Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 372-395
Received 15 April 2021; Accepted 05 May 2021.

Prognostication of Diabetes Diagnosis Based on Different Machine Learning Classification Algorithms

[1.] S. Prasanth, [2.] M. Roshni Thanka,[3.] E. Bijolin Edwin, [4]V. Ebenezer
[1,2,3,4.] Department of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu, India.
*sprasanth19@karunya.edu.in

Abstract— Diabetic disease is one of the most deadly and infectious illnesses that induce blood glucose (sugar) levels to rise. If this disease is left unchecked and undiagnosed, it can lead to a slew of consequences. Diabetes is the creator of various diseases like diabetic retinopathy, heart problems, etc. The time-consuming identification procedure leads to a patient's initial consultation to a clinical facility and consultation with a practitioner. Data science techniques always have opportunity to aid other scientific areas by avoiding existing concerns on traditional issues. Machine learning is a recent field of AI which looks at how machines learn via their experiences. The emergence of ML strategies on the other hand, addresses this crucial issue. If accurate earlier detection is achievable, the health risk and intensity of disease may be greatly decreased. Owing to the small range of labelled records and the inclusion of data points and missing values in diabetic databases, stable and reliable diagnosis prognostication is extremely difficult. The aim of this research is to build a model that can accurately predict the risk of developing diabetes in individuals. Exploratory data analysis, data pre - processing, feature importance, feature engineering, and different Machine Learning classifiers are all part of a plan that we introduce for diabetes to diagnose. For diabetes prognostication, our paradigm outshines the various strategies mentioned in the study. It can also provide better results from the same data source, leading to better diagnostic of prediction performance.

Keywords: Data Science, Machine Learning strategies, Pre-processing, Feature engineering, Classifiers.

I. INTRODUCTION

Diabetes diagnosis is a group of chronic illnesses wherein glucose levels or sugar levels stay abnormally high for extended spans of time Frequent urination, elevated appetite, and increased hunger are signs of high blood sugar. Diabetes can cause multiple problems if left unchecked. Diabetic ketoacidosis, hyperosmolar hyper-glycaemic condition, or death may be acute complications. Cardiovascular disease, stroke, progressive kidney disease, foot ulcers, and eye injury are serious long-term risks. Diabetes mellitus, also known as diabetes, is another metabolic disorder that causes elevated blood sugar. The enzyme insulin transports sugar from the blood into the cells for absorption or use for eating. The body either doesn't have sufficient insulin for diabetes or doesn't use the insulin it produces effectively. But also untreated excessive blood sugar from diabetes can be affected in the liver, eye, lungs, kidneys, and other organs. One of the most severe and chronic conditions that cause blood sugar to increase is considered to be diabetes. Several risks occur as diabetes remains unchecked and unexplained. The exhausting identification process consists of a patient visiting and consulting a consultant at a medical centre. Yet this key issue is answered by the rise in approaches to machine learning.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 372-395
Received 15 April 2021; Accepted 05 May 2021.

Classification approaches are widely used in the medical field to classify data subject to such constraints according to a single classifier in multiple classes. Diabetes is a disorder that inhibits the bodies natural capacity to regulate insulin receptors, which in turn induces unhealthy carbohydrate metabolism and increases blood glucose levels. High blood pressure is typically caused by diabetes. Intensifying hunger, intensifying appetite, and frequent urination are all the signs induced by increased blood sugar. Because diabetes is left uncontrolled, it causes a slew of issues. Metabolic acidosis and – anti metabolic derangement syndrome are two important signs. Diabetes is recognized as a global public health issue in which calculating sugar consumption is challenging. However, diabetes is not only caused by various variables such as height, weight, genetic factor, and insulin, but the main reason is also the concentration of sugar among all variables. Early warning is the safest solution to staying away from the issues. Several researchers use numerous of machine learning approach for classification algorithms to run disease diagnosis trials such as MLP, SVM, Naive Bayes, Decision Tree, Decision Table, etc. Machine-learning algorithms have been shown to be better at diagnosing various illnesses. Data mining and machine learning algorithms are gaining attention as a result of their ability to handle a large number of data sources from many formats, blending data and inserting historical data into the study.

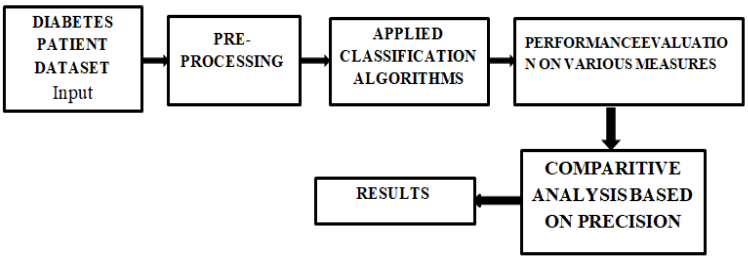

Fig 1.1.Overview of Diabetes prediction process

1.1 Pre-Processing in Diabetes diagnosis

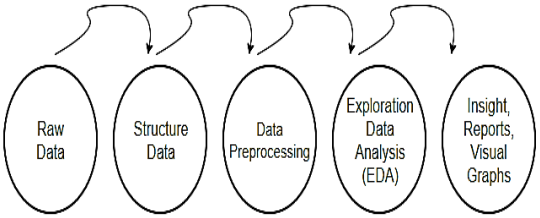

Fig 1.2.Pre-Processing process in diabetes diagnosis

Every repository is a collection of objects with related documents. Sample data, events, observations, and records are all appropriate names. Interestingly, every of them has been characterised by a set of features. In data science terminology, these are referred to as characteristics or features are until creating a model with these attributes, data Preprocessing is needed. By data Preprocessing, we:

- Improve the accuracy of our database. We remove any values that are inaccurate or missing as a result of medical errors or defects.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 372-395
Received 15 April 2021; Accepted 05 May 2021.

- Consistency should be improved. The precision is harmed when there are data inconsistencies or duplicate entries.
- Make the data as achieve as possible. If necessary, we could even replace in the missing characteristics.
- The collected data should be clean. We find things simpler to use and perceive this way.

1.2 Feature Importance and Feature Engineering in diabetes diagnosis:

Feature importance corresponds to a collection of strategies for awarding metrics to input data or features in a particular classification, indicating their quantitative significance while generating predictions. Tactics that give a ranking to inputs attributes depending about however effective there were for anticipating specific variables are known as feature significance. Standardized performance metrics, parameters measured as result of existing systems discriminant analysis, and variation significance grades are some of the most common forms and indicators of function importance results. In a statistical analytics plan, aspect significance metrics perform a significant position in delivering analysis of the information, insights towards another system, and the foundation for computational complexity restriction and feature engineering, which can increase the reliability and efficacy of a coherent system on the issue.

Feature engineering is the tactic of removing functionality via original data through technical information and data mining strategies, these characteristics can help ML algorithms function better Feature engineering may be thought of as a type of ML and data science.

1.3 Approaches and Methods for classifications(classifiers) used in this project

Machine learning methods are used in multiple prognostication classifiers. The machine learning process of training a mapping function an input to even an output relying on instance input-output pairs is known as supervised learning. It uses named training data and a series of training experiences to conclude a feature.
Define the Features to their respective characteristics using classification algorithms. In, we use some typical classifiers like:

- Multi-layer Perceptron
- K-Nearest Neighbour
- Gaussian Mixture Model
- Hidden Markov Model
- Support Vector Machine
- singular value Decomposition
- Artificial Neural Network
- Random Forest
- Decision Trees
- Boosting
- Naïve Bayes
- Classification And Regression Trees
- Light GBM

In data science, machine learning is the classification of a supervised theory of learning that separates a set of data into categories. Voice, detection, facial identification, and other classification problems are the most critical. It may be a problem with binary sorting or a

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 372-395
Received 15 April 2021; Accepted 05 May 2021.

problem with different groups. There are several common machine learning algorithms for classification and recognition in machine learning.

1.4Objective

- The motivation of this project is to create a simplest yet complete representative of prediction of diabetes diagnosis detection.
- This study uses a distinctive classifier to determine whether or not a patient will develop diabetes. As previously mentioned, the dataset includes the labels for the model's dummy variable Result. The goal is to predict the risk of diabetes based on a subset of variables such as blood pressure, insulin levels, glucose, skin thickness, and BMI.
- The objectives of this work is to develop a modern technologies that can combine the impacts of multiple machine learning tactics to provide a more precise earlier start diabetes prognostication for a service user. SVM, Logistic regression, ANN, MLP, Random Forest, Decision Tree, Boosting, KNN, and other machine learning techniques were used in this project plan to prognosticate diabetes.

II. LITERATURE SURVEY

2.1Related Works

There has been a significant amount of study has been done in the area of Machine learning based prediction on diabetes diagnosis detection.

In [1] they have a proposed a Prediction of Diabetes using Classification Algorithms and The focus of this study is to build a model that can more accurately assess the threats of diabetes in patients. As a result, this thesis hires three machine learning classification algorithms and techniques also a tentative diagnosis of diabetes. The results of all three algorithms are calculated using various metrics. Accuracy is measured in terms of correctly categorized and incorrectly classified cases. The findings show that Naive Bayes outperforms other algorithms, with an accuracy of 76.30 percentage growth.

In [2] they have recommended theMachine Learning Based Unified Framework for Diabetes Prediction. We suggested a system for the estimation, tracking and execution of diabetes in real time. Our mission is to create an optimised and effective framework for machine learning (ML) that can identify and forecast the state of diabetes effectively. In this research, the five most important classification strategies for machine learning were considered for predicting diabetes. Fortunately, in order to evaluate the feasibility of these classification processes, more research is needed. As compared to the other four classifiers, the analysis showed that Nave Bayes had the highest performance, with an F1 measure of 0.74.

In [3] they have proposed an Analysis and Prediction of Diabetes based on Machine Learning. The key objective is to detect new trends and then to analyse these patterns and provide users with relevant and usable data. Diabetes leads to heart failure, kidney disease, blindness and nerve damage. A main issue is the effective mining of diabetes data. The techniques and techniques of data mining will be discovered to identify the required methods and techniques for effective diabetes dataset classification and extraction of useful patterns. To construct an accurate model, the dataset was studied and analysed. Diabetes disease detection and diagnosis. In this research, we plan to use the resampling technique of

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 372-395
Received 15 April 2021; Accepted 05 May 2021.

bootstrapping to improve precision and then add Naïve Bayes classifier, Decision Trees algorithm and Knn classifiers and compare their efficiency.

In [4] they have proposed a Machine learning framework for diabetes disease diagnosis and detection. The main goal of this study is to develop a machine learning-based approach for forecasting diabetic patients. Primarily focused on the p value and odds ratio, logistic regression is used to classify risk factors for disease prediction. .To determine diabetic patients, we used four distinct classifiers. These protocols have also been followed and replicated in 20 trails by three groups of partition protocols (K2, K5, and K10). The precision of these classifiers is used to test their reliability.

In [5] they have explored implementing Machine Learning Techniques to predict diabetes. In this study, various machine learning strategies were added to the input, and detection was performed using various classifiers, with Logistic Regression ensuring the finest of 96 % accuracy. Pipeline implementation offered the Adaboost classifier with 98.8 % accuracy as the best model. There was also a variation in performance between the machine learning model and two separate samples. In comparison to current data sets, it is apparent that the methodology improves the reliability as well as accuracy of diabetes prediction with this dataset.

In [6] A Comparative Study of Machine Learning Methods to Predict Diabetic Mellitus was discussed. For the estimation of diabetic mellitus in adult population data, we use four widely used machine learning algorithms: Support Vector Machine, Naive Bayes, K-Nearest Neighbour, and C4.5 Decision Tree. In comparison to other machine learning models, the C4.5 decision tree obtained better accuracy according to our results.

In [7] they suggested the Detection of diabetes Disease Prediction Based ML on Public health care Big Data analytics. This work aims to build a binary classifier using the WEKA tool to predict future diabetes diagnosis using various classifiers, and the Base model is CART classifier.Centered on promising actual outcomes, the study intends to recommend the best model for diabetes disease diagnosis. Taking a look at the test findings of each classifier in the dataset. It was discovered that the SVM performed well in disease diagnosis, with the greatest precision the highest, which is 79.13. SVM has a significantly better accuracy rate in anticipating diabetes than other classification algorithm.

In [8] they discussed by using auto tuning multi-layer perceptron to build an accomplished method for diabetes prognostication. The thesis builds an unique outlier recognition approach by combining an AutoMLP mechanism instead of an Enhanced Class Outlier Detection exceptional case recognition system that uses a distance-based methodology. The Automation Multi-Layer Perceptron proposed approach is auto-tunable, which means it instantaneously optimizes parameters during in the training process of the test, which would otherwise require human interference. Outlier detection is performed by our system during the pre-processing of data. The machine achieved an accuracy of 88.7%, which was higher than any previous findings.

In [9] they've suggested incorporating machine learning approaches to evaluate diabetes. In the world of data science, ML is a modern area of research that explores how computers learn from their experiences. The aim of this research is to develop technologies which could incorporate the results of different machine learning methods to provide a more early risk diabetes predictor for an individual. This article intends to anticipate diabetic using various supervised machine learning tactics SVM, logistic regression, and ANN.This initiative seems

to have the aim of recommending an appropriate strategy. We selected SVM to anticipate diabetes as it is an efficient method for comparative classifier. The justification for this is that SVM is well known for its discriminatory identification power, specifically when a substantial majority of classification methods are concerned in the features, and the component of the function in our case is 7.

In [10] they have proposed a Prediction of Type 2 Diabetes using Machine Learning Classification Methods.This article investigated the likelihood of diabetic in individuals based on their diet and personal health records. Using multiple machine learning algorithms, the probability of type 2 diabetes was estimated because these algorithms are highly reliable, which is really important in the health care profession. When the model is educated with reasonable specificity, people will determine the risk of diabetes on their own. The precision of our dataset's Random Forest is 94.10%, which is the best among the others. For the PIMA dataset, Random Forest often has the best precision. Learning algorithms applied to six different machines of classifiers

In [11] they suggested a Diabetes Analysis Using Various Machine Learning Methodologies. The aim of this research is to develop a tool for determining a patient's diabetes perceived risk which is more reliable. Pattern is applied using classification techniques such as Decision Trees, ANNs, and SVMs.The models for Decision Tree have a performance of 85 % 77 % for NB, and 77.3 % for SVM. The results point to a high level of accuracy in the techniques.

In [12] they suggest implementing machine learning approaches to classify diabetes medical data. The objective of this work is to discover insights by observing the trends in the repository and using predictive analytics to classify these diseases. In addition, the solution to the neural network is often used to identify current diabetic patient data to predict the disease of the patient based on qualified data that can contribute to the detection of various levels of diabetes affected people. It is often compared to the extraction system of collaboration rule for model evaluation to ensure that the classification is accurate.

In [13] they suggested to use data science of ML tactics to recognize diabetes in data sets. They expect to apply the bootstrapping-like approach in this analysis to upgrade the accuracy and then apply it to different classification strategies and learn about their application. After Bootstrapping (Accuracy Rates %) comparing all classifiers the SVM and Ada boost gives highest precision of 94.44%

In [14] they proposed Analyzing Machine Learning Techniques and Achieving the Greatest Performance for Diabetes Prognostication in Female.We worked to identify the best fitting algorithm for this function in order to successfully predict and diagnose diabetes. To achieve the highest precision, the key purpose is to compare the various algorithms. Few classifiers were compared for finding the best outcome. With the aid of K-Fold and Cross Validation, the final outcome gave us an accuracy of 81.1%.

In [15] they proposed a Research on Diabetes Prediction Method Based on Machine Learning. We use supervised ML algorithms such as SVM, Classifier Naive Bayes classifier and LightGBM in this manuscript to train based on the real data of 520 diabetic and possible diabetic patients between the ages of 16 and 90. The efficiency of the support vector machine is the greatest, by comparative study of classification and recognition accuracy.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 372-395
Received 15 April 2021; Accepted 05 May 2021.

In [16] A Study on Diabetic Predictions Using Machine Learning has been suggested by them. It aims to use machine learning models such as SVM and Naive Bayes. Using such system to determine diabetes will help save other time and provide more accurate outcomes.

In [17] they have proposed diabetes to Heart Disease: A Survey. In addition to glucose level, heart rhythm, BMI, era, circulatory pressure, we understand heart disease by the diabetes side effects. The suggested system is to classify the heart condition for which we implement the prediction system based on diabetes and separate the heart disease using SVM measurement.

In [18] they have recommended a Survey on recent developments in automatic detection of diabetic retinopathy. Here we use various tactics on deep learning, machine learning and medical image processing. DR is diagnosed with a CAD protocol. It is used to analyse medical images, Medical diagnosis Image Recognition.

In [19] they recommended a machine learning method for predicting and diagnosing possible diabetes threat We attempted to focus forward in this review on the onset of diabetes, and is one of the global 's highest degenerative illnesses, according to the Health Organization. We have attempted to show numerous approaches including certain Classification Algorithms like GB, LR, and NB, which can be used to diagnose diabetes disease with 86 %accuracy for Gradient Boosting, % accuracy for Logistic Regression, and % accuracy for Naive Bayes.

In [20] they suggested a machine learning methodology for diabetes disease prognostication and diagnosis. The combination of a classifier based on LR and RF works well. For the prediction of diabetic patients, this mixture would be really useful. The ML-based system's overall precision is 90.62%. The fusion of LR-based feature selection and Random Forest based classification algorithm provides 94.25 % accuracy and 0.95 precision for the K10 protocol.

2.2 Title, classifiers used, Performance Metrics, and Merits of an existing system are:

Ref No	Classification (Classifier) and Algorithm used	Performance Metrics	Merits	Demerits
[1]	Decision Tree classifier , SVM and NB	The findings reveal that Naive Bayes outperforms other classifiers, with an accuracy of 76.30 %. These results are verified using Receiver Operating Characteristic curves in a reasonable and validated manner.	Constant Improvement. ML algorithms are able to learn from the knowledge that we have.	The model's performance can be impacted by algorithm selection.
[2]	Five classifiers are used	The findings of the study show that Naïve Bayes obtained the best efficiency, achieving the F1 measure of 0.74, compared with the other classifiers.	Time-consuming	There's a risk that ML model just might make a significant error.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 372-395
Received 15 April 2021; Accepted 05 May 2021.

[3]	Naive Bayes Decision Tree, KNN	The WEKA software has been used to diagnosis diabetes as a mining tool. High precision with an accuracy score of 90.36 % and Stump's judgement offered less precision than some by offering 83.72 %.	Handling multi-dimensional and multi-variety data	For train and test the data, we collect a large volume of data. Such procedure can occasionally result in data uncertainty.
[4]	NB , Decision Tree, Ada Boost, RF classifier	The cumulative precision of the ML-based model is 90.62%. The combined effect of LR-based function choice and RF-based classifier offers 94.25 % of accuracy and 0.95 AUC for the K10 protocol.	Easily identifies trends and patterns	ML can consider taking time and effort to achieve outputs
[5]	Decision Tree Gaussian NB , SVM ,Random Forest ,Extra Trees AdaBoost ,MLP ,LR , Gradient Boost Classifier , KNN ,Bagging	Logistic regression offers 96 % precision. The pipeline application gave the Ada Boost classifier 98.8 % as the best model.	Helps in Identifying Diseases and Diagnosis	In order to assess the usefulness of ml strategies it is therefore essential to analyse the data.
[6]	SVM, Naïve Bayes, KNN, C4.5	According to the investigation, and discussion the C4.5 decision tree outperforms those certain diabetes input classifier model by 73.5 %.	Helps in Identifying Diseases and Diagnosis	Over Fitting of the Training Samples
[7]	Naïve Bayes, SVM, RF, simple CART	On the highest, which is 0.7913, is according to the Classification Accuracy of SVM. SVM's overall success in predicting diabetes is higher than that of NB, Random Forest classifier and Simple CART.	Improving quality of life	Interpretation of the findings
[8]	Auto MLP	Auto MLP Outlier Identification has an	Improving efficiency and quality for care.	It necessitates a considerable

		accuracy of 88.7%, a Mean Score Recall of 88.5, and a Weighted Mean Precision of 85.8%.		variety of data.
[9]	SVM, Logistic regression, ANN	SVM is an effective approach for binary supervised learning, so we chose SVM to predict diabetes.	Fast Processing and Real-Time Predictions	When the sample data contains more distortion, such as conflicting target classes, SVM classifier does not accomplish efficiently.
[10]	LR, KNN, SVM,NB,Decision Tree and RF classifiers were used	Hence the precision of used dataset's Random Forest is 94.10 % which is the best among the others. For the PIMA dataset, Random Forest often has the best precision. Learning algorithms applied to six different classifiers	Machine Learning Improves Over Time	Utilizing the KNN classifier The consistency of the data is determined by its efficiency.
[11]	Decision tree classifier, ANN, NBand SVM model.	The models with the good success are Decision Tree of 85% accuracy, Naïve Bayes obtains of 77%, and SVM had given 77.3%precision.	Contingency ratings that are reliable and appropriate allow reliable as well as consistent resource distribution, resulting in high levels of efficiency results.	We can resolve the difficulties of exhibiting the challenge to the network by using ANN.
[12]	Rule-based strategies Neural network	Rule-based method gives 88.6% of accuracy and Neural network gives 88.5% precision	No Human Intervention Needed (Automation)	It's complex and time-consuming to enumerate any of the rules.
[13]	SVM, Decision Trees , Ada boost , Linear regression	After Bootstrapping (Accuracy Rates %) comparing all classifiers the SVM and Ada boost gives highest precision of 94.44%	machine learning enables computers to access hidden insights	Linear regression classifier only considers the dependent variable's mean value.
[14]	DT, Logistic Regression classifier, NB, SVM and KNN algorithm	With the aid of KFold and Cross Validation, the final outcome gave us an accuracy of 81.1%.	Wide Applications are used here	The concept of linearity here across variables of the study is a significant drawback of LR classifier.
[15]	SVM, Naïve Bayes, Light	SVM has the highest accuracy rate, with an	With a strong margin of differentiation, SVM fits	The presumption of autonomous

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 372-395
Received 15 April 2021; Accepted 05 May 2021.

	GBM	accuracy rate of 96.54%. This illustrates that the most effective diabetes classification algorithm SVM is a forecast.	pretty well. In high-dimensional spaces, it is efficient.	predictive properties is one of NB classifier key flaws.
[16]	SVM and Naïve Bayes classifiers were used.	It could conserve resources and time by doing so to achieve more reliable results by using these algorithms to anticipate diabetes.	This data analysis methodology is used by many hospitals to determine admissions patterns.	It requires more time to process.
[17]	Machine Learning Classifiers, SVM	The suggested system is to classify the heart condition for which we implement the prediction system based on diabetes and separate the heart disease using SVM measurement.	Unlimited Resources Incorporating Information. With timely analysis and evaluation, machine learning can effectively ingest infinite volumes of data.	Automation is a con
[18]	ML model, Deep learning algorithm, Medical Image processing	CAD system is used for diagnosing of DR	It is used to analyse medical images Medical diagnosis Image Recognition.	When it comes to bringing existing models into practice, ML remains a challenge.
[19]	Gradient Boosting, Logistic Regression and Naive Bayes	The obtained accuracy for the model are 86 % for Gradient Boosting, 79 % for Logistic Regression and 77 % for NB the diagnosis of diabetes.	No human activity was required (automation) Multi-dimensional and multi-variety information handling.	It generally requires innovation, exploration, and perseverance.
[20]	Naïve Bayes algorithm , Decision tree classifier, RF and Adaboost	The ML-based system's overall precision is 90.62 % . For the K10 protocol, The integration of LR based selecting features and an RF based classification model yields a reliability of 0.95 and an accuracy of 94.25 %.	Continuous development, continuous improvement in precision and efficacy	Hence the possibility of riskon mass underemployment though they are trying to replace humankind in certain areas.

Table 2.1Title, classifiers used, Performance Metrics, Merits of an existing system

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 372-395
Received 15 April 2021; Accepted 05 May 2021.

2. 3 Data Sources used are:

The Datasets are used from public and open-source platforms like kaggle, etc.

S.no	Tends to require Database (Dataset)
1.	Pima Indian diabetes Database

Table 2.2 Data Sources used

2.4 Review on Findings (Summary):

The focus of the survey is to compare and assess the optimum tools and techniques and advanced features focused on Prediction of diabetes diagnosis detection based on machine learning tactics and algorithm and detection of diabetic retinopathy using AI. The overview of detection of diabetes diagnosis and diabetic retinopathy is shown in below fig 2.1 and 2.2.

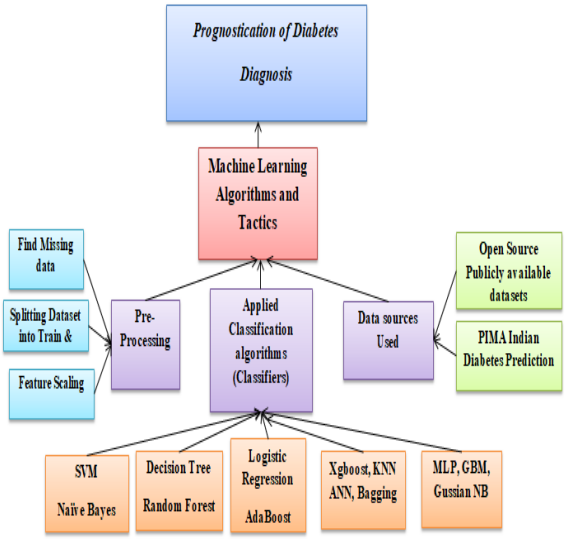

Fig 2.1.Overview of Prognostication of diabetes diagnosis

III. METHODOLOGY

The paper's primary aim is to build a paradigm for diabetes prognostication focused on machine learning methodologies adequate classifier for diagnosis, and accuracy by using PIMA Indian datasets, which we can evaluate and diagnose in an optimal method**.** The method is applied point by point to accomplish the aforementioned objective. The main objective of this project is to build a clear but complete portrayal of diabetic diagnostic identification estimation through ensemble and supervised learning strategies with the System may also be used as a classification algorithm to increase recognition accuracy.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 372-395
Received 15 April 2021; Accepted 05 May 2021.

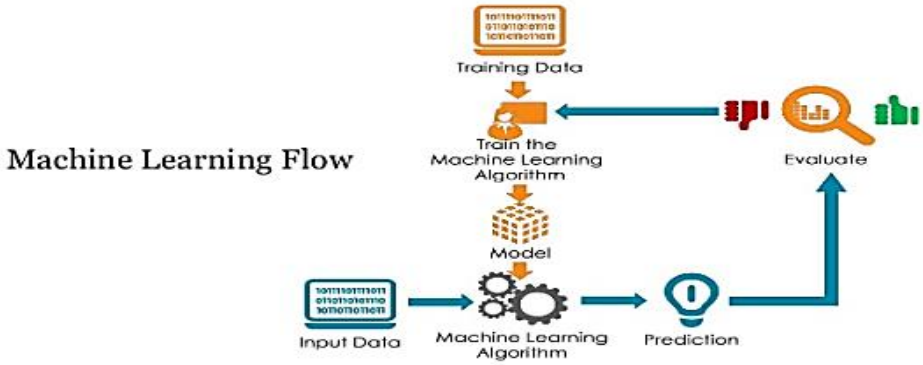

Fig 3.1. Machine Learning Flow

We propose a classification method with increased precision for predicting people with diabetes. Random Forest, Decision Tree, Ada boost, Naive Bayes, voting classifier, KNN, LightGBM and Logistic Regression Judgment were among the classification methods used in this system. The primary aim is to use the possibly the best standard diabetic database to boost the accuracy of the PIMA Data source collection, that consisted of several attributes.

3.1 Architecture Diagram for diabetes diagnosis

The goal of this project was to develop classifier prototype for the disease collection of data, as well as to evaluate when an individual is seriously ill by implementing models to achieving full validity scoring in all of those models. Then, by precisely prognosticating the trained model that corresponds to the classification of the model can effectively produce effective training performance. Figure 3.2 depicts the described strategy.

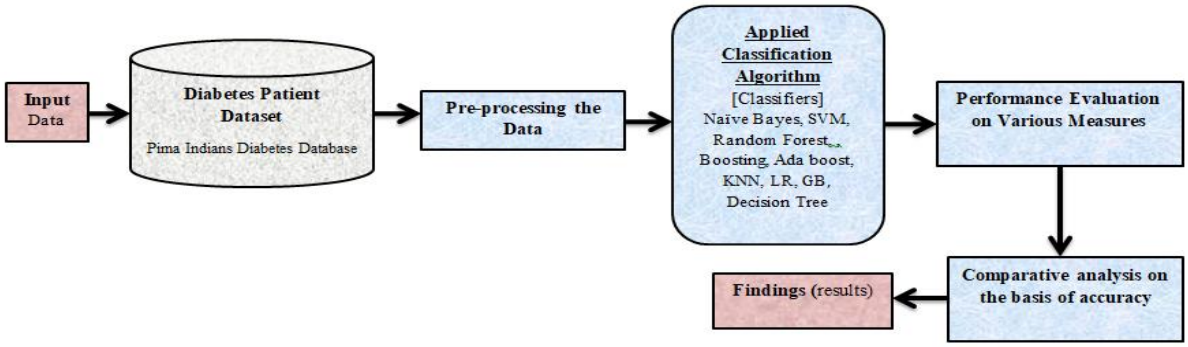

Fig 3.2. Architecture Diagram: Prediction of Diabetes Diagnosis Detection

When using datasets, the system will determine if the individual has diabetic disease. People may use various ML algorithm to recognize the disorder The classifiers that are used were logistic regression, LightGBM, Decision tree, SVM, KNN, Random forest, Ada boost, and gradient boosting . The system's architectural design will ultimately assist in the diabetes diagnosis.

3.2 Illustration of Flow Model for diabetes diagnosis

A flow diagram, also known as a flowchart, is indeed a category of operation illustration which depicts a series of behaviour or activities occurring inside a complicated process .Also

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 372-395
Received 15 April 2021; Accepted 05 May 2021.

flow diagram are a valuable resource for deciding the best directions for individuals, artifacts, or records. A type of diagram that represents that represents a process, computer, or machine's algorithms. They're widely used in a variety of areas to track, analyse, plan, optimize, and interact often complicated mechanisms in plain, convenient visuals.

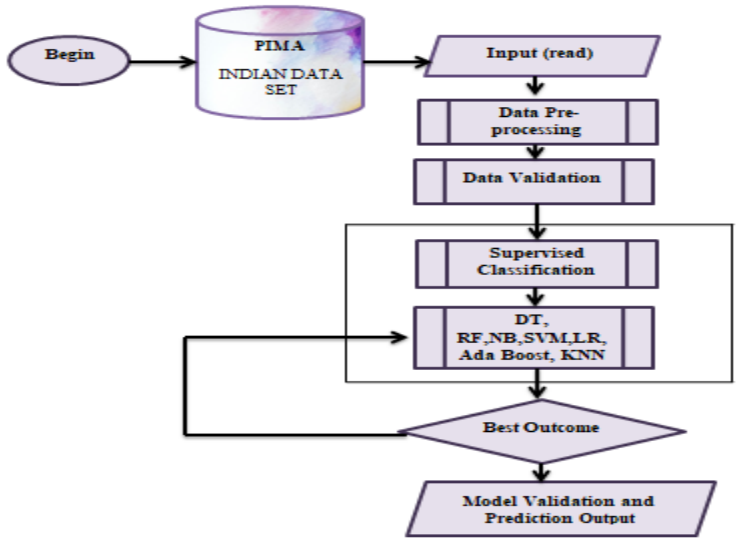

Fig 3.3. Flow Diagram for diabetes diagnosis

3.3. Prognostication of diabetes diagnosis diseases detection Methodology

Steps of detection:

- Data collection process
- Defining data , that is exploratory data analysis
- Data Pre-processing
- Building model
- Analysis
- Results

Algorithm:

- Importing the libraries
- Dataset importing
- Defining dataset
- Preprocessing the data
- Feature Engineering and Importance
- Training and testing on dataset
- Performing the algorithms
- Evaluation and comparison of the outcome results

Data Collection

Collection of data is the act of gathering and processing relevant information from a number of resources. In addition to be using the data we collect to construct usable AI and

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 372-395
Received 15 April 2021; Accepted 05 May 2021.

ML applications, this must be gathered and analyzed in a way that works for the specific nature of the problem.

This dataset consists several variables:

- Pregnancies: No. of times pregnant.
- Glucose: Plasma glucose concentration a two hours in an oral glucose tolerance test Blood Pressure: Diastolic blood pressure (mm Hg)
- Skin Thickness: The triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- Diabetes Pedigree Function: Diabetes pedigree function
- Age: Age (years)
- Outcome: Class variable (0 or 1) which is a binary classification.

Pre-Processing

In ML, data Preprocessing means the practice of data cleaning and filtering relevant data in order to optimize it for the creation and training of ML models.

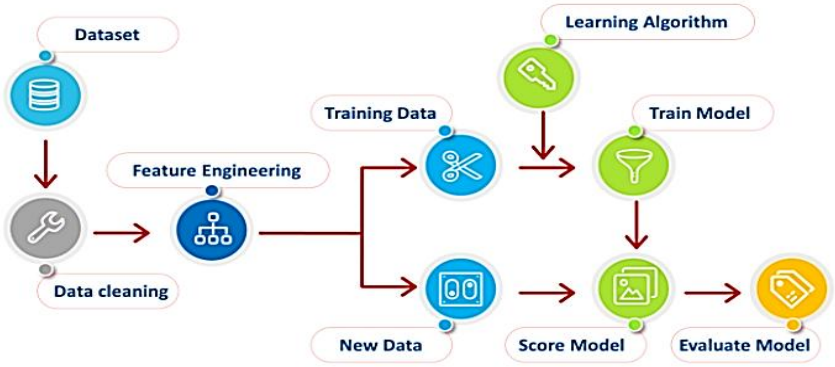

Fig 3.4.Pre-processing action

It involves:

- **Getting the dataset:**Since a ML algorithm is entirely based on information, the very first aspect they need to construct one is a repository. The dataset consists of data in a certain structure for a specific issue.
- Now for instance, unless we're to construct a machine learning algorithm for work purposes, the data source would be distinct from the sample needed for a diabetic patient. As a result, growing datasets is distinct from others. We normally save the database as a csv format so that we can include it in our application. It involves data cleaning process too.
- **Finding the Missing Data:** The following phase inside the data Preprocessing stage is to deal with insufficient data in the repositories. If any of the information in our data source is incomplete, this might pose a significant dilemma for our classification model. As a consequence, handling missed attribute values is needed.
- **Splitting the data into train and test set:** A training collection and a testing dataset are generated from our repository. It is an important stage in Preprocessing since it improves the accuracy of our classification model.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 372-395
Received 15 April 2021; Accepted 05 May 2021.

Preprocessing entails a variety of cleaning, incorporation, optimization, and elimination strategies, and Then i'll focus on finding the missing value tactics here. Via the technique, the characteristics of the information gathered could therefore become clearly understood

Classification (classifiers):

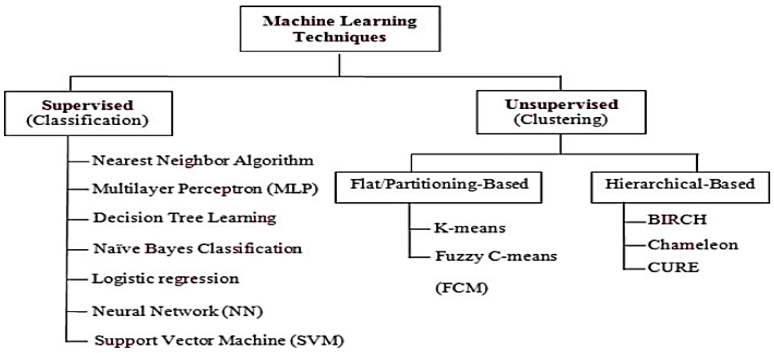

Fig 3.9.Classification Method

Data can be classified in two ways: structured and unstructured. Classification is a process for categorizing data into a set of groups or class. The key aim of a classification challenge is to determine that type or class relevant data is belong to.

Few of the terminologies encountered in machine learning – classification:

- **Classifier:** The methodology that assigns a group to the input information.
- **Feature:** A feature is an observable attribute of an anomaly under investigation
- **Binary Classification:** There are two possible scenarios in this labeling process. outcome can be 1 or 0, True or False etc.
- **Initialize** the classifier to be used.
- **Train the classifier**
- **Predict the target outcome.**

IV. APPLICATIONS USED

Identifying Diseases and Diagnosis	Smart Health Records	Predict chronic disease
Personalized Treatment & Behavioural Modification	Clinical Trial Research	Medical Diagnosis

Table 3.1:Applications

V. TECHNIQUES AND TECHNOLOGIES

- Data science Libraries such as, pandas, matplotlib,seaborn,numpy, etc.
- Machine Learning library ,sklearn are used with many deferential models such as model classifiers, and metrics etc.
- Google Colab used for implementation.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 372-395
Received 15 April 2021; Accepted 05 May 2021.

- Data set available on open source platforms
- Programming Language: Python.

VI. EXPERIMENTAL RESULTS AND IMPLEMENTATION DETAILS

The main objective of the proposal was to use the ML tactics to determine if a person has diabetic or not using diagnoses from the repository. The aim of this work is to build classifier model for disease information gathering and to determine when a patient will become ill by identifying patterns and achieve greater accuracy outcomes within these proposed modelling techniques. The diabetic data presents a binary labelling challenge where so many of the dataset's features available are used to decide that patient has infected or not condition. On the Pima repository, various Preprocessing, extraction and classification, feature engineering, and strategies for determining the severity of diabetic are used, as well as ML and data science approaches. Type 1, type 2, and gestational diabetes are perhaps the most prevalent complication of diabetes. The information will be loaded, features retrieved, and the dataset divided into train and test sets. After that, we'll train the model and activate the classifier. Also we compare the performance of different classifiers.

6.1 Dataset used here:

- **Simulated Database :** Pre – recorded by well trained and well performed data.
- Here we will use the PIMA dataset; Pima Indians Diabetes Database and Predict the onset of diabetes based on diagnostic measures of data.
- This datasets consists of several medical predictor (independent) variables and one target (dependent) variable, which is the Outcome.
- The number of pregnancies the patient has had, their BMI, insulin level, age, and so on is all possible outcomes of independent variables.

For Prognostication of diabetes diagnosis:

We have to load the necessary libraries:

- Main: Pandas ,Numpy
- For visualization (plot): seaborn, matplotlib
- For Models: Sklearn

6.2 Steps for implementing the model:

Prognostication of diabetes diagnosis:

Step1: Import the requisite tasks. Read the data from libraries that have been loaded.

Step2: Read the data and load the dataset using pandas data frame

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Table 6.1: Read and EDA of input data

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 372-395
Received 15 April 2021; Accepted 05 May 2021.

Step3: Review the Descriptive Analysis and the count of the target variables after looking through the results.

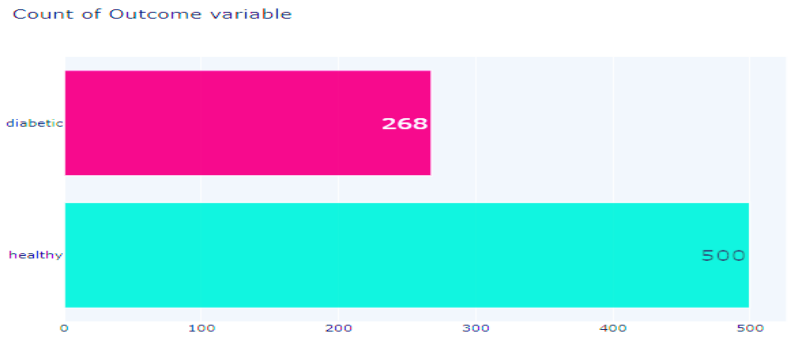

Fig 6.1. Total Target variable

Step4: Do the EDA: The information label's organized data were analysed. The function types of the sample are studied. The scale of the repository was calculated. The dataset set's missing values are 0s and may produce random errors. NaN indices are often used to consider replacing the 0 values. The qualitative data of the data collection are analysed.

Fig 6.2.Histogram

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 372-395
Received 15 April 2021; Accepted 05 May 2021.

Step5: Data Preprocessing Process: The median parameters upon whether growing vector was diabetes or not were used to fill in the NaN values missing findings. Outliers were identified and eliminated using the Local outlier factor.

```
[ ]  missing_values_table(df)

                n_miss  ratio     type
Insulin            374  48.70  float64
SkinThickness      227  29.56  float64
BloodPressure       35   4.56  float64
BMI                 11   1.43  float64
Glucose              5   0.65  float64

There are 5 columns with missing values
```

Fig 6.3.Finding the missing values

Step7: Find the new features and do EDA to get new dataset.

Step8: Prepare Dataset and find the Standard Scalar: **Standard Scaler and label encoder:**

- Computecorrelation Matrix: A matrix of attribute correlations. is a figure that shows the coefficient of association amongst ranges of factors for the table's random variables is compared also with table's certain values. This helps users to assess growing combinations get the greatest association.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.129459	0.141282	0.015304	-0.003258	0.017683	-0.033523	0.544341	0.221898
Glucose	0.129459	1.000000	0.152590	0.162184	0.423109	0.221071	0.137337	0.263514	0.466581
BloodPressure	0.141282	0.152590	1.000000	0.181240	0.075271	0.281805	0.041265	0.239528	0.065068
SkinThickness	0.015304	0.162184	0.181240	1.000000	0.255293	0.514121	0.158611	0.025308	0.217639
Insulin	-0.003258	0.423109	0.075271	0.255293	1.000000	0.216141	0.168702	0.050598	0.259293
BMI	0.017683	0.221071	0.281805	0.514121	0.216141	1.000000	0.140647	0.036242	0.292695
DiabetesPedigreeFunction	-0.033523	0.137337	0.041265	0.158611	0.168702	0.140647	1.000000	0.033561	0.173844
Age	0.544341	0.263514	0.239528	0.025308	0.050598	0.036242	0.033561	1.000000	0.238356
Outcome	0.221898	0.466581	0.065068	0.217639	0.259293	0.292695	0.173844	0.238356	1.000000

Table 6.2: Correlation matrix

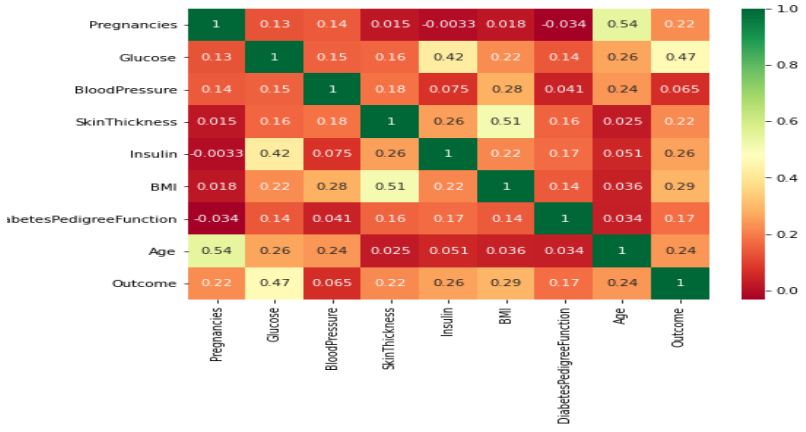

Fig 6.4.Correlation between features

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 372-395
Received 15 April 2021; Accepted 05 May 2021.

- Now we will define x and y.
- Here we consider Base Model as the LightGBM :
  LightGBM is a fast gradient boosting method which uses tree based learningapproaches.
  It gives faster training acceleration and higher reliability and Better accuracy.

- Splitting dataset into training and test data:

- The classifiers here we used here are Linear Regression, KNN,CART,GBM, XGB,LightGBM , Random Forest

Step9: Model Building:

The base model was developed first, followed by a review of the study results. The model was then updated with new functionality and predictor data were updated. Now, we'll Train few Classification models on the Training set which produces the highest accuracy.From comparing the 7 models, we can say that our base model Light GBM has achieved **88.02%**

Feature Importance: It assigns a ranking to all of your data's features; the greater the level, its most significant or noteworthy the functionality is to their outcome parameter

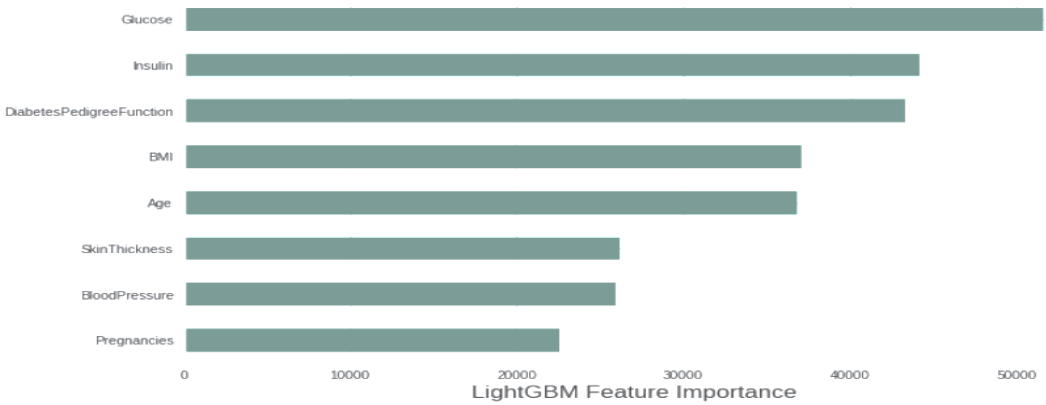

Fig 6.5Feature Importance

Feature Engineering:

- It is the method of extracting functionality through actual data utilizing technology information gathering predictive analytics.
- Such characteristics can help ML algorithms do well. Feature engineering may also be thought of as a type of machine learning and data science.
- One instance of hot encoding is the translation of categorical data into a format that can be fed into machine learning methods to enhance classification accuracies.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 372-395
Received 15 April 2021; Accepted 05 May 2021.

Insulin/Age	BMI/Age	Pregnancies/Age	Ins*Glu	New_BMI	New_BloodPressure	New_Glucose	NewInsulinScore
3.390000	0.672000	0.120000	25086.0	Obes	Normal	Prediabetes	Abnormal
3.306452	0.858065	0.032258	8712.5	Overweight	Normal	Normal	Normal
5.296875	0.728125	0.250000	31018.5	NormalWeight	Normal	Prediabetes	Abnormal
4.476190	1.338095	0.047619	8366.0	Overweight	Normal	Normal	Abnormal
5.090909	1.306061	0.000000	23016.0	Obes	Normal	Normal	Abnormal

Table 6.3:After Feature engineering the new datasets

- A ML model's performance can be improved by tuning it. Adjustment can be described as the method of improving the model's output without causing any controversy or over fitting of a variation. Also check for overfitting problem.

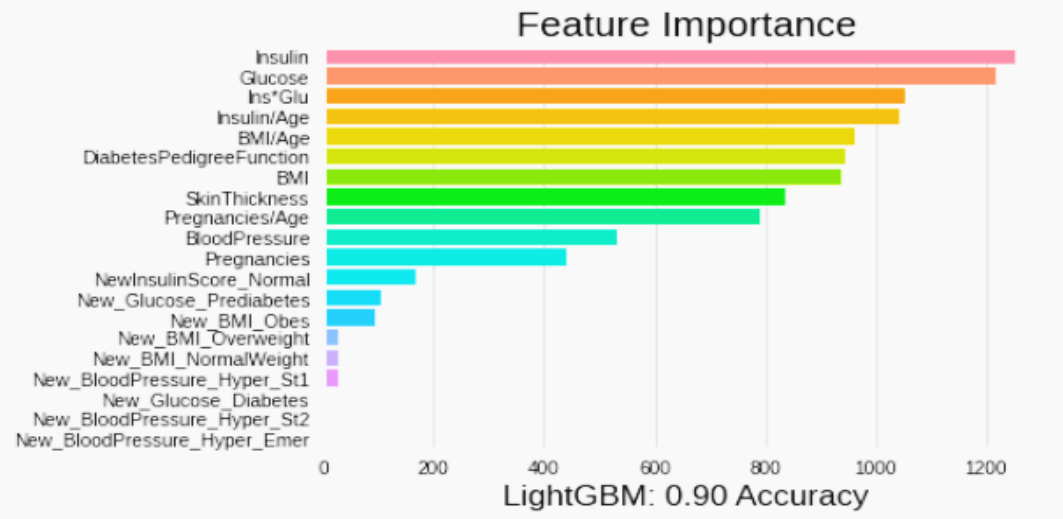

Fig 6.6.Feature Importance

5.3 Experimental Results

The model with the highest score after Hyper Parameter optimization was LGBM with 90.01 % accuracy and 0.90 of cross validation score

Model (Classifiers)	Accuracy
Logistic Regression	77.34%
KNN	85.02%
CART	84.76%
RF	87.76%
GBM	88.15%
XGB	88.02%
Light GBM(BM)	88.28%

Table 6.4:Classification model of accuracy results

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 372-395
Received 15 April 2021; Accepted 05 May 2021.

After Feature Engineering Process:

Model (Classifiers)	Accuracy
RF	87.90%
GBM	88.17%
XGB	89.61%
Light GBM(BM)	88.43%

Table 6.5:Classification model of accuracy results after feature engineering process

After Tuning the Model:

Model (Classifiers)	Accuracy
GBM	88.96%
XGB	89.71%
Light GBM(BM)	90.01%

Table 6.6:Classification model of accuracy results after Model Tuning

After Checking over fitting model:

Model (Classifiers)	Accuracy
LOGISTIC REGRESSION	79.36%

Table 6.7:Classification model of accuracy results after over fitting

- Also, the classification model has been tried using voting classifiers
- A Voting Classifier is a ML algorithm which learns from an ensemble of modeling techniques and determines an outcome (category) dependent on the greatest likelihood of the outcome being the desired class.
- It essentially adds up the results of every classifier passing into Voting Classifier and prognosticates the outcome class based on the most votes. Rather than making individual designated models and assessing their reliability, we form a new model that train on all these models and predictions performance based on the cumulative majority of votes for each level of output.

Here we tried the proposed work using voting classifier to build the model

- LogisticRegression, RandomForestClassifier, SVM, DecisionTreeClassifier, KNeighborsClassifier
  BaggingClassifier, GradientBoostingClassifier, AdaBoostClassifier

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 372-395
Received 15 April 2021; Accepted 05 May 2021.

```
LogisticRegression 0.7792207792207793
RandomForestClassifier 0.8831168831168831
SVC 0.7857142857142857
DecisionTreeClassifier 0.8636363636363636
KNeighborsClassifier 0.8181818181818182
BaggingClassifier 0.9025974025974026
GradientBoostingClassifier 0.935064935064935
AdaBoostClassifier 0.8961038961038961
```

Fig 6.7. Accuracy of each classifier on applying voting classifiers

- On applying the voting classifier, Gradient boosting classifier has achieved greatest accuracy of overall proposed model.

Evaluation of the Model:

	precision	recall	f1-score	support
0	0.96	0.94	0.95	99
1	0.89	0.93	0.91	55
accuracy			0.94	154
macro avg	0.93	0.93	0.93	154
weighted avg	0.94	0.94	0.94	154

Fig 6.8.Performance and Classification Report of GB model, which is selected using ensemble method.

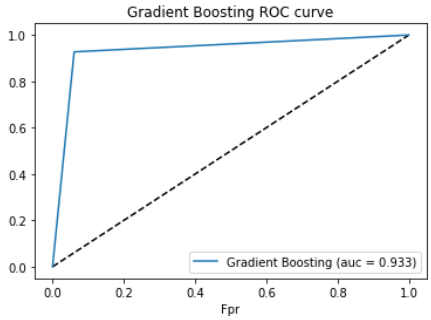

Fig 5.9.ROC –Performance metrics

Model (Classifiers)	Accuracy
GRADIENT BOOSTING	94%

Table 6.8:Accuracy report

VII. CONCLUSION

The Study on Prognostication of diabetes diagnosis is still evolving, due to the difficulty of disease diagnosis and detection synopsis and modelling. Different scholars and researchers

are working hard to find an advanced method, using multiple techniques. The results from this study will support the upcoming researchers which will give an idea of the various techniques of diabetes diagnosis. These has been accomplished by various pre-processing, feature importance and engineering, classification as classifiers, machine learning techniques and exceptional precision in detection of diabetes diagnosis On the data source, various ML tactics were implemented, and classification being performed via different algorithm, with LGBM receiving the greatest performance after Hyper Feature efficiency with 90.01 % accuracy and 0.90 CV score. Gradient boosting has reached an AUC of 0.933 efficiency by using ensemble approach of implementing voting classifier and achieved 94% of accuracy. A comparative analysis of machine learning classifier of prognostication accuracy with dataset is also being found here. With this data source, it is clear that the model increases diabetes diagnosis accuracy and consistency using a variety of approaches.

REFERENCES

1. Sisodia, D. and Sisodia, D.S., 2018. Prediction of diabetes using classification algorithms. Procedia computer science, 132, pp.1578-1585.
2. Mahmud, S.H., Hossin, M.A., Ahmed, M.R., Noori, S.R.H. and Sarkar, M.N.I., 2018, August. Machine Learning Based Unified Framework for Diabetes Prediction. In Proceedings of the 2018 International Conference on Big Data Engineering and Technology (pp. 46-50).
3. Saru, S. and Subashree, S., Analysis and Prediction of Diabetes Using Machine Learning (April 2, 2019). International Journal of Emerging Technology and Innovative Engineering, Volume 5, Issue 4, April 2019, Available at SSRN: https://ssrn.com/abstract=3368308
4. Maniruzzaman, M., Rahman, M.J., Ahammed, B. and Abedin, M.M., 2020. Classification and prediction of diabetes disease using machine learning paradigm. Health Information Science and Systems, 8(1), p.7.
5. Mujumdar, A. and Vaidehi, V., 2019. Diabetes prediction using machine learning algorithms. Procedia Computer Science, 165, pp.292-299.
6. M. F. Faruque, Asaduzzaman and I. H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, 2019, pp. 1-4, doi: 10.1109/ECACE.2019.8679365.
7. A. Mir and S. N. Dhage, "Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-6, doi: 10.1109/ICCUBEA.2018.8697439.
8. M. Jahangir, H. Afzal, M. Ahmed, K. Khurshid and R. Nawaz, "An expert system for diabetes prediction using auto tuned multi-layer perceptron," 2017 Intelligent Systems Conference (IntelliSys), London, 2017, pp. 722-728, doi: 10.1109/IntelliSys.2017.8324209.
9. Joshi, T.N. and Chawan, P.P.M., 2018. Diabetes Prediction Using Machine Learning Techniques. Ijera, 8(1), pp.9-13.
10. Tigga, N.P. and Garg, S., 2020. Prediction of Type 2 Diabetes using Machine Learning Classification Methods. Procedia Computer Science, 167, pp.706-716.
11. P. Sonar and K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 367-371, doi: 10.1109/ICCMC.2019.8819841
12. Singh, P.P., Prasad, S., Das, B., Poddar, U. and Choudhury, D.R., 2018. Classification of diabetic patient data using machine learning techniques. In Ambient Communications and Computer Systems (pp. 427-436). Springer, Singapore.
13. Aada, A. and Tiwari, S., 2019. Predicting diabetes in medical datasets using machine learning techniques. Int. J. Sci. Eng. Res, 5(2).
14. Agarwal and A. Saxena, "Analysis of Machine Learning Algorithms and Obtaining Highest Accuracy for Prediction of Diabetes in Women," 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2019, pp. 686-690.

15. Xue, J., Min, F. and Ma, F., 2020, November. Research on Diabetes Prediction Method Based on Machine Learning. In Journal of Physics: Conference Series (Vol. 1684, No. 1, p. 012062). IOP Publishing.
16. Amulya, K.J., Divya, S., Deepali, H.V. and Ravikumar, V., 2021. A Survey on Diabetes Prediction Using Machine Learning. In ICCCE 2020 (pp. 1049-1057). Springer, Singapore.
17. R. A. Canessane, R. Dhanalakshmi, B. Muthukumar, C. V. Sailaja, B. Jyothi and S. Dhamodaran, "Diabetes To Heart Disease: A Survey," 2019 Fifth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 2019, pp. 69-74, doi: 10.1109/ICONSTEM.2019.8918854.
18.Bilal, A., Sun, G. and Mazhar, S., 2021. Survey on recent developments in automatic detection of diabetic retinopathy. Journal Français d'Ophtalmologie.
19. Birjais, R., Mourya, A.K., Chauhan, R. and Kaur, H., 2019. Prediction and diagnosis of future diabetes risk: a machine learning approach. SN Applied Sciences, 1(9), pp.1-8.
20.Maniruzzaman, M., Rahman, M.J., Ahammed, B. et al. Classification and prediction of diabetes disease using machine learning paradigm. Health Inf Sci Syst 8, 7 (2020). https://doi.org/10.1007/s13755-019-0095-z

Biographies

S. Prasanth is currently pursuing his Master of Engineering in Computer Science and Engineering at Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu, India. His area of interests includes Data Science, Machine Learning, Image Processing and Big Data Analytics. He is a member of Computer Society of India.

M. Roshni Thanka is presently being an Assistant Professor in the Department of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore. She has received her Ph.D degree in Cloud Computing from Anna University, B.E and M.E degree from affiliated colleges of Anna University. Her research interest is mainly based on Networking, Cloud Computing, Artificial Intelligence and IoT. She has published papers in reputed journals and also delivered guest lectures in FDPs. She is aLife time member of Computer Society of India.

E. Bijolin Edwin is currently working as Assistant Professor at the Department of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore. India. He received his Ph.D degree in Cloud Computing from Anna University, Master of Engineering from affiliated college of Anna University, Chennai, India. His research interests include Cloud Computing, Artificial Intelligence, and Image Processing. He is a Life time member of Computer Society of India.

V. Ebenezer received his B. Tech degree in Information Technology and M.E degree in Computer Science and Engineering from Anna University, Chennai in the year of 2009 and 2012. He also received his PhD in Information and Communication Engineering from Anna University, Chennai in the year of 2020. He is currently working as Assistant professor in the department of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore Tamilnadu, India. He has published many research papers in the various International / National Conferences and Journals. His area of interests include Cloud computing, Body Area Networks, Data Structures and distributed systems etc.