annunziata, K. (2014). Characterization of type 2 diabetes mellitus burden by age and ethnic groups based on a nationwide survey. *Clinical therapeutics, 36*(4), 494-506.

[3] Sabariah, M.M.K., Hanifa, S.A., & Sa'adah, M.S. (2014). Early detection of type II Diabetes Mellitus with random forest and classification and regression tree (CART). *In International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA)* (pp. 238-242). IEEE.

[4] B.M. Patil, R.C. Joshi and Durga Toshniwal,"Association Rule for Classification of Type-2 Diabetic Patients", ICMLC '10 Proceedings of the 2010 Second International Conference on Machine Learning and Computing, February 09 - 11, 2010.

[5] Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing,2015.

[6] Dost Muhammad Khan1, Nawaz Mohamudally2, "An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm ", Journal Of Computing, Volume 3, Issue 12, December 2011.

[7] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly," Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.

[8] K. Rajesh and V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.

[9] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," J. Big Data, vol. 6, no. 1, Dec. 2019.

[10] Y. Li, H. Li, and H. Yao, "Analysis and Study of Diabetes Follow-Up Data Using a Data-Mining-Based Approach in New Urban Area of Urumqi, Xinjiang, China, 2016-2017," Comput. Math. Methods Med., vol. 2018, 2018.

[11] Q. Sun, M. V. Jankovic, L. Bally, and S. G. Mougiakakou, "Predicting Blood Glucose with an LSTM and Bi-LSTM Based Deep Neural Network," in 2018 14th Symposium on Neural Networks and Applications, NEUREL 2018.

[12] P. P. Singh, S. Prasad, B. Das, U. Poddar, and D. R. Choudhury, "Classification of diabetic patient data using machine learning techniques," in Advances in Intelligent Systems and Computing, 2018, vol. 696, pp. 427–436.

[13] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," Front. Genet., vol. 9, Nov. 2018.

[14] S. Apoorva, K. Aditya S, P. Snigdha, P. Darshini, and H. A. Sanjay, "Prediction of Diabetes Mellitus Type-2 Using Machine Learning," 2020, pp. 364–370.

[15] G. Swapna, R. Vinayakumar, and K. P. Soman, "Diabetes detection using deep learning algorithms," ICT Express, vol. 4, no. 4, pp. 243–246, Dec. 2018.

[16] S. B. Kotsiantis, "Decision trees: A recent overview," Artificial Intelligence Review, vol. 39, no. 4. Springer, pp. 261–283, 29-Apr-2013.

[17] Larabi-Marie-Sainte, Aburahmah, Almohaini, and Saba, "Current Techniques for Diabetes Prediction: Review and Case Study," Appl. Sci., vol. 9, no. 21, p. 4604, Oct. 2019.

[18] A. Negi, V. Jaiswal
A first attempt to develop a diabetes prediction method based on different global datasets
2016 fourth international conference on parallel, distributed and grid computing, PDGC) (2016), pp. 237-24.

[19] N. Murat, E. Dünder, M.A. Cengiz, M.E. Onger
The use of several information criteria for logistic regression model to investigate the effects of diabetic drugs on HbA1c levels
Biomed Res, 29 (2018), pp. 1370-1375.

[20] H.N. Merad-boudia, M. Dali-Sahi, Y. Kachekouche, N. Dennouni-Medjati
Hematologic disorders during essential hypertension," diabetes & metabolic syndrome
Clinical Research & Reviews (2019).