DIABETES PREDICTION USING MACHINE LEARNING

Abhishek M B[1], Manasa R[2], Vibha T G[3], and Bindu H M[4]

[1] Assistant Professor,Department of Electronics Engieering, Dayananda Sagar College of Engineering, Bangalore, Karnataka

[2] Assistant Professor,Department of Electronics Engieering, Dayananda Sagar College of Engineering, Bangalore, Karnataka

[3] Assistant Professor,Department of Electronics Engieering, Dayananda Sagar College of Engineering, Bangalore, Karnataka

[4] Assistant Professor,Department of Electronics Engieering, Dayananda Sagar College of Engineering, Bangalore, Karnataka

abhishek.mb@gmail.com

Abstract: Diabetes is an associate malady caused as a result of high aldohexose level in the human body. If Diabetes is ignored or untreated then it may lead to serious health issues which may be heart related problems, blood pressure, kidney problems and also eye damage. A key aspect in treatment of Diabetes is it's detection at an early stage. If predicted earlier Diabetes can be controlled effectively. Hence, this led us to propose a diabetes prediction model for classifying diabetes in the right category taking into account numerous external factors which may be responsible for causing diabetes. The regular factors such as Age , Insulin , Glucose ,etc. also play a vital role in increasing the percentage of people falling under the Diabetes category. Random selection of a certain number of health data of both diabetic patients and healthy people , as a training set was conducted. Selection of valid categorical variables and the most general classifier with high accuracy score were the driving force for determining the machine learning method . Hence a set of algorithms will be compared against using some standard measurement techniques like confusion matrix, classification report, and accuracy score. Finally, we will compare standard existing machine learning predictive algorithms and select the method that is best suited for our diabetes prediction. The results obtained suggest Random Forest classifier and XGBoost as the most efficient and general classifier as against other models.

Keywords: Machine Learning, Prediction algorithm, Support Vector Machine, Random Forest Classifier, Decision tree Classifier, Logistic Re- gression Classifier, K-neighbours Classifier, Decision tree Classifier, Gra- dient Boosting Classifier, and XGBoost Classifier.

Number: 10.14704/nq.2022.20.7.NQ33372 **Neuro Quantology 2022; 20(7):2932-2945**

1 INTRODUCTION

Diabetes is an associate malady caused as a result of high aldohexose level in the human body. If Diabetes is ignored or untreated then it may lead to serious

health issues which may be heart related problems , blood pressure , kidney problems and also eye damage. A key aspect in treatment of Diabetes is it's detection at an early stage. If predicted earlier Diabetes can be controlled ef- fectively. As of present situation the growing rate of the disease in the world population gives an insight about the upcoming future wherein it is predicted that by 2040, 642 million people across the world will be affected by Diabetes meaning that out of ten adults , one is suffering from diabetes. Such an alarming figure definitely deserves immediate attention. In the field of machine learning , medical health has witnessed a wide range of developments.

Hence in this project, our research led us to propose a diabetes prediction model for classifying diabetes in the right category taking into account numerous ex- ternal factors which may be responsible for causing diabetes. The regular factors such as Age , Insulin , Glucose ,etc. also play a vital role in increasing the per- centage of people falling under the Diabetes category. Random selection of a certain number of health data of both diabetic patients and healthy people , as a training set was conducted. Related works have been analysed and suitably manipulated and applied to give much accurate results on mostly any healthcare dataset.

Selection of valid categorical variables and the most general classifier with high accuracy score were the driving force for determining the machine learning method . Hence a set of algorithms will be compared against using some standard measurement techniques like confusion matrix, classification report , accuracy score etc . Finally, we will compare standard existing machine learning predictive algorithms and select the method that is best suited for our diabetes prediction [10].The results obtained suggest Random Forest classifier and XGBoost as the most efficient and general classifier as against other models.

Diabetes mellitus may be regarded as a chronic illness mainly described by the distinctive feature of hyperglycaemia. More or less 2-5 million patients once a year lose their lives because of the disease. It's the same that by 2045 this can rise to 629 million. In the near future the onset of the disease is calculated to rise and can create a striking difference. Over the years the care sectors have gathered massive volume databases based on the patients health report .Dia- betes has been classified into the category of NonCommunicable disease.

Generally, Diabetes is classed as - Type-1 Insulin-Dependent diabetes (IDDM). The patient's inability to come up with enough hypoglycemic agents is the rea- son behind this sort of DM and thus it's needed to inject hypoglycemic agents called insulin to the body. Patients with type1 of the disease square measure unremarkably younger, principally but thirty years recent. Generally monitored and registered symptoms square measure high glucose levels, hyperbolic thirst, uncontrolled perspiration and frequent excretion. The treatment prescribed un- der this sort of the disease with oral medications alone are not effective in the long run and therefore the patients square measure needed to possess insulin medical care.

Type-2 conjointly referred to as type II diabetes Mellitus (NIDDM). This sort of the disease is seen once body cells seem unable to make a proper use of the

insulin. This type happens a lot ordinarily in old and aged individuals, that is usually related to the incidence of avoirdupois,dyslipidemia, high blood pressure, induration, and different other serious diseases.

Type-3 is Gestational Diabetes. This is predominantly common in pregnant ladies due to increase in glucose levels wherein if the disease is left undiscov- ered in the early stages may lead to this type of DM.

A technique referred to as, Predictive / Prognostication Analysis, embodies a spread of machine learning algorithms, data processing techniques and applied math strategies that makes use of past and current information to seek out data and predict impending events related to one's health.Application of prognos- tication analysis on health care information, vital choices are often taken and predictions are often created. Machine learning and regression techniques are the most commonly and widely used methods for Prognostication analysis. This technique aims at identifying the illness with excellent accuracy, enhancing the standards of the nursing methods,resource optimization beside up clinical out- comes. Machine learning is taken into account to be one among the foremost vital computer science options for the development of such computer systems that can effectively accumulate data from past events without having to pro- gramme for each and every case which can be a very tedious job to implement. Machine learning is taken into account overviewing the present scenario in order to reduce human efforts and flaws [9], [14] Centre of attention in this paper is the use of standard machine learning algorithms to set up such predictive models for Diabetes prediction.

2 LITERATURE SURVEY

The analysis of previous research work provides results on varied healthcare datasets. Different prediction models were developed and enforced by varied re- searchers exploiting variants of machine learning algorithms. In [1], enforced a system exploitation Hadoop and Map Reduce techniques for analysis of Dia- betic knowledge. This method predicts a sort of diabetes disorder . In the [12] used classification techniques to detect underlying patterns in the dataset. Na¨ıve Bayes and Decision Trees were utilized in this model. The models were compared against each other and their effectiveness were displayed as a result. In [2] used the C4.5 Decision tree rule which helps in classifying expeditiously . [3] proposed Random Forest Classifier as one of the best models for Diabetes prediction . In

[5] used C4.5 Decision tree rule, Neural Network, K-means clustering rule and visual image to predict polygenic disorder.

As of recent works during this domain -[6] conducted a comprehensive study on a polygenic disorder dataset with Random Forest (RF), SVM, k-NN, CART and LDA algorithms. The achieved results show that RF is giving additional cor- rect predictions compared to alternative algorithms. In [7] Mitushi Soni and Dr. Sunita Varma used Machine Learning Classification and ensemble techniques on a dataset to predict polygenic disorders or diabetes that are KNN , LR , DT , SVM, Gradient Boosting and Random Forest (RF). Their results additionally

showed that Random Forest achieved higher accuracy compared to alternative machine learning techniques. In [8] Jobeda Jamal Khanam and Simon Y. Foo used seven mil algorithms on the dataset to predict polygenic disorder. They used a similar PIMA dataset . They engineered the NN model with a special hidden layer with varied epochs and determined that the NN with 2 hidden lay- ers provided 88.6% accuracy.

The task of selecting a machine learning rule includes feature matching of the info to be learned and supported existing approaches. Taxonomy of machine learning algorithms is mentioned below Machine learning has various algorithms that are classified into 3 categories as shown in Figure **??**: supervised learning, unsupervised learning, Semi-supervised learning [11], [13].

– The supervised Learning/Predictive Models This type of algorithm is accus- tomed to construct prognostication models. A prognosticative model pre- dicts missing values exploitation alternative values given within the dataset. Supervised learning rule contains both input and output file, based upon which information a model is built to give much accurate responses to real time data. Supervised learning includes Decision Tree, Bayesian methodol- ogy, Artificial Neural Network, Instance based mostly learning, Ensemble methodology. These are booming techniques in Machine learning.

– Unsupervised Learning / Descriptive Models In this model we've got a no- table set of inputs however output remains untold. This methodology in- cludes bunch algorithms like k-Means Clustering and k-Medians Clustering.

– Semi-supervised Learning Semi supervised learning methodology uses each tagged and untagged knowledge on a training dataset. Classification, Re- gression techniques fall under this category.

From the literature review,

In current diagnosing methodology, 3 sorts of errors are-

– False-Negative - The patient is actually a diabetic patient however the check report tells otherwise. This kind of error is the most common and fatally dangerous.

– False-Positive - The patient isn't actually a diabetic patient however check reports wrongly classifies the patient as diabetic.

– Unidentifiable - In this type a particular case cannot be diagnosed by the common diagnosing machine or method. Unfortunately such a thing takes place only when meager information is extracted from past knowledge.

These errors that occur during diagnosing might cause spare treatments or no treatments in the least once needed. Hence to avoid or at least bring down any significant impact, there's a desire to form a system based on machine learning algorithms which can offer correct results and cut back human efforts.

Currently existing solutions for Diabetes detection include research lab tests most popularly being fasting aldohexose/blood sugar, random plasma glucose (RPG) and A1C test. Nevertheless this methodology is time intensive.

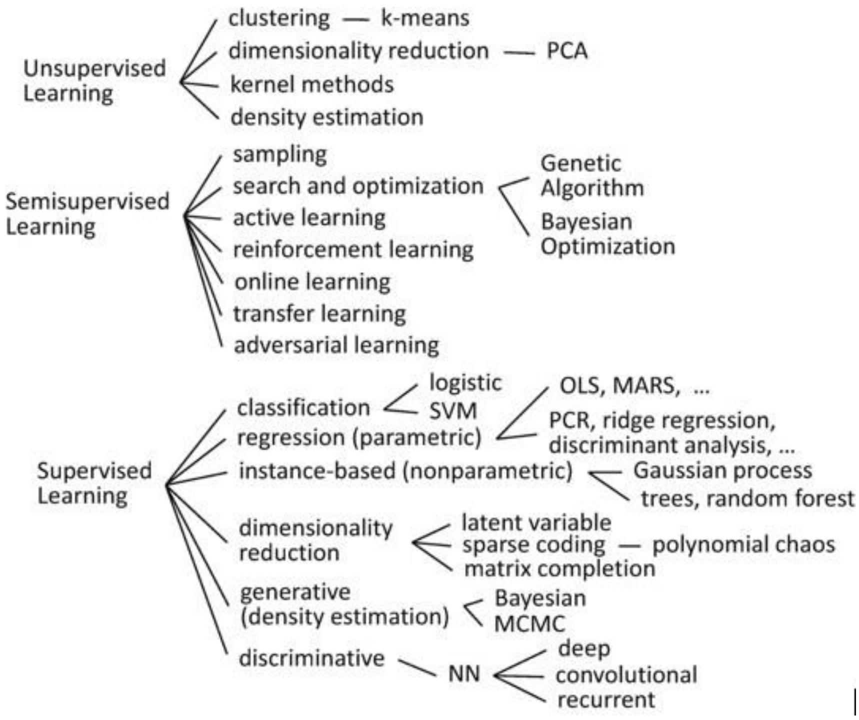

Fig. 1. Classification

3 METHODOLOGY

3.1 DATA DESCRIPTION

The information is gathered from UCI repository that is called as PIMA Indian Diabetes Dataset. The dataset has several attributes of 768 patients.

PYTHON - Jupyter Notebook Platform Python comes square measure the foremost trending topics for educational python project development. With the assistance of python 3.8.3 we've enforced the DM prediction model. The platform we have a tendency to use is the terribly basic Jupyter Notebook.

– NumPy- Library which works with arrays .NumPy aims to produce associate degree array objects that's up to 50x quicker than ancient Python lists.

– Matplotlib.pyplot- it's a set of functions that create matplotlib work like MATLAB. Every pyplot operation makes some modification to a figure.

– Seaborn-Seaborn may be a Python information image library supported by matplotlib.

– Pandas- An open supply Python package that's mostly employed for infor- mation science/data analysis and machine learning tasks. It's designed on prime of the Numpy package, and provides support for multi-dimensional arrays.

– SciPy- SciPy may be a scientific computation library that uses NumPy be- neath. SciPy stands for Scientific Python. It provides additional utility func- tions for optimisation, stats and signal processing. Like NumPy, SciPy is open source thus we are able to use it freely.

– XGBoost (eXtreme Gradient Boosting) - is a sophisticated implementation of gradient boosting formula.

DATA LOADING AND ANALYSING CSV (comma-separated value) files are unit a standard file format for transferring and storing information .The basic method of loading information from a CSV file into a Pandas DataFrame is achieved by victimising the "read.csv" operation in Pandas.If any informa- tion is found to be missing victimisation mean/median methodology or any imputation strategies betting on its behavior of the dataset . We will addition- ally apply algorithms that support missing values. In our case missing values were found thence we will proceed with data preprocessing for missing infor- mation.Pandas dataframe.groupby() perform is one in every of the foremost helpful functions within the library; it splits the info into teams supported columns/conditions (here 'outcome') so applies some operations here size() that counts the amount of entries / rows in every cluster . Here it provides the scale of rangethe amount—the quantity of 0's and also the number of 1's so prints.

DATA CLEANING Data Cleaning means that the method of characteris- tic is inaccurate, incomplete, inaccurate, inapplicable or missing a part of the

information then modifying, exchanging or deleting them consistent with the requirement. Information cleansing is a foundational part of basic information science.Missing information will occur once no data is provided for one or a lot of things or for a full unit. Missing information could be a terribly huge drawback in real-life eventualities. Missing information may be asked as NA(Not Avail- able) values in pandas. In DataFrame generally several datasets merely arrive with missing data, either as a result of it existing and wasn't collected or it ne'er existed.In Pandas missing information is diagrammatic by :

– NaN : NaN (a word form for Not a Number), could be a special floating- point value recognized by all systems that use the quality IEEE floating-point illustration.

DATA VISUALIZATION

– Seaborns .setstyle() makes the plots interesting by adjusting the color of the axes,enabling a grid or not accordingly , etc.

– plt.figure() creates a new figure where parameter figure size takes in values ten for breadth and ten for height in inches.

– seaborn.countplot() methodology is employed to point out the counts of observations in every feature. x- name of variable (i.e target - Outcome) in information. data-Dataset for plotting.

– palette- colors for various levels of the categorical variable. (eg:magmar,Accent,YlGnBu,etc)

– .corr() function is used for the computation of pairwise correlation of at- tribute columns. The .corr() function returns a matrix where along the di- agonal the matrix is symmetric and has value equal to 1 .

– np.triu(corrmax) returns the upper triangle of the array corrmax with re- spect to k=0 i,e it also returns the non-zero diagonal values.

– sns.heatmap() plots rectangular data. Since the parameter of the heatmap function is a correlation matrix only half of the matrix is required with diagonal elements also being treated as redundant elements as they give the max correlation i.e. 1. Therefore the upper triangle of the correlation matrix is used as a mask. Other parameters like vmax , vmin ,centre, cmap are for anchoring colour maps whereas parameters square and linewidth are used for creating the cells in the heatmap.

SPLITTING DATASET For randomly splitting matrices or arrays into 2 subsets - training and test set. In order to fulfill our expectation of an unpreju- diced evaluation a test set is quite required. Our model is developed by training it on the training dataset. Here the random parameter controls shuffling if applied and test size determines the splitting point of the dataset .

– Xtrain : Training subset of the independent variable i.e. x

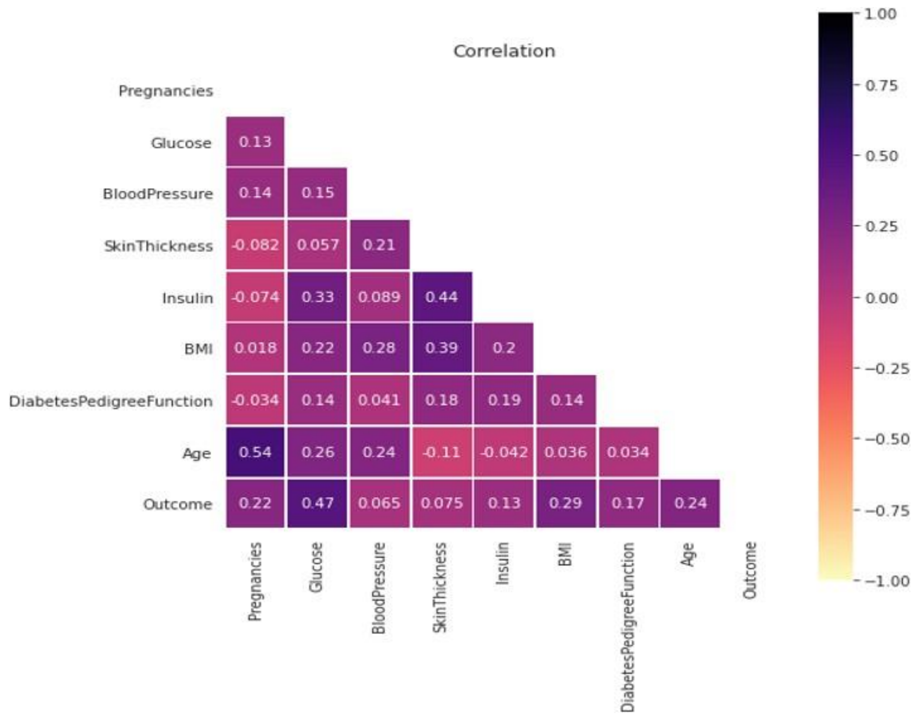

Fig. 2. Classification

– Xtest : Test subset of the independent variable i.e. x

– ytrain : Training subset of the dependent variable i.e. y

– ytest : Test subset of the dependent variable i.e. y Displaying the dimensionality of the xtrain and ytrain.

4 MODEL BUILDING AND ACCURACY MEASUREMENT

4.1 SUPPORT VECTOR CLASSIFICATION

“Support Vector Machine” (SVM) could be a supervised machine learning for- mula which might be used for each classification or regression challenge. However, it’s largely employed in classification issues. Within the SVM formula, we have a tendency to plot every information item as a degree in n-dimensional area (where n is that the range of options you have) with the worth of every feature being the worth of a specific coordinate. Then, we have a tendency to perform classi- fication by finding the hyper-plane that differentiates the 2 categories finely.

C-SVC is imported from the support vector machines available in scikit-learn

.The default kernel i.e. rbf is used for fitting the model to the training dataset. Prediction is performed on the test dataset using the predict() method .For an unbiased evaluation of the final model , a test set is necessary. accuracyscore() function imported from the metrics library in scikit-learn returns the Accuracy classification score .

GridSearchCV is a library operation that’s included in sklearn’s model selection package. Looping through preconceived hyperparameters and working the reck- oner (model) on the training set. The most effective parameters from the enlisted hyperparameters are selected based on scores. svc is the classifier model/estimator used here. The estimator found out using the best parameter values is fitted to the training data .The accuracy classification score is calculated.The classifica- tion report operation displays the main metrics of the classification model .The confusion matrix gives an overview of the actual result .It gives the actual count in 4 boxes depending on actual and predicted values. The confusion matrix is vi- sualized using heatmap. (In the parameter grid , C – Regularization parameter, Gamma – rbf’s kernel coefficient .)

4.2 RANDOM FOREST CLASSIFIER

The embodiment of two averaging algorithms in the sklearn.ensemble module is particularly based on the Extra-Trees method as well as RandomForest algo- rithm. A random forest is a model which has another model as it's parameter, a characteristic of the Ensemble Learning method. Random forest in the core uses a decision tree estimator the number of which can be specified in its parame- ter . These estimators are then fitted on diverse sub-samples created randomly from the dataset. Random forests have a tendency to over-fitting which can be controlled by averaging. Averaging also improves the predictive accuracy of the

model.rmfr is the classifier created and fitted to the training data. The predicted results are compared to the test results and an accuracy classification score is calculated. Just like in the above model , for Random forest Classifiers also we perform HYPERTUNING. Then the most effective parameters from the enlisted hyperparameters are selected based on scores .

4.3 DECISION TREE CLASSIFIER

A Decision Tree Classifier is built by taking a splitting point / threshold point into account. That point is determined by a questionnaire based on the attributes of the predefined dataset. The conclusion of the questionnaire is marked until a particular class label of the record is found which contains an almost pure class of the dependent variable. The above discussed method culminates into a decision tree - a hierarchical structure built using directed edges and nodes.A class label is assigned to each terminal node of the Decision tree. To separate data points characterized by diverse features, non-terminal nodes hold the test conditions of a particular attribute.

Our first step should be importing the DecisionTreeClassifier module and then using the DecisionTreeClassifier() function with parameters to create a Decision- Tree estimator object. Then, using the fit() method fit the model on the training data set . Prediction is performed on the test dataset using the predict() method

.For an unbiased evaluation of the final model , a test set is necessary. accuracy score() function imported from the metrics library in scikit-learn returns the Accuracy classification score .

GridSearchCV - a library function included in the sklearn's model selection package. The estimator is fitted on the training set after looping through the preconceived hyperparameters .The best parameter as against the other listed hyperparameters is decided using scores .

4.4 K-NEIGHBOURS CLASSIFIER

K-Nearest Neighbour is based on the Supervised Learning technique and is also counted as a simple Machine Learning algorithm. The K-NN algorithm checks for the similarity between the available cases and the new case and accordingly the new case is categorised based on the decision that determines the most sim- ilar category to the already available categories.

First step is to import the KNeighborsClassifier module then using the KNeigh- borsClassifier () function we should create a KNeighbors classifier object. Then, using the fit() method fit the model on the training data set . Prediction is performed on the test dataset using the predict() method .For an unbiased eval- uation of the final model , a test set is necessary. accuracy score() function imported from the metrics library in scikit-learn returns the percentage of pre- dictions correctly classified.

GridSearchCV - a library function included in the sklearn's model selection pack- age. The estimator is fitted on the training set after looping through the precon-

ceived hyper parameters .The best parameter as against the other listed hyper parameters is decided using scores .

4.5 LOGISTIC REGRESSION

Logistic regression is a Supervised Learning technique and the most popular as well as most basic Machine Learning algorithm. It helps in the prediction of the categorical dependent variable which are of the form 0, 1 or yes,no , etc. i.e.

discrete values using a predefined set of independent variables or features.

First step is to import the Logistic Regression module then using the LogisticRe- gression() function a Logistic Regression classifier object is created. Then, using fit() method fit the model on the training data set . Prediction is performed on the test dataset using the predict() method .For an unbiased evaluation of the final model , test set is necessary. accuracy score() function imported from the metrics library in scikit-learn returns the Accuracy classification score i.e. the predictions which are correctly classified.

GridSearchCV - a library function included in the sklearn's model selection package. The estimator is fitted on the training set after looping through the preconceived hyperparameters .The best parameter as against the other listed hyperparameters is decided using scores .

4.6 GRADIENT BOOSTING CLASSIFIER

Gradient boosting algorithm is a powerful algorithm based on the Ensemble Learning technique. It learns from its previous errors and builds a better model each time after learning.Errors in machine learning algorithms are classified into Variance Error and Bias Error. Being a boosting algorithm itself , Gradient boosting cuts down the model's bias error.

The very step is to import the gradient boosting classifier module then using the gradient boosting classifier() function a gradient boosting classifier object is created. Then, using the fit() method fit the model on the training data set . Prediction is performed on the test dataset using the predict() method .For an unbiased evaluation of the final model , test set is necessary. accuracy score() function imported from the metrics library in scikit-learn returns the Accuracy classification score i.e. the predictions which are correctly classified.

GridSearchCV - a library function included in the sklearn's model selection package. The estimator is fitted on the training set after looping through the preconceived hyperparameters .The best parameter as against the other listed hyperparameters is decided using scores.

4.7 XGBOOST CLASSIFIER

XGBoost where X means extreme i.e. Extreme Gradient Boosting which in turn clearly means a better version of the Gradient boosting algorithm. It's basic lies as simple as the Decision Tree Classifier hence its ensemble learning technique

is tree based . It's speciality or advantage over the Gradient Boosting algorithm is that it is a scalable learning system.. It uses more accurate approximations to find the best tree model.

First of all we import the XGBClassifier module then use the XGBClassifier() function. XGBclassifier object is created. Then, using the fit() method fit the model on the training data set . Prediction is performed on the test dataset us- ing the predict() method .For an unbiased evaluation of the final model , a test set is necessary. accuracy score() function imported from the metrics library in scikit-learn returns the Accuracy classification score i.e. the predictions which are correctly classified.

GridSearchCV - a library function included in the sklearn's model selection package. The estimator is fitted on the training set after looping through the preconceived hyperparameters .The best parameter as against the other listed hyperparameters is decided using scores .

4.8 CAT BOOST CLASSIFIER

– "CatBoost is a high-performance open source library for gradient boosting on decision trees."

– So, CatBoos algorithm is the same as the gradient boosting algorithm. It is an ensemble learning technique based on decision trees.

– It would automatically deal with categorical variables.

– Dissecting the word "CatBoost" into 2 gives - "Cat" for Category and "Boosting".

Our first move should be to import the CatBoostclassifier module and then using the CatBoostClassifier() function a cat boost classifier object is created. Then, using the fit() method fit the model on the training data set . Prediction is performed on the test dataset using the predict() method .For an unbiased evaluation of the final model , test set is necessary. accuracy score() function imported from the metrics library in scikit-learn returns the Accuracy classifi- cation score i.e. the predictions which are correctly classified.

GridSearchCV - a library function included in the sklearn's model selection package. The estimator is fitted on the training set after looping through the preconceived hyperparameters .The best parameter as against the other listed hyperparameters is decided using scores.

5 BLOCK DIAGRAM and IMPLEMENTATION

Steps involved in diabetes prediction as shown in Figure 3:

– STEP1 : DATASET COLLECTION
– STEP 2 : EXPLORATORY DATA ANALYSIS
– STEP 3 : DATA PREPROCESSING
– STEP 4 : MODEL BUILDING
– STEP 5 : MODEL EVALUATION

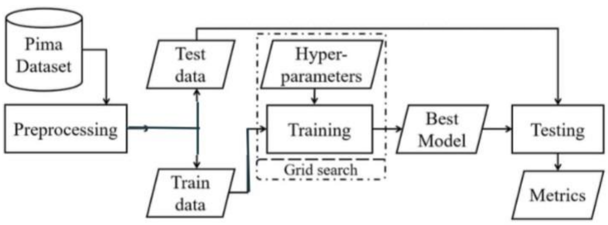

Fig. 3. Block Diagram

5.1 Dataset Assortment

This module includes information assortment and analysis of the information to follow any particular patterns and trends which helps in better prediction and generating more accurate results. This polygenic disease dataset contains 768 records and 8 attributes.

5.2 Exploratory Data Analysis

It is additionally referred to as information exploration. It's a step within the information analytics method wherever no.of techniques area units want to per- ceive the information getting used. It additionally involves information images. It's been conducted using victimisation count plots, matrix heatmap, combine plots and boxplots.

5.3 Data Pre-processing

This part of the model handles inconsistent information so as to urge a lot of correct and precise results. The dataset we used contains missing values. Therefore we have a tendency to impute missing values for a couple of selected attributes as a result of these attributes cannot have values zero. Then we have a tendency to scale the dataset to normalize all values.

5.4 Model Building

This is the foremost vital part which incorporates model building for predic- tion of Diabetes disease. During this we've enforced numerous machine learning algorithms for polygenic disease prediction. These algorithms embody Support Vector Classifier,Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, XGB Classifier, K - Neighbours Classifier, Logistic Regres- sion and Cat boost Classifier.

5.5 Model Evaluation

This is the ultimate step of the prediction model. Here, we have a tendency to assess the prediction results of numerous analysis metrics like classification report, confusion matrix and accuracy score .

CONFUSION MATRIX The main utility of Confusion Matrix is that it gives an overview of model performance. Accuracy for the matrix may be calculated by taking the typical of the values lying across the most diagonal. It's given as shown in Figure 4

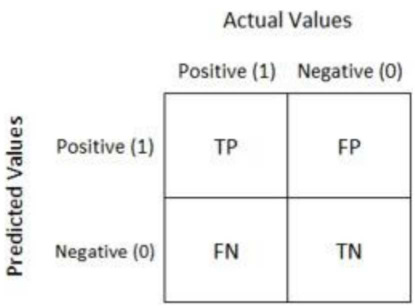

Fig. 4. CONFUSION MATRIX

Where,TP: True Positive, FP: False Positive,

FN: False Negative, TN: True Negative,

Accuracy = $(TP + FP)/N$, N:Total number of samples

6 RESULT AND DISCUSSIONS

The results of the outcome of the project work could be summarized as follows:

6.1 SVC

The Classification report for SVC is as shown in Table 4 and Confusion matrix is shown in Figure 5

– CLASSIFICATION REPORT :

– CONFUSION MATRIX :

Table 1. SVC-CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.82	0.89	0.85	133
1	0.84	0.72	0.78	98
accuracy			0.82	231
macro avg	0.83	0.81	0.81	231
weighted avg	0.82	0.82	0.82	231

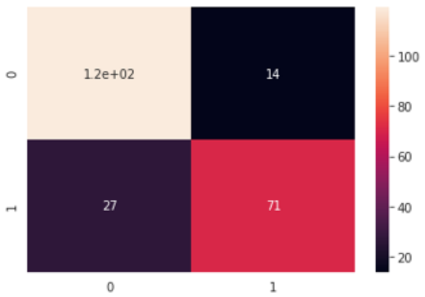

Fig. 5. CONFUSION MATRIX : SVC

6.2 RANDOM FOREST CLASSIFIER:

The Classification report for RFC is as shown in Table 2 and Confusion matrix is shown in Figure 6

– CLASSIFICATION REPORT :

– CONFUSION MATRIX :

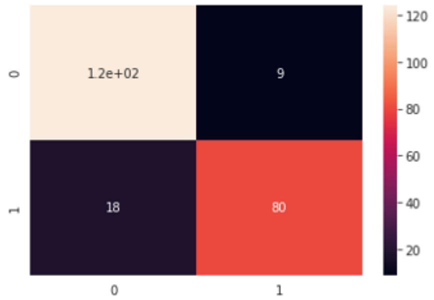

Fig. 6. CONFUSION MATRIX : RFC

Table 2. RFC-CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.87	0.93	0.90	133
1	0.90	0.82	0.86	98
accuracy			0.88	231
macro avg	0.89	0.87	0.88	231
weighted avg	0.88	0.88	0.88	231

6.3 DECISION TREE:

The Classification report for Decision Tree is as shown in Table 3 and Confusion matrix is shown in Figure 7

– CLASSIFICATION REPORT :

Table 3. DECISION TREE :-CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.87	0.93	0.90	133
1	0.90	0.82	0.86	98
accuracy			0.88	231
macro avg	0.89	0.87	0.88	231
weighted avg	0.88	0.88	0.88	231

– CONFUSION MATRIX :

6.4 K-NEIGHBOURS CLASSIFIER :

The Classification report for KNN Classifier is as shown in Table **??** and Con- fusion matrix is shown in Figure 8

- CLASSIFICATION REPORT :
- CONFUSION MATRIX :

6.5 LOGISTIC REGRESSION :

The Classification report for Logistic Regression is as shown in Table 5 and Confusion matrix is shown in Figure 9

- CLASSIFICATION REPORT :
- CONFUSION MATRIX :

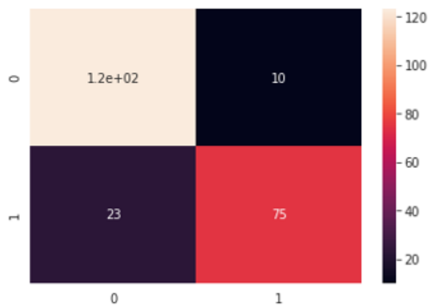

Fig. 7. CONFUSION MATRIX : Decision tree

Table 4. K-NEIGHBOURS CLASSIFIER :- CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.83	0.92	0.88	133
1	0.88	0.74	0.81	98
accuracy			0.85	231
macro avg	0.86	0.83	0.84	231
weighted avg	0.85	0.85	0.85	231

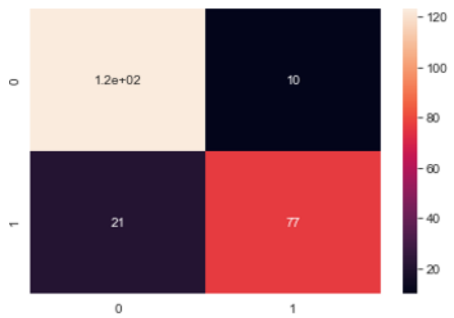

Fig. 8. CONFUSION MATRIX : K-NN

6.6 GRADIENT BOOST CLASSIFIER :

The Classification report for Gradient Boost Classifier is as shown in Table 6 and Confusion matrix is shown in Figure 10

- CLASSIFICATION REPORT :
- CONFUSION MATRIX :

Table 5. LOGISTIC REGRESSION :- CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.74	0.87	0.80	133
1	0.77	0.58	0.66	98
accuracy			0.75	231
macro avg	0.75	0.73	0.73	231
weighted avg	0.75	0.75	0.74	231

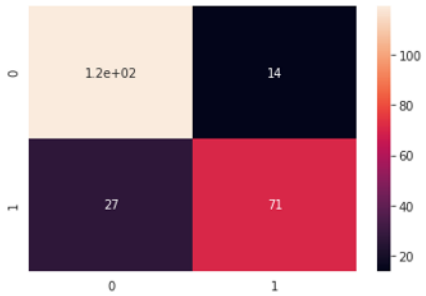

Fig. 9. CONFUSION MATRIX : Logistic Regression

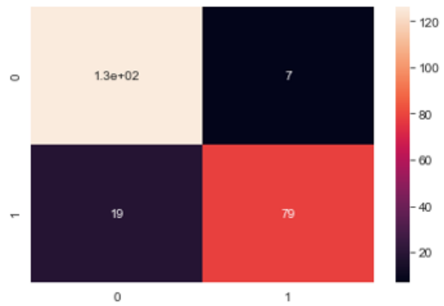

Fig. 10. CONFUSION MATRIX : Gradient Boost Classifier

6.7 XGBOOST CLASSIFIER :

The Classification report for XG Boost Classifier is as shown in Table 7 and Confusion matrix is shown in Figure 11

- CLASSIFICATION REPORT :

– CONFUSION MATRIX :

Table 6. GRADIENT BOOST CLASSIFIER :- CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.87	0.95	0.91	133
1	0.92	0.81	0.86	98
accuracy			0.89	231
macro avg	0.89	0.88	0.88	231
weighted avg	0.89	0.89	0.89	231

Table 7. XGBOOST CLASSIFIER :- CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.89	0.94	0.91	133
1	0.91	0.84	0.87	98
accuracy			0.90	231
macro avg	0.90	0.89	0.89	231
weighted avg	0.90	0.90	0.90	231

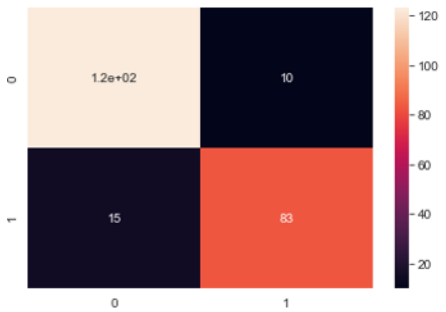

Fig. 11. CONFUSION MATRIX : XG Boost Classifier

6.8 CAT BOOST CLASSIFIER :

The Classification report for Cat Boost Classifier is as shown in Table 8 and Confusion matrix is shown in Figure 12

– CLASSIFICATION REPORT :

– CONFUSION MATRIX :

Table 8. CAT BOOST CLASSIFIER :- CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.87	0.93	0.90	133
1	0.90	0.82	0.86	98
accuracy			0.88	231
macro avg	0.89	0.87	0.88	231
weighted avg	0.88	0.88	0.88	231

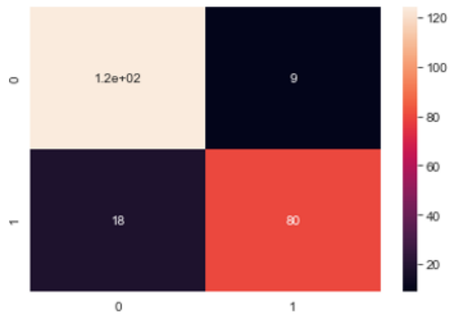

Fig. 12. CONFUSION MATRIX : Cat Boost Classifier

6.9 VOTING CLASSIFIER :

The Classification report for Voting Classifier is as shown in Table 9 and Con- fusion matrix is shown in Figure 13

– CLASSIFICATION REPORT :

– CONFUSION MATRIX :

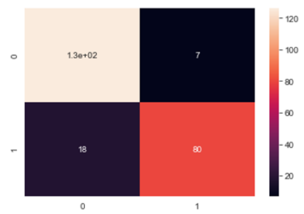

Fig. 13. CONFUSION MATRIX : Voting Classifier

Table 9. VOTING CLASSIFIER :-CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.88	0.95	0.91	133
1	0.92	0.82	0.86	98
accuracy			0.89	231
macro avg	0.90	0.88	0.89	231
weighted avg	0.89	0.89	0.89	231

6.10 RESULTS AFTER MODEL EVALUATION

The accuracy score and classification report of standard existing machine learn- ing predictive models is as shown in the Table 10, Figure 14 and comparing and selecting the best suited algorithm for diabetes prediction.

Table 10. Model Evaluation Table

MODEL	SCORE
XGB CLASSIFIER	89.177489
GRADIENT BOOSTING CLASSIFIER	88.744589
RANDOM FOREST CLASSIFIER	88.311688
CAT BOOST CLASSIFIER	88.311688
DECISION TREE CLASSIFIER	86.580087
K-NEIGHBORS CLASSIFIER	84.415584
SVC	82.683983
LOGISTIC REGRESSION	75.324675

7 Conclusion

In this paper, Polygenic or Diabetes Disease prediction has been achieved for the PID dataset using the projected RandomForest and the XGBoost Classifier model, where preprocessing plays an important role in ensuring a stable and exact prediction. Data preparation enhanced the dataset's quality (standardis- ation and filling missing values). The correlation between attribute and desired

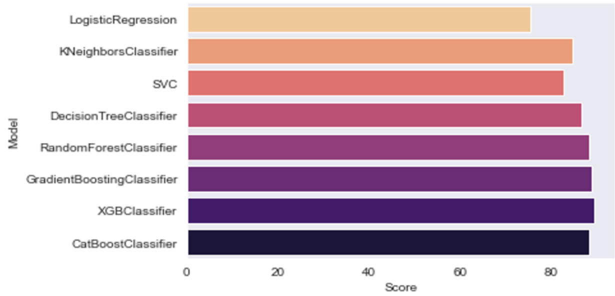

Fig. 14. Classification

outcome will improve if the connection is mostly dependent on attribute choice. The instructional capability of these classifiers will be driven by hyperparame- ters of multiple classifiers, which were tuned using a grid search approach in our models. The outcomes of all the models were evaluated, and the final output of our proposed framework outperformed other models identified in previous re- search works, indicating that the PID dataset has a great potential for Diabetes prediction.

8 FUTURE WORK

Our model's capacity to predict patients with Diabetes by exploiting several commonly used scientific laboratory data is strong, with acceptable sensitivity. In the future, the trained model will be used to create an online app with a user- friendly interface to help clinicians anticipate patients with potential polygenic illness prevalence and provide appropriate preventative measures. In addition, the proposed framework will be tested in a variety of medical settings to ensure its universality and flexibility in predicting disease categories.

References

1. Anand, A. and Shakti, D., 2015, September. Prediction of diabetes based on per- sonal lifestyle indicators. In 2015 1st International Conference on Next Generation Computing Technologies (NGCT) (pp. 673-676). IEEE.
2. Nithya, B. and Ilango, V., 2017, June. Predictive analytics in health care using ma- chine learning tools and techniques. In 2017 International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 492-499). IEEE.
3. Eswari, T., Sampath, P. and Lavanya, S.J.P.C.S., 2015. Predictive methodology for

diabetic data analysis in big data. Procedia Computer Science, 50, pp.203-208.

4. Joshi, T.N. and Chawan, P.P.M., 2018. Diabetes prediction using machine learning techniques. Ijera, 8(1), pp.9-13.

5. Faruque, M.F. and Sarker, I.H., 2019, February. Performance analysis of machine learning techniques to predict diabetes mellitus. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-4). IEEE.

6. Kumar, P.S. and Pranavi, S., 2017, December. Performance analysis of machine learning algorithms on diabetes dataset using big data analytics. In 2017 Interna- tional Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions)(ICTUS) (pp. 508-513). IEEE.

7. Khanam, J.J. and Foo, S.Y., 2021. A comparison of machine learning algorithms for diabetes prediction. ICT Express.

8. Suresh, K., Obulesu, O. and Ramudu, B.V., 2020. Diabetes Prediction using Ma- chine Learning Techniques. Helix, 10(02), pp.136-142.

9. Abhishek, M.B. and V Shet, N.S., 2018. Data transmission unit and web server interaction to monitor water distribution: A cyber-physical system perspective.

10. Abhishek, M.B. and Shet, N.S.V., 2019. Cyber physical system perspective for smart water management in a campus. Desalination and Water Treatment, 147, pp.296-307.

11. Abhishek, M.B., 2021. Conceptualization and Modelling the Functioning of The Networking Unit in Cyber Physical System for Smart Water Management. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(10), pp.4277-4286.

12. Abhishek, M.B. and Shet, N.S.V., 2019, November. Data Processing and deploying missing data algorithms to handle missing data in real time data of storage tank: A Cyber Physical Perspective. In 2019 1st International Conference on Electrical, Control and Instrumentation Engineering (ICECIE) (pp. 1-6). IEEE.

13. Bindu, H.M. and Gupta, K.A., 2020, December. Wideband Digital Channelizer based on Spectrum Sensing. In 2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC) (pp. 1-6). IEEE.

14. Abhishek, M.B., Tejashree, S., Manasa, R. and Vibha, T.G., 2021, May. Smart Agriculture Management System Using Internet of Things (IoT). In Proceedings of International Conference on Sustainable Expert Systems: ICSES 2020 (Vol. 176, p. 363). Springer Nature.