Project Report

On

DIABETICS PREDICTION USING MACHINE LEARNING MODELS AND FEATURE ENGINEERIG

Submitted

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

APPLIED STATISTICS AND DATA ANALYTICS

by

PALLAVI PRASAD T

(Register No. SM23AS011)

(2023-2025)

Under the Supervision of

JESNA BABU

DEPARTMENT OF MATHEMATICS AND STATISTICS

ST. TERESA'S COLLEGE (AUTONOMOUS)

ERNAKULAM, KOCHI – 682011

APRIL 2025

CERTIFICATE

This is to certify that the dissertation entitled, DIABETICS PREDICTION USING MACHINE LEARNING MODELS AND FEATURE ENGINEERINGis a bona fide record of the work done by PALLAVI PRASAD T under my guidance as partial fulfilment of the award of the degree of Master of Science in Applied Statistics and Data Analytics at St. Teresa's College (Autonomous), Ernakulam affiliated to Mahatma Gandhi University, Kottayam. No part of this work has been submitted for any other degree elsewhere.

Date

Place: Ernakulam

JESNA BABU,

Assistant Professor,

Department of Mathematics and Statistics,

St. Teresa's College (Autonomous)

Nisha Oommen

Assistant Professor & HOD,

Department of Mathematics and Statistics,

St. Teresa's College (Autonomous),

Ernakulam.

External Examiners

1 Sangeetha Chandran

30.04.25

2 Anju N B

30 04 25

ST. TERESA'S COLLEGE (AUTONOMOUS) DEPARTMENT OF MATHEMATICS AND STATISTICS ERNAKULAM

DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of **Ms. JESNA BABU,** Assistant Professor, Department of Mathematics and Statistics, St. Teresa's College (Autonomous), Ernakulam and has not been included in any other project submitted previously for the award of any degree.

Place: Ernakulam **PALLAVI PRASAD T**

Date: **SM23AS011**

ACKNOWLEDGMENT

I must mention several individuals who encouraged me to carry out this work. Their continuous invaluable knowledgeable guidance throughout this study helped me to complete the work up to this stage.

I am very grateful to my project guide Jesna Babu for the immense help during the period of work.

In addition, the very energetic and competitive atmosphere of the Department had much to do with this work. I acknowledge with thanks to the faculty, teaching, and non-teaching staff of the department and Colleagues.

am also very thankful to HoD Ms. Nisha Oommen for their valuable suggestions, critical examination of work during the progress.

ABSTRACT

In the entire world, diabetes is one of the most common chronic illnesses. Diabetes can cause kidney failure, heart disease, eyesight loss, and other catastrophic conditions if left untreated. Patients should therefore seek an early diagnosis of diabetes in order to lessen the disease's effects. The study's data set comes from BRFSS 2015 and the open-source Kaggle platform. XG Boost, Random Forest, Decision Tree, Naïve Bayesian classifier, and logistic regression are fiveexamples of machine learning models. The feature importance method is applied. The accuracy score is used to evaluate the model, and the performance of the model with the highest accuracy score is assessed using the confusion matrix and roc curve.

ST.TERESA'S COLLEGE (AUTONOMOUS)
ERNAKULAM

Certificate of Plagiarism Check for Dissertation

Author Name	PALLAVI PRASAD T
Course of Study	MSc. Applied Statistics & Data Analytics
Name of Guide	Ms. Jesna Babu
Department	PG> Dept of Mathematics & statistics
Acceptable Maximum Limit	20
Submitted By	library@teresas.ac.in
Paper Title	DIABETICS PREDICTION USING MACHINE LEARNING MODELS AND FEATURE ENGINEERING
Similarity	10% AI-18%
Paper ID	3422096
Total Pages	40
Submission Date	2025-03-21 14:48:27

Signature of Student

Signature of Guide

Checked By
College Librarian

ST. TERESA'S COLLEGE LIBRARY ERNAKULAM

* This report has been generated by DrillBit Anti-Plagiarism Software

Contents Page number

Chapter 1

Introduction

The inability to regulate blood sugar levels is one of the traits of a disease coupled with a particular lifestyle – diabetes. Diabetes is commonly defined as either a deficiency in insulin production or inefficient utilization of insulin. The pancreas produces insulin from digested food parts during the process of digestion.

Serious complications like heart and kidney diseases, and blindness can occur as a consequence of very high sugar level within the human body. The impact of diabetes can be minimized through exercises and maintaining a healthy weight, though, the disease isn't curable in the traditional sense. Early diagnosis can significantly increase the effectiveness of treatment.

There are three different kinds of diabetes; Type 1, an autoimmune disease attacking the insulin producing cells, Type 2 where cells become resistant to insulin, and gestational diabetes which develops during pregnancy and causes higher probability of type 2 diabetes later on.

Machine learning, a subset of Artificial Intelligence (AI) is able to automatically make predictions or decisions based on data without needing precise programming. The different aspects of machine learning include supervised, unsupervised, and reinforcement learning. Supervised learning is employed to predict an outcome based on an input and output pair. It is used spam detection, disease diagnosis, or stock price prediction. In unsupervised learning, there is no labelled data, but rather methods are created to find patterns and associations. This is commonly used for customer segmentation or PCA, where dimensionality is reduced. An agent or robot learns through trial and error in reinforcement learning. Good behaviours are rewarded with positive reinforcement, while negative actions lead to punishment. This type of learning is used in game playing, robotics, or autonomous controlled cars.

The focus of this particular work is to use the machine learning model that has the capability to predict diabetes and aid in early detection. This project employs a classification algorithm, which is a technique in machine learning that attempts to separate data into distinct classes. Some commonly used types of classifications are binary, multiclass and multilabel.

In this project, at least five machine learning models are used including Naive Bayes classifier, Decision Tree, Logistic Regression, Random Forest, and XG Boost. These machine learning models are evaluated against each other to choose the best model based on the accuracy score, F1 score, confusion matrix, and ROC AUC curve. The calculation of feature importance is carried out along with the analysis of overall health as a contributing factor in predicting diabetes.

1.1Objective of the study

1. To conduct an exploratory data analysis on the data set to gain insights about the data,
1. To conduct principal component analysis for dimensionality reduction.
2. To compare Naive Bayesian classifier, Decision tree, Neural network, Random forest, Logistic Regression and XG Boost.
3. To detect the most important factor that contributes to the disease

Chapter 2

LITERATURE REVIEW

This chapter gives the recent and contributory works done in the field of diabetes prediction using machine learning and deep learning techniques.

Mujumdar and Vaidehi (2019) analysed on the topic 'Diabetes prediction using machine learning algorithms. Various machine learning algorithm were used of which Logistic Regression gave highest accuracy of 90%. Application of pipeline provided AdaBoost classifier as optimal model with accuracy of 98.8%. We have observed comparison of accuracies of machine learning algorithms with two different datasets. It is evident that the model enhances accuracy and precision of diabetes prediction with this dataset in comparison to current dataset.

Soni and Varma (2020) analysed 'Diabetes Prediction using Machine Learning Techniques'. They retrieved data set from Pima Indian Diabetes Dataset and applied various machine learning techniques and ensemble techniques. Method for utilizing different classification and ensemble learning technique where SVM, KNN, Random Forest, Decision Tree, Logistic Regression and Gradient Boosting classifiers are used for which random Forest provided the highest accuracy.

Rani (2020) examined 'Diabetes prediction using machine learning'. The objective of this project is to create a system that can make early prediction of diabetes for a patient with greater accuracy by integrating the output of various machine learning methods. The algorithms such as K nearest neighbour, Logistic Regression, Random Forest, Support vector machine and Decision tree are employed. Experiments are conducted on John Diabetes Database. Experimental findings decided the effectiveness of the designed system with a achieved accuracy of 99% employing Decision Tree algorithm.

Hasan et.al (2020) conducted a study on 'Diabetes Prediction Using ensemble of Die rent Machine Learning Classifiers'. Here the suggested ensemble classifier is the optimal performing classifier with the sensitivity, specificity, false omission rate, diagnostic odds ratio, and 0.789, 0.934, 0.092, 66.234, and 0.950 respectively that performs better compared to the state-of-the-0.789, 0.934, 0.092, 66.234, and 0.950 respectively. prediction wherein the outlier rejection, missing values filling, standardization of data, feature selection, K-fold cross-validation, and various Machine Learning (ML) classifiers (k-nearest Neighbour, Decision Trees, Random Forest, AdaBoost, Naive Bayes, and XG Boost) and Multilayer Perceptron (MLP) were utilized. Weighted ensemble of various ML models is also suggested here, to better predict diabetes where the weights are learned from the respective Area Under ROC Curve (AUC) of the ML model. AUC is selected as the performance metric, which is maximized with hyperparameter tuning using the grid search method. All the experiments, in this paper, were performed under identical experimental conditions on the Pima Indian Diabetes Dataset. Here the suggested ensemble classifier is the optimal performing classifier with the sensitivity, specificity, false omission rate, diagnostic odds ratio, and 0.789, 0.934, 0.092, 66.234, and 0.950 respectively that performs better compared to the state-of-the-0.789, 0.934, 0.092, 66.234, and 0.950 respectively.

Xue et al. (2020) conducted a project on ‘Research on Diabetes Prediction Method Based on Machine Learning’. The data set in this article comes from the open-source standard test data set website UCI. The data set was obtained by direct questionnaires from 520 patients at the Sylhet Diabetes Hospital in Sylhet, Bangladesh, and was approved by doctors. The data set is divided into 17 attributes including age, gender, polyuria, depression, Sudden weight loss、Weakness、Polyphagia、Genital thrush、Visual blurring、Visual blurring、Itching、Irritability、Delayed healing、Partial paresis、Muscle stiffness、Alopecia、and Obesity. Although, in this project the naive Bayes classifier is the most popular classification algorithm, the final accuracy rate on our data set is only 93.27%. SVM has the highest accuracy rate, with an accuracy rate of 96.54%. The accuracy of Light GBM is only 88.46%. Therefore, SVM has the highest accuracy through the confusion matrix evaluation test.

Khanam and Foo ,(2021) published an article on the topic ‘A comparison of machine learning algorithms for diabetes prediction ‘.es is Datamining, machine learning (ML) algorithms, and Neural Network (NN) techniques are employed to predict diabetes in our work. We employed the Pima Indian Diabetes (PID) dataset for our work, which was gathered from the UCI Machine Learning Repository. The dataset comprises details of 768 patients and the respective nine distinct attributes. Seven ML algorithms were employed on the dataset to predict diabetes by them. They identified that the model with Logistic Regression (LR) and Support Vector Machine (SVM) is suitable for diabetes prediction. They constructed the NN model using a different hidden layer with different epochs and saw that the NN with two hidden layers yielded 88.6% accuracy.All models are yielding good results for some measures such as accuracy, precision, recall, and F-measure. All models gave an accuracy of more than 70%. LR and SVM gave around 77%–78% accuracy for train/test split and K-fold cross-validation technique. We also applied the NN model for diabetic prediction of PIDD. We applied the 1, 2, 3 hidden layers in the neural network model changing the epochs 200, 400, 800. Hidden layer 2 with 400 epochs gave 88.6% accuracy, which is the highest accuracy among our executed model for PIDD. Of all the models proposed, the NN with two hidden layers is found to be the most effective and promising in analysing diabetes with an accuracy level of around 86%.

Ramesh et al, (2021) conducted a project on the topic ‘A remote healthcare monitoring framework for diabetes prediction using machine learning ‘. Remote patient monitoring may enable effective intervention and treatment paradigms through existing technology. This paper presents an end-to-end remote monitoring system for automated diabetes risk prediction and management, based on personal health devices, smart wearables and smartphones. A support vector machine was implemented for diabetes risk prediction from the Pima Indian Diabetes Database, following feature scaling, imputation, selection and augmentation. This work attained the performance parameters of accuracy, sensitivity and specificity values of 83.20%, 87.20% and 79% respectively using the tenfold stratified cross validation technique, which is in comparison to current techniques. Patients have the ability to utilize several healthcare devices, smartphones and smartwatches to monitor vital parameters, prevent the development of diabetes and close the communication loop with healthcare professionals. The given framework allows the healthcare professionals to make decisions informed by the recent diabetes risk prediction and lifestyle data while achieving unobtrusiveness, lowered cost, and vendor interoperability Ahmed (2022) released an article titled 'Prediction of Diabetes

Empowered with Fused Machine Learning'. The dataset adopted in this research is from the UCI Machine Learning Repository.

At the Data Acquisition phase, a dataset with sufficient features can be employed to predict diabetes. Here the number of instances is 520, and consists of 17 attributes based on diabetic symptoms. on. In this study, a machine learning based diabetes decision support system has been proposed using decision level fusion. Two common machine learning approaches are combined in the proposed model using the fuzzy logic. The suggested fuzzy decision system has attained accuracy of 94.87, which is greater than the other available systems.

Febrian et al (2023) did a project on the topic ‘Diabetes prediction using supervised machine learning’. The data in this study uses a public dataset originating from Kaggle that can be access in with the name of the dataset used using a public dataset originating from Kaggle with the dataset name Pima Indians Diabetes Database. The Pima Indians diabetes database dataset consists of 768 data which is divided into 8 attributes and 2 classes with a total of class 1 (268) and class 0 (500). The study concluded by comparing two k-Nearest neighbour algorithms and the Naive Bayes algorithm to predict diabetes based on several health attributes in the dataset using supervised machine learning. According to the results of our experiments and evaluating algorithm using Confusion Matrix, the Naive Bayes algorithm outperforms KNN, with an average value of 76.07 percent accuracy, 73.37 percent precision, and 71.37 percent recall in Naive Bayes and an average value of 73.33 percent accuracy, precision 70.25 percent, and recall of 69.37 percent in KNN. As a result, it can be concluded that the Naive Bayes algorithm is preferable to the KNN algorithm for predicting diabetes using the Pima Indians dataset. For future research can be done by adding other algorithm like neural network and other techniques in order to produce an accuracy value and better precision also by Adding technique Particle Swarm Optimization for optimize the results and using application program development.

Tasin et al (2023) conducted a study on ‘Diabetes prediction using machine learning and explainable AI techniques. The open-source Pima Indian and a private dataset of female Bangladeshi patients have been used in this work. SMOTE and ADASYN preprocessing techniques have been applied to handle the issue of imbalanced class problems. This research paper reported different performance metrics, that is, precision, recall, accuracy, F1 score, and AUC for various machine learning and ensemble techniques. The XG Boost classifier achieved the best performance with 81% accuracy and an F1 score and AUC of 0.81 and 0.84, respectively, with the ADASYN approach. Next, the domain adaptation technique has been applied to demonstrate the versatility of the proposed prediction system. Finally, the best-performed XG Boost framework has been deployed into a website and smartphone application to predict diabetes instantly.

Chapter 3
Materials and Methods

3.1 Dataset

This study used dataset from open-source website Kaggle which was collected by BRFSS 2015.The data set contains 253680 survey responses which was cleaned by BRFSS.The target variable is classified into 3 classes: - no diabetes ,pre diabetes and diabetes. The original dataset contains 441455 responses from individuals of America. These features are either questions directly asked of participants or estimated based on individual responses. The target class is classified into pre diabetic, diabetic and non diabetic

3.2 Machine learning

A branch of artificial intelligence that creates algorithms to analyse patterns and correlations in data sets to get meaningful insights about the data and to forecast and decide without programming in an explicit manner. By training the data for classification or prediction, machine learning helps to makes meaningful insights about the data. The result can be utilised by businesses, health sector, educational institutes, finance sector reduce and so on to reduce risks and make decisions in the future.

Types of machine learning

3.2.1 Supervised learning

Supervised learning algorithm trains the data using labelled dataset. Each training consists of an input(features) and corresponding(label). The model maps input to output and makes prediction on the new data.

Regression and classification are two examples of supervised learning algorithms. Regression to make numerical predictions. There are different kinds of regression such as linear, multiple, polynomial, logistic, and so on. Linear maps one independent variable to the dependent variable where as multiple has one independent and multiple dependent variables.

Classification: The output variable is classified as a category variable. Various classification techniques include random forest, decision tree, logistic regression, and others. In this the predicted class is already given. The model trains the data and then predict accordingly.

3.2.2 Unsupervised learning

In unsupervised learning, an algorithm analyses unlabelled data to find patterns, structures, or correlations among the data. There are no clear target variables, unlike supervised learning.

Finding hidden structures or groupings within the data is the main objective of un supervised learning algorithms

Types of unsupervised algorithms: -

Clustering - Clustering is a technique used to group similar data points based on their features. It is widely used for clustering data points with similar characteristics, pattern recognition, and anomaly detection. Common clustering algorithms are K nearest, DBSCAN, hierarchical, KNN, K means and so on.

Dimensionality reduction –These techniques are used to reduce the number of features in a dataset while retaining useful information to perform analysis types of tasks. This is useful when dealing with high-dimensional data. It makes the data easier to visualize and analyse. Examples of dimensionality reduction include principal component analysis, recurrent feature elimination and so on.

Association rule – This learner is interested in finding some sort of relationships between variables. It is commonly used in market basket analysis, where businesses analyses purchasing behaviour of the customer. One of the common associationrules is apriori algorithm.

1.2.3 Semi supervised learning

In machine learning, semi-supervised learning is a combination of supervised and unsupervised learning. To increase accuracy, a significant amount of unlabelled data is combined with a small amount of labelled data. This is useful in situations when a high volume of unlabelled data is available but labelling the data is costly or time-consuming.

Semi supervised learning is generally used in cases where labelled data is costly or difficult to attain. Very few labelled samples can be found in the real world. where as unlabelled data is widely available and less costly. Using a very small amount of labelled data increases performance, even when unsupervised learning alone might not be able to identify significant trends. This combines between learning that is monitored and learning that is unsupervised. This benefits from large-scale unlabelled data while using a limited set of labelled data to guide learning.

Some applications are Image and speech recognition, fraud detection , natural language processing (NLP),healthcare and medical diagnosis and so om

1.2.4 Reinforcement learning

RL is a form of machine learning where an agent learns through goal-based actions in a specific environment. The agent applies the learned actions and gets rewards or punishments for them. In reinforcement learning, models learn without pre-existing labels and receive positive feedback or penalties aiding in outcome-based decision making. This is different from supervised learning where models rely on labelled data.

Fundamental Principles of Reinforcement Learning

Agent - The individual making the interactions with the environment.

Environment - The world that the agent works within and from which feedback is obtained.

State (S) - An environment that describes the agent at any period in time.

Action (A) - Considered in terms of the state, defined as the set of all possible actions for the agent.

Reward (R) - In basic terms, reward can be understood as the outcome perceived from an action taken which determines further learning.

Positive reward - Encouraging feedback.

Negative reward - An action discouraging feedback.

Policy (π) - A set of rules in which the agent executes the defined set of actions.

Deterministic Policy - A policy in which, given a state, a specific action is guaranteed to executed.

Stochastic Policy - A policy which is based on probability as an approach to any action.

Value Function (V) - Predicts how favourable it is to be in a specific state.

Q – function (Q) - Anticipates the worth of performing an action in a specific state.

Reinforcement Learning can be applied to robotics, computer games, autonomous vehicles, finance and stock market trading, and so on.

1.3 Exploratory data analysis

This is a vital step in data preprocessing and is crucial in the analysis of the structure, patterns, and relationships inherent in the dataset even before any machine learning models have been implemented. The process starts off with loading the dataset, after which an overview of it is prepared with the help of descriptive statistics, data types, and missing values. One of the most important steps is handling missing values, as this problem will certainly affect any machine learning models you plan to design. Common methods of address with missing values include simply dropping them, filling with dispersion means such as the mean or median, and even using predicted values. Attention to outlier detection is important to identify where there are unusual values or areas of noise. The removed part should be analysed. But in some cases, like fraud detection, the abnormal spots are useful in understanding the unexpected actions. By employing techniques such as the interquartile range (IQR), Z-score and several others, including visual ones like boxplots and scatter plots, it is possible to identify outliers and get rid of them. EDA, together with graphical analyses, includes the knowledge of the distribution of numerical variables through histograms and density plots and categorical variables by means of bar charts and frequency tables. Examination of the relationship between numerical variables is known as correlation analysis, which can be made visible through the use of heatmaps, a commonly used visualization tool.

Another major part of exploratory data analysis is feature engineering which comprises creation of new features, encoding categorical features and transforming variables. As a last step, EDA is complemented by visualizations such as histograms, pair plots, bar charts, which show the relationships within data, correlations, and outliers in a much deeper level. In summary, EDA is one of the most important techniques of data analysis or machine learning because it is the first step of the process of any machine learning projects.

1.4 Classification models

There are different classification models that helps in classifying the data. Various models used in this study are naïve Bayesian classifier, decision tree, XG boost, random forest and logistic regression.

3.4.1 Naïve Bayesian classifier

The Naïve Bayes classifier is a probabilistic machine learning algorithm based on Bayes' Theorem. This is used primarily in classification type tasks. Naïve Bayes often performs well in applications like spam detection, sentiment analysis ,and medical diagnosis.

Naïve Bayes is based on Bayes' Theorem. Bayes theorem describes the probability of an event based on the knowledge that some event has already been occurred:

$P(A|B) = (P(B|A) * P(A))/P(B)$

Were

$P(A|B)$ = Posterior probability

$P(B|A)$ = Likelihood

$P(A)$ = Prior probability

$P(B)$ = Evidence

3.4.2 XG Boost

XG Boost (Extreme Gradient Boosting) is a powerful machine learning algorithm based on gradient boosting, designed for speed, accuracy, and scalability. It is generally used for both classification and regression tasks. XG Boost is designed to handle large datasets efficiently

and is generally effective in reducing errors by combining multiple weak learners into a single strong predictive model.

By minimizing the residual errors from earlier trees, XG boost creates an ensemble of decision trees through a technique called gradient boosting. By using regularization techniques, parallel computation, and optimized tree learning, XG Boost overcomes the normal boosting methods and becomes one of the most accurate and efficient algorithms available today. The main concept of XG boost is to reduce error and to improve prediction capacity by iteratively training trees on the mistakes made by the prior trees. XG Boost is a flexible option for various machine learning problems because it performs well with both structured and unstructured data and is suitable for handling large data sets.

1.4.3 Decision Tree

Decision Tree is widely used machine learning algorithm that can be utilized for both classification and regression problems. It is a decision treelike model. Every internal node corresponds to a feature test, every branch corresponds to an outcome, and leaf node corresponds to a final decision or prediction or a class label. Decision Trees are interpretable and is appropriate for handling numerical and categorical data.

The decision Trees divide the dataset into subtrees according to feature values, employing a process called recursive partitioning. The algorithm chooses the optimal feature in which to split the data at each step, in order to build subtrees containing most of the data from one class. This procedure repeats until the stopping condition is reached, i.e., maximum depth is attained or a node contains too few samples. This is appropriate when dealing with huge data sets. Components of decision tree are: -

Root Node – The starting point of the tree, representing the entire dataset.

Internal Nodes – Represent feature-based decision points.

Branches – Represent possible outcomes of a decision.

Leaf Nodes – output or decision (class label or continuous value).

3.4.4 Random Forest

Random Forest is a supervised machine learning algorithm that can be useful for classification and regression tasks. It is an ensemble method that makes multiple decision trees and combines their outputs to improve accuracy and reduce overfitting. The idea behind the algorithmis a technique named Bootstrap Aggregation (Bagging), where each tree is trained on a random subset of the dataset and selects a random subset of features at each split. For classification, the final prediction is determined by majority voting. For regression, the final prediction is obtained by averaging the outputs of all trees. One of the limitations of random forest is that

though it can handle large data set, the processing time is slow since each label has to be passes through multiple trees.

3.4.5 Logistic Regression

Logistic Regression is a supervised machine learning algorithm mainly used for classification tasks. It is a statistical model that is used to predicts the probability of a categorical output. It is widely used for binary but can be used for multi-class classification using techniques like One-vs-Restor , softmax Regression., Logistic Regression uses a sigmoid function that maps predicted values to probabilities between 0 and 1(in case of binary classification).

Types of Logistic Regression

Binary Logistic Regression - Classifies between two categories

Multinomial Logistic Regression - Handles multiple classes without having any order.

Ordinal Logistic Regression - Used when ordinal dependent variables are based on one or more order.

1.5 Model Evaluation

To ensure the correct and optimal performance of a model, it is important to evaluate the model. This is to identify the best model. The methods used here are accuracy score, Confusion matrix, ROC curve and F1 score.

3.5.1 Accuracy Score and F1 score

The percentage of all categories that were accurate, whether positive or negative, is known as accuracy.

It is defined by the formula: -

Accuracy = Correct Classification / Total Classification

$$= (TP+TN)/(TP+TN+FP+FN)$$

Were,

TP represents True positive

TN represents True Negative

FP represents False Positive

FN represents False Negative

It is often used as the default and widely used evaluation metric since it makes use of all four outcomes of a confusion matrix.

F1 Score

F1 score is one of the important metrics in machine learning. F1 score provides a balanced measure of a model's precision and recall, particularly when class distribution is unbalanced. The F1 Score is obtained from the harmonic mean of precision and recall. This is an important metric which can be commonly used in classification tasks to evaluate the model performance.

Precision:

It refers to the ratio of correct true Positives and the sum of true positive and false positive. It can be interpreted as the measure to find the accuracy of the positive predictions.

Precision = True Positive / (True Positive +False Positive)

Recall (Sensitivity or True Positive Rate):

It is also known as sensitivity or true positive rate. It is the measure of the ratio of true positive and the sum of true positive and false negatives.

Recall = True Positive / (True Positives + False Negatives)

3.5.2 Confusion Matrix

A confusion matrix is a chart used to compare actual outcomes with the predictions of a classification model in order to show just how well it is functioning. There are four classes used to categorize the predictions and these are true positives and true negatives as correct predictions for both classes and false positives and false negatives as incorrect predictions. This helps determine where the model requires improvement. The matrix displays the number of instances produced by the model on the test data.

True Positive (TP): The model correctly predicted a positive outcome

True Negative (TN): The model correctly predicted a negative outcome

False Positive (FP): The model incorrectly predicted a positive outcome. Also called as a Type I error.

False Negative (FN): The model incorrectly predicted a negative outcome. Also called as a Type II error.

A confusion matrix helps is used to see how well a model is working by showing correct and incorrect predictions.

The confusion matrix for binary classification is show in the table 3.1

Table 3.1

	Predicted	Predicted
Actual	True Positive	False Negative
Actual	False Positive	True Negative

Since the target class is multi class, the confusion matrix for multi class is in the table 3.2

Table 3.2

	Predicted	Predicted	Predicted
Actual	True Positive (TP)	False Negative (FN)	False Negative (FN)
Actual	False Negative (FN)	True Positive (TP)	False Negative (FN)
Actual	False Negative (FN)	False Negative (FN)	True Positive (TP)

Metrics based on Confusion Matrix Data are

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Precision = TP/(TP+FP)

Recall = TP/(TP-FP)

F1Score = (2*precision*recall)/ (precision + recall)

3.5.3 ROC – AUC curve

ROC -AUC curve is a graph utilized to assess the performance of binary classification models. It measures the True Positive Rate against the False Positive Rate at various thresholds to analyse the ability of the model to discriminate between two classes like positive and negative.

It gives a graphical view of the model performance to differentiate between two classes like positive class for the presence of a disease and negative class for the absence of a disease.
Key Terms in AUC-ROC:

True Positive Rate-The ratio of correctly predicted positive instances.

False Positive Rate - The ratio of incorrectly predicted negative instances.

Specificity: The proportion of actual negatives correctly identified by the model.

Sensitivity/Recall: The proportion of actual positives correctly identified by the model.

A typical ROC – AUC curve is shown by the figure 3.1

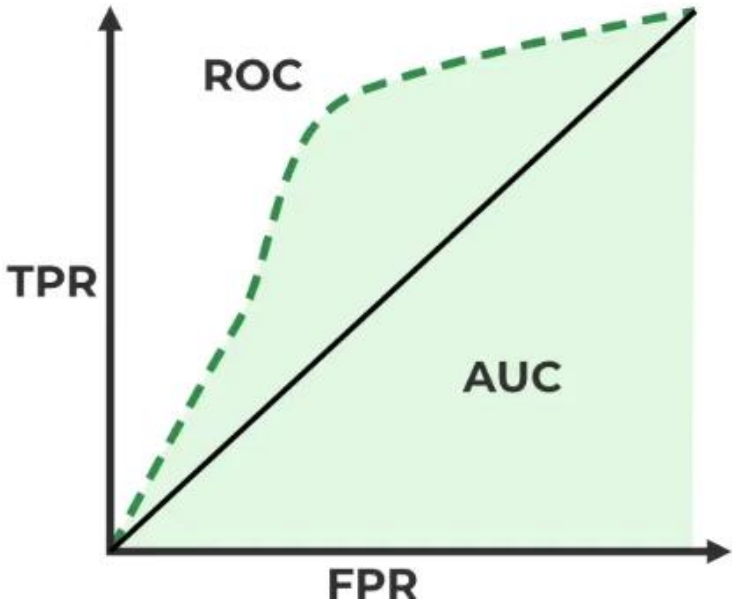


Figure 3.1

An example for ROC – AUC curve in multiclass is shown in figure 3.2

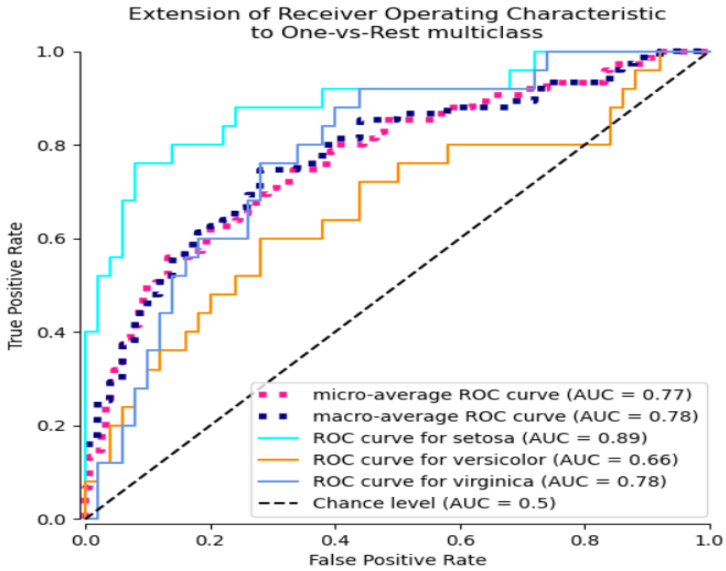


Figure 3.2

3.6 Principal Component Analysis

PCA is a statistical technique developed by mathematician Karl Pearson in 1901. The idea is to map high-dimensional data into a lower-dimensional space such that the variance (or spread) of the data in the new space is maximized. This is achieved by retaining the most significant patterns, features and relationships in the data.

Steps in PCA are

Standardization of data - Ensure that all features are on the same scale. Standardizing our dataset to ensures that each variable has a mean of 0 and a standard deviation of 1.

Find Relationships -Compute how features vary together from a covariance matrix. Covariance quantifies the degree of shared variability between two or more variables, how much they vary relative to one another.

Find the Principal Components - PCA finds new axes along which the data is most spread out.

1st Principal Component (PC1) using eigen values and eigen vectors

2nd Principal Component (PC2).

These directions are determined by Eigen values and Eigenvectors where: eigenvectors, and their significance is ranked by Eigen values

Select the top Directions & Reduce Data –

This is an application of unsupervised learning algorithm ,where the machine learns without any labels from the data set. It is commonly used in exploratory data analysis and machine learning to reduce the dimension of large datasets without losing critical information.

Scree Plot

A common graphical visualisation method for determining the number of principle components a scaled a scree plot. A scree plot is a graph of eigenvalues against the corresponding Principal component number. The number of PCs retained is then subjectively determined by locating the point at which the graph shows a slope or an elbow.

An example of scree plot is shown in the figure 3.3

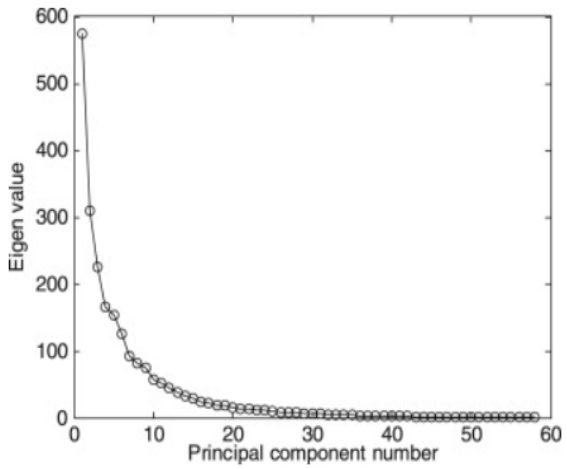


Figure 3.3

3.7 Feature Importance

A feature importance plot is a visual representation of the importance of various variables in a dataset when utilizing a machine learning model to predict results. In data analytics, statistics, and machine learning, this plot is frequently used to help determine which features in a given dataset are most important for the model's prediction. In addition to providing insight into the relationship between features and target variables, feature importance plots can also be used to explain model results . An example of feature important plot is given in the below figure.

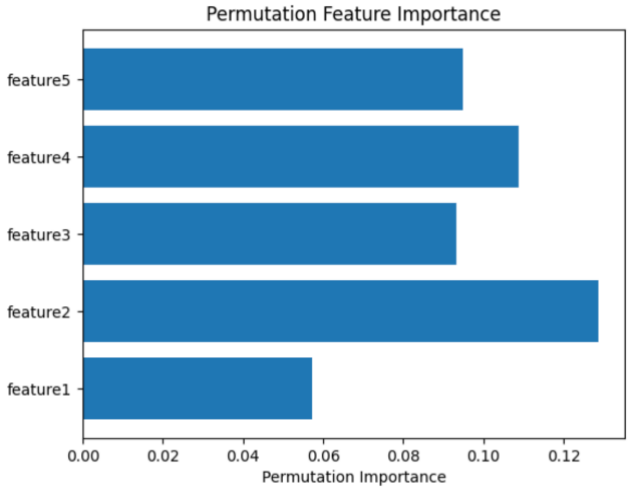
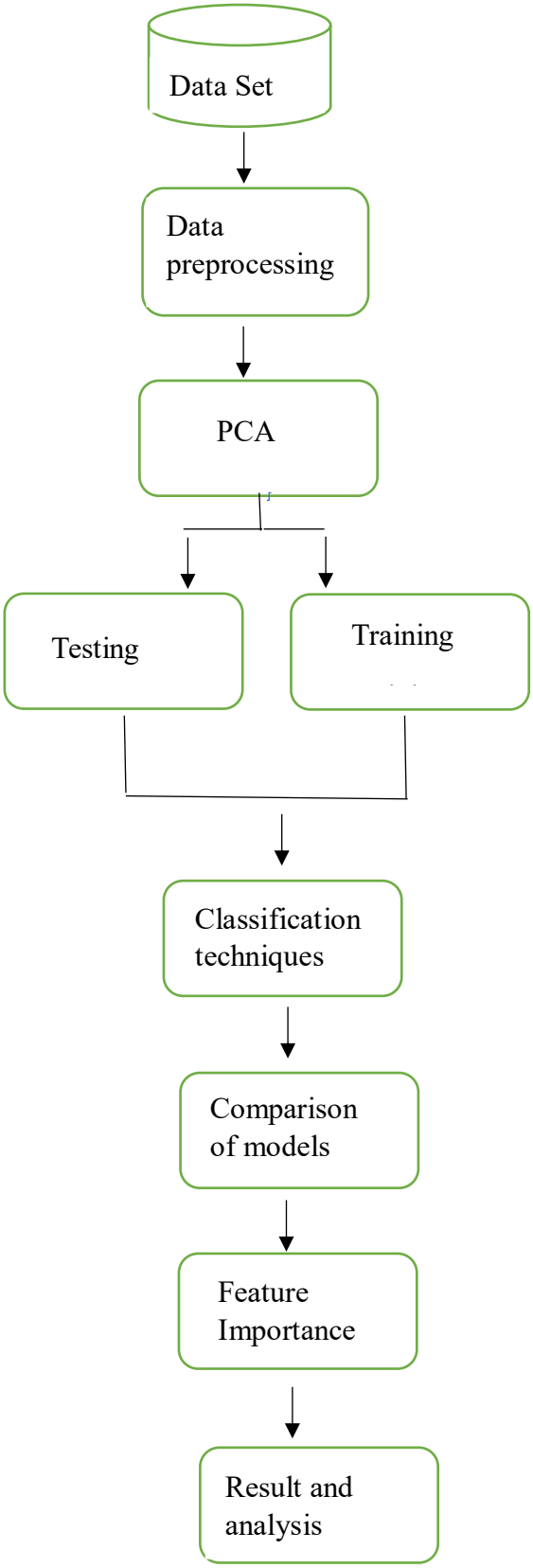

Permutation Feature Importance
feature5
feature4
feature3
feature2
feature1
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Permutation Importance

Figure 3.4

3.7 Methodology



Chapter 4

Data Description and Exploratory Data Analysis

4.1 Attributes

For this project a Kaggle data set of 22 variables for diabetes prediction is used. The target variable in the study is 'Diabets_012' and the rest of the variables are:

If the patient has diabetes or not categorized into 0 ,1 and 2 indicating diabetes, pre diabetes and no diabetes

High Chol which is in 0,1 format where 0 indicate no high cholesterol and 1 indicate high cholesterol.

. Chol Check which is in 0,1 format where 0 indicate no cholesterol check 5 years and 1 indicate cholesterol check had happened in 5 years.

. BMI of the patient.

If the patient is smoker or not.

Ever had stroke or not.

. Ever had coronary heart disease or myocardial infraction or not.

Is the patient physical active in the past 30 days or not.

Does the patient consume fruits or not.

Does the patient include vegetables in their diet or not.

Is the patient an alcoholic or not.

Does the patient have any kind of health care coverage or not

Was there a time in the past 12 months when the patient needed to see a doctor but could not because of no cost

General health in a scale of 1-5.

Mental health of the patient.

Physical health of the patient.

Does the patient struggle to walk or to climb stairs or not.

Sex of the patient.

Age of the patient

Education level of the patient.

Income of the patient.

Does the patient have high BP or not.

The data set is a cleaned data set from BRFSS 2015 from 253880 survey responses.

4.2 Exploratory Data Analysis

Conducting exploratory data analysis for ideas regarding the data set.

4.2.1 Data Cleaning and Analysis

The data is checked for any null values.

Since it is already cleaned it has no null values. Anomaly detection using isolation forest is conducted and eliminated 76791 extreme or rare cases.

Conducting exploratory data analysis for ideas regarding the data set. For deeper insight look for the data visualization.

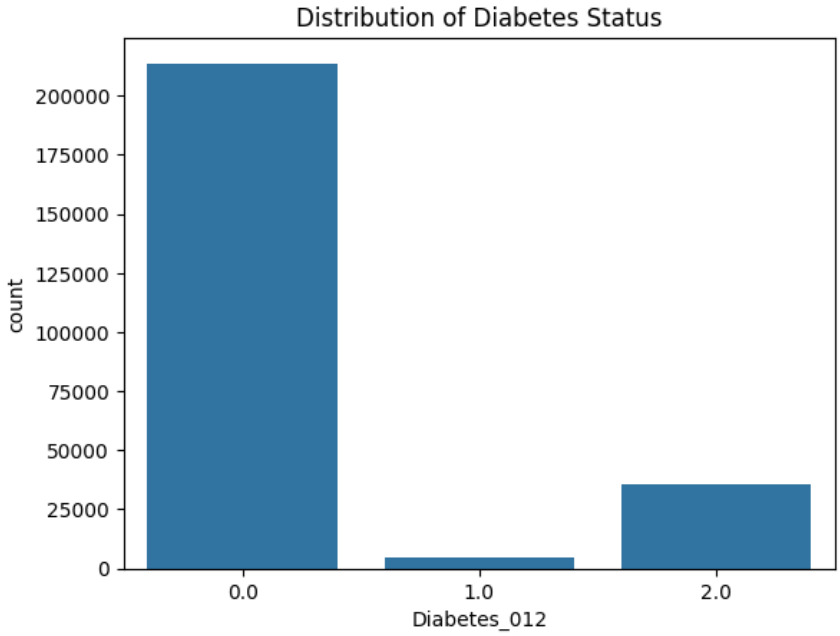


Figure 4,1

Figure 4.1 is a graph displaying a bar chart for the diabetes distribution with diabetes on the x axis and diabetes count on the y axis. About 20,000,00 people do not have diabetes. Between 25,000 and 50,000 cases are non-diabetic, and the number of prediabetes cases is too low.

Figure 4.2 shows the box plot of diabetes vs BMI.

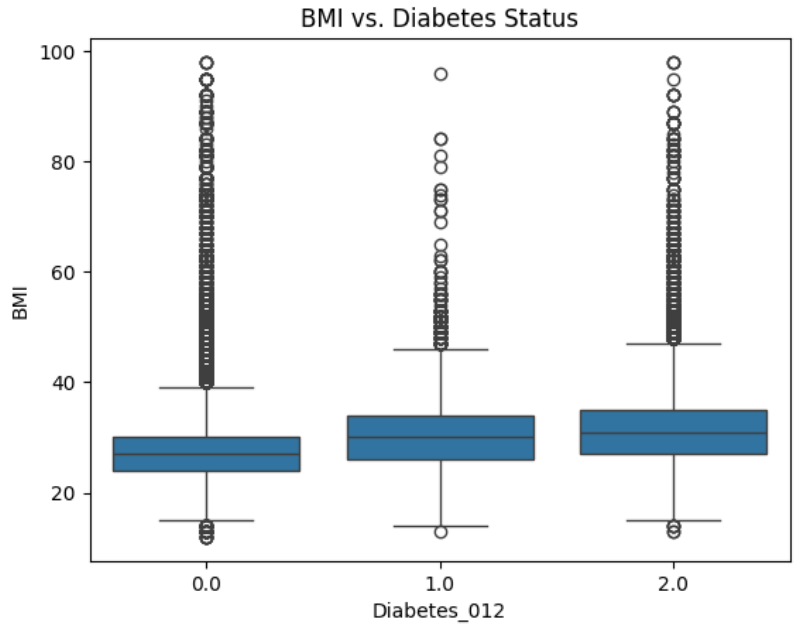


Figure 4.2

This graph is a box plot showing the relationships between BMI taken on the x axis and Diabetes taken on the y axis. Each box plot summarizes the distribution of BMI vs diabetes.

In a box plot, box indicates the inter quartile range. The line inside the box represents the median. The whisker is extended from the box to the farthest data points within 1.5 times the inter quartile range. The points outside the whisker are outliers.

People in all diabetes status categories appear to have a wide range of BMI values, according to the box plots. Individuals who are diabetic, have a less dispersed distribution and a slightly higher median BMI than those with non diabetic or prediabetic. All groups contain outliers, which suggests that some people have BMIs that are significantly higher than the normal range for their particular diabetes condition

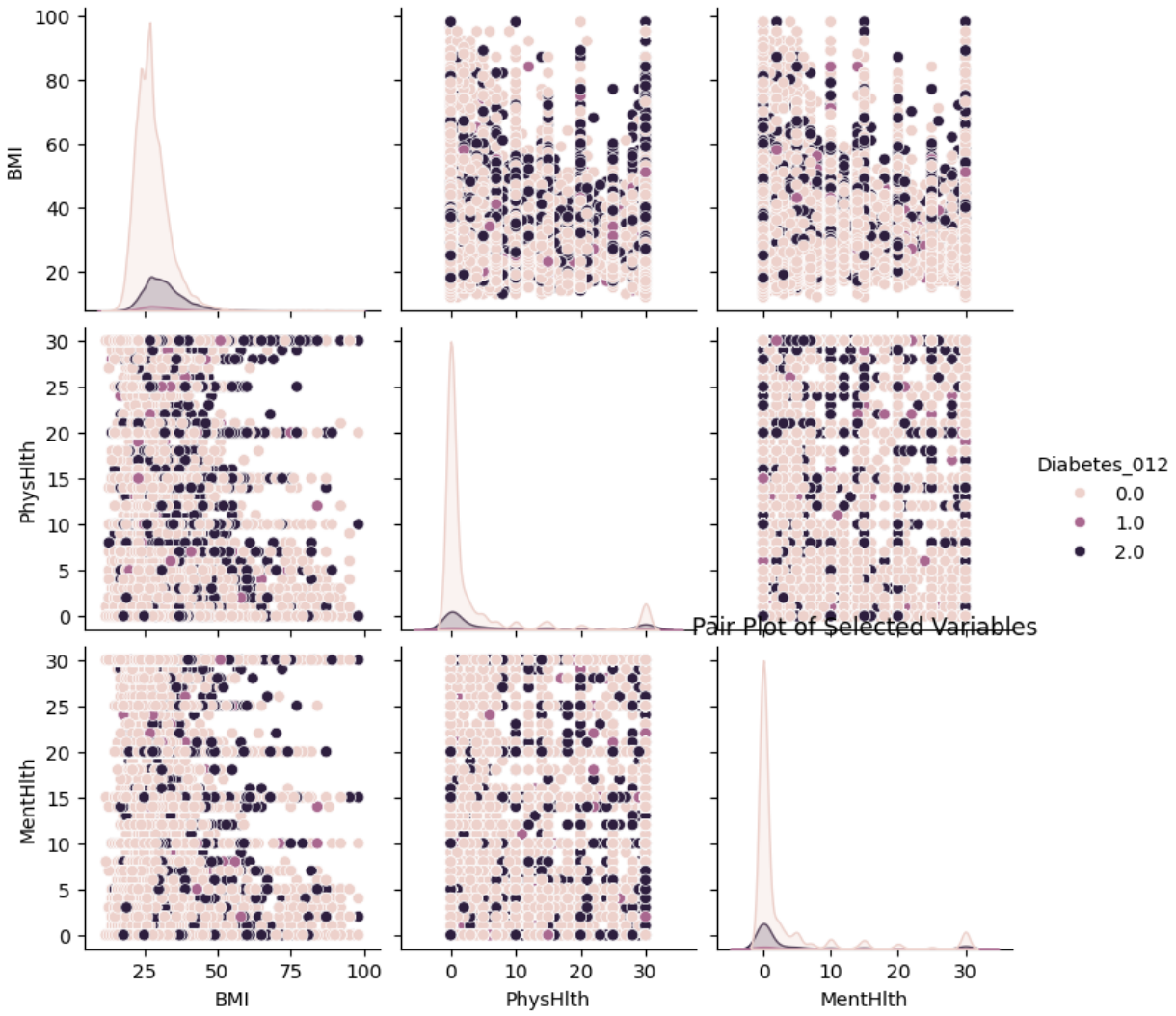


Figure 4.3

The picture displays a pair plot that was produced using feature engineering and machine learning techniques to visualize the relationships between various variables in a diabetes prediction dataset. The plot displays histograms for the distribution of variable along the diagonal as well as scatter plots for every variable pair. The variables are Diabetes (Diabetes statuses, non diabetic, pre and diabetic), Physical Health, and BMI With points coloured based on the Diabetes category. The scatter plots display the distribution of data points for every pair of variables whereas the frequency distribution of every variable is shown by histograms.

Chapter 5

Result And Discussions

5.1 Model Building

5.1.1PrincipalComponent Analysis

Principle component analysis is a dimensionality reduction technique. Given below is the scree plot for visualizing PCA in the figure 5.1

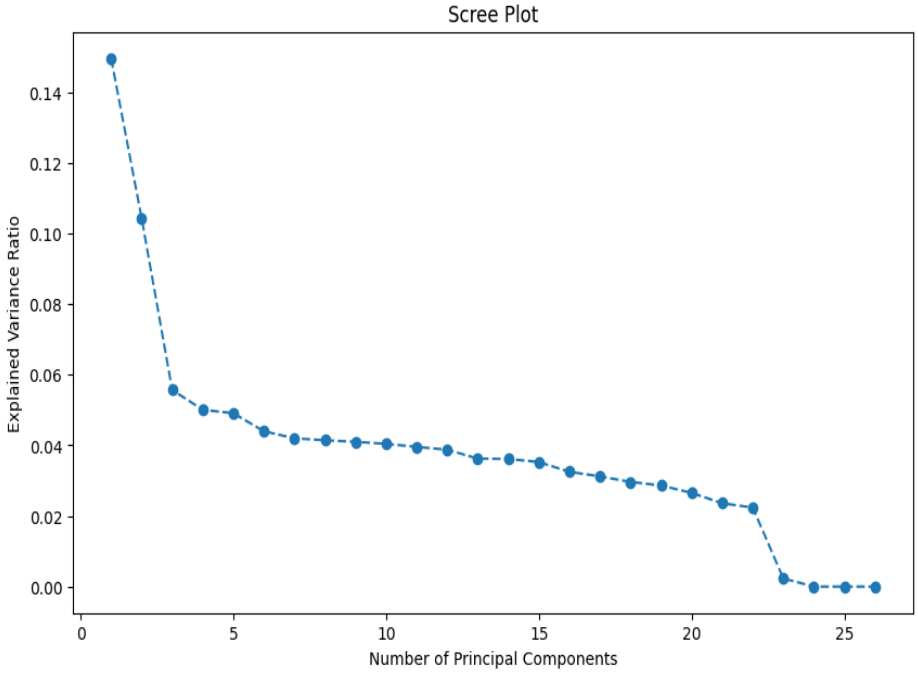


Figure 5.1

The first three to four principal components would capture the majority of the data's significant information. From this it is clear that there are more than 25 principal components.

5.1.1 Naïve Bayesian Classifier

Naïve Bayesian classifier has an accuracy of 0.48 and F1 score of 0.59. The confusion matrix of Naïve Bayesian shows that the model correctly predicts 14900 class 0 cases ,17602 true instances for class 1. Further it shows that 399 cases as wrongly predicted as class 0 when they belonged to class 1 and 22 cases were wrongly predicted as class 0. THE ROC -AUC curve shows that though the model predicts class 0 and 2 well but it fails in predicting class 1.So, with the low accuracy score and F1 score and from insights gained from ROC -AUC curve, this model is not a best fit to predict diabetes. Confusion matrix and ROC – AUC curve is given below. figure5.2 and 5.3 shows the confusion matrix and ROC -AUC curve of Naïve Bayesian classifier.

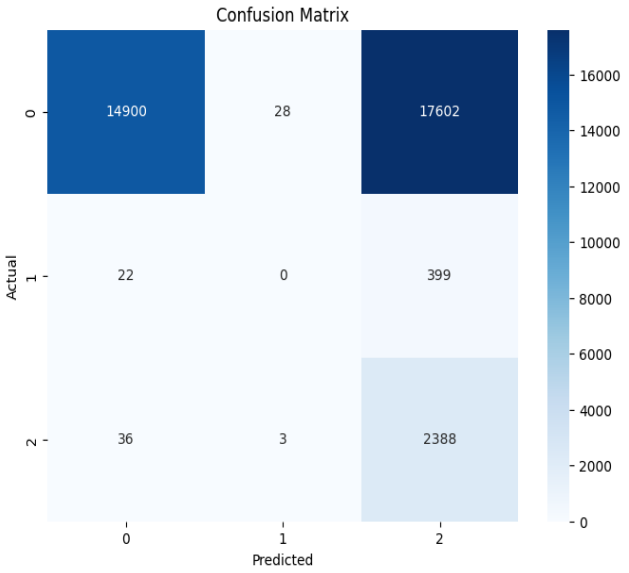


Figure 5.2

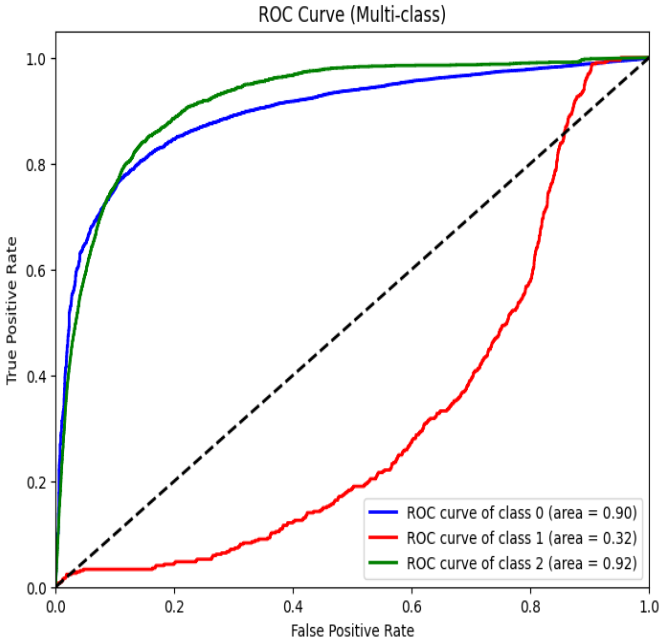


Figure 5.3

5.1.2 Decision Tree

Decision Tree classifier has an accuracy and F1 score of around 97percent. The confusion matrix of decision tree indicates that 32188 cases were correctly classified as 0 ,127 as 1 and 2137 as 2. The off-diagonal elements of confusion matrix indicate misclassification. That is 166 instances were incorrectly predicted as class 0, but were actually class 1, about 134 instances were incorrectly predicted as class 0, but were actually class 2. About 208 instances were incorrectly predicted as class 1, but were actually class 0. About 156 instances were incorrectly predicted as class 1, but were actually class 2. About 134 instances were incorrectly predicted as class 2, but were actually class 0. From ROC -AUC curve, it is understood that class 0 and 2has AUC score of 0.94 which indicates that it performs well for predicting non diabetic and pre diabetic but fails at predicting class 1. Figure 5.4 and 5.5 shows the confusion matrix and ROC – AUC curve of decision tree.

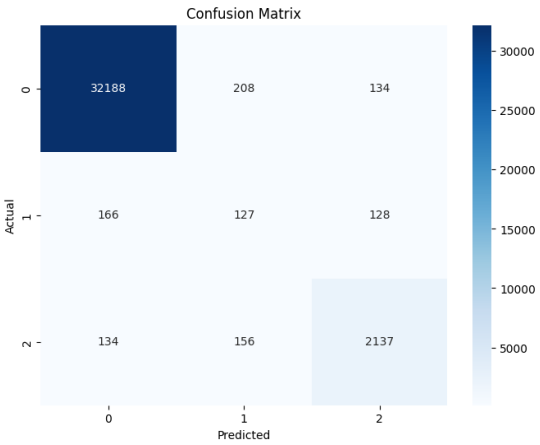


Figure 5.4

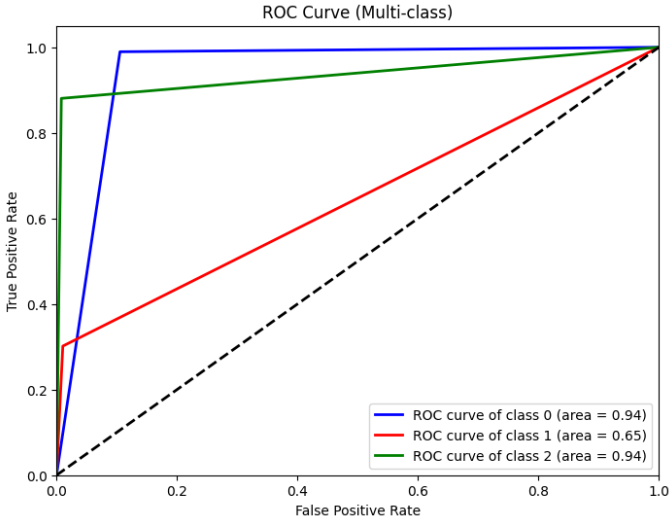


Figure 5.5

5.1.3 XG Boost

The XG Boost classifier has an accuracy and F1 score of around 99 percent.

From the confusion matrix, it is clear that the model correctly predicted 32529 were as 0,201 as 1 and 2418 as 2. The off diagonals elements give the misclassifications which is very low. The ROC -AUC shows that model is 100 percent accurate for predicting class 0 and 2 and 99 percent for class1. This indicates that the model works well for predicting the disease. The confusion matrix and area under the curve is shown below.

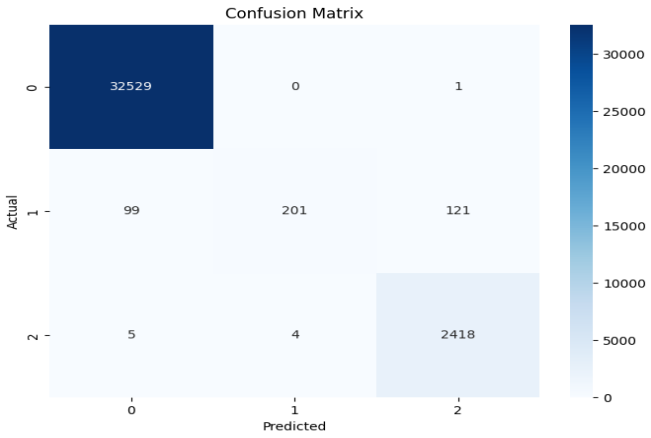


Figure 5.6

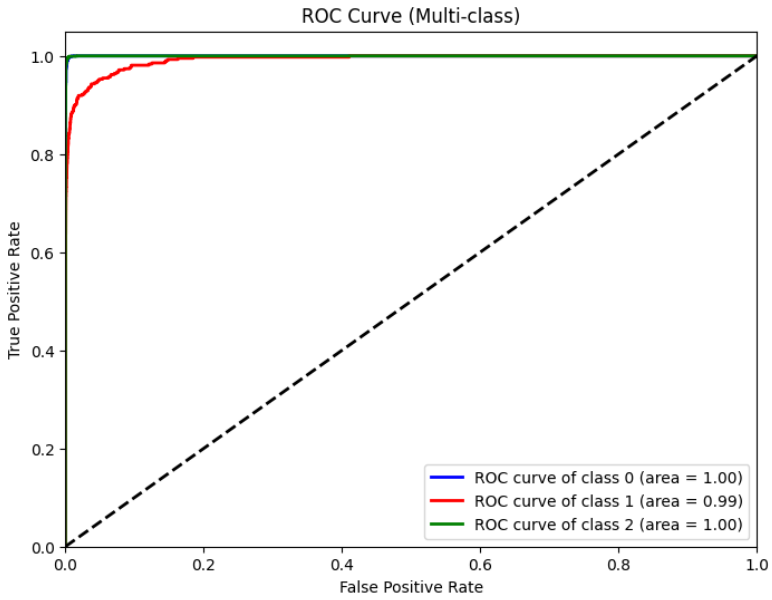


Figure 5.7

5.1.4 Random Forest

The Random Forest classifier has an accuracy of around 84.1 percentand F1 score of 80.5 per cent. Confusion matrix of random forest assesses that the model could correctly predict 41460 instances a class 0, no instances for class 1 and 1234 for class 2. The off-diagonal elements are the incorrect predictions. The ROC-AUC curve shows that the model could predict around 79 percent for class 0, 62 percent for class 1 and 80 percent for class 2. Figure 5.8 and 5.9 shows the confusion matrix and area under the curve for random forest.

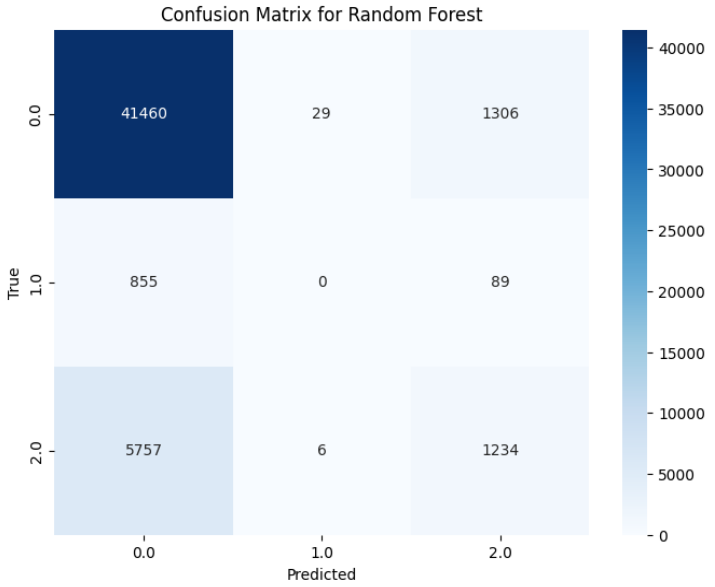


Figure 5.8

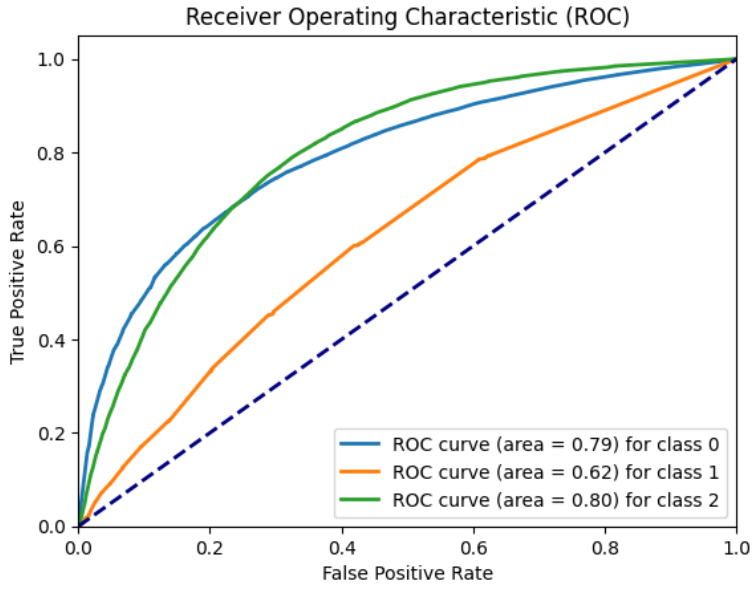


Figure 5.9

5.1.5 Logistic Regression

The logistic regression has an accuracy of around 84.8 percent and F1 score of around 80 percent. The confusion matrix of logistic regressions shows that the model correctly classified

41823 as 0, no instances for 1 and 1211 for class 2. On analysing further around 5788 instances are classified to class 0 when they were class 2 which is a significant amount. And the ROC - AUC curve also indicates that though it can predict class 0and 2 around 81 and 82 percents, this model is not a fit for predicting class 1 since it could only predict 69 percent of class 1. The confusion matrix and roc curve are shown in figures 5.10 and 5.11

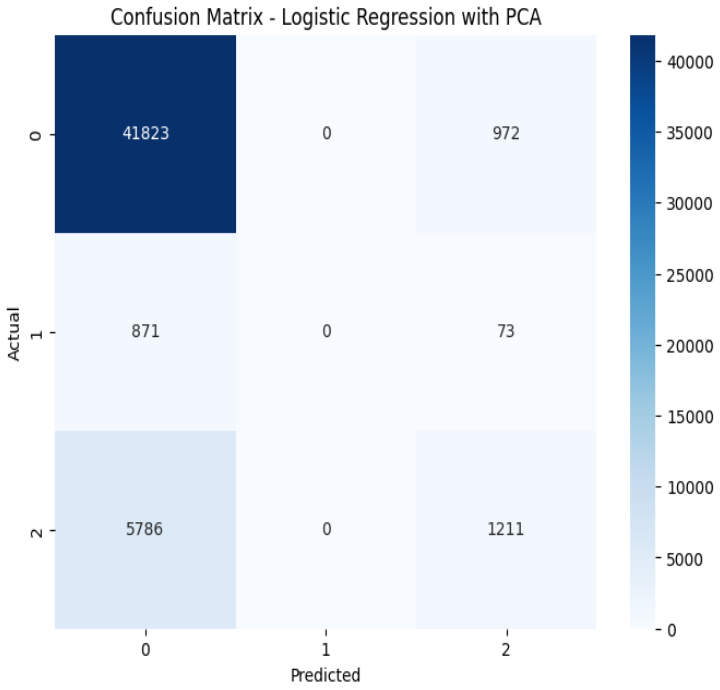


Figure 5.10

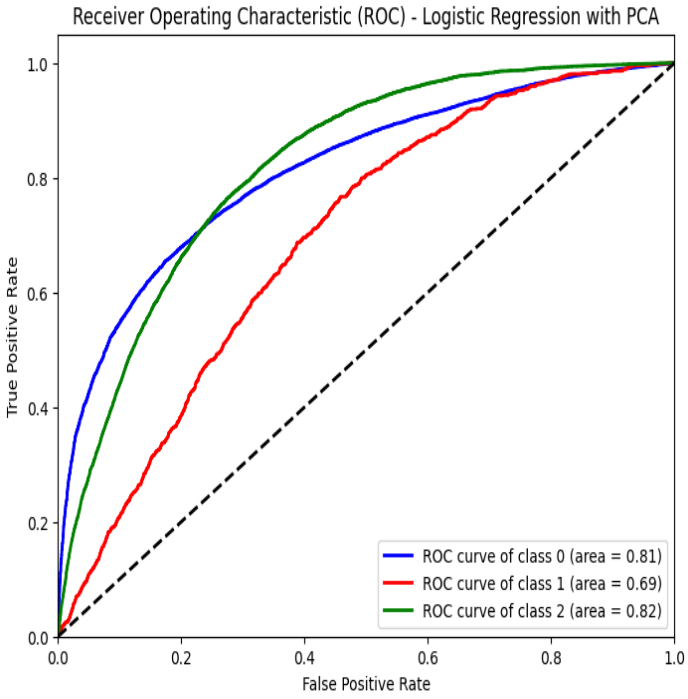


Figure 5.11

5.2 Comparison of Models

For the diagnosis of diabetes ,we build different models and XG Boost was the best model with accuracy score of 0.99 and F1 score of 0.99. Table 5.1 shows the comparison of 5 machine learning models.

Table 5.1

Models	F1 Score	Accuracy Score
Naïve Bayesian Classifier	0.48	0.59
Decision Tree	0.97	0.97
XG Boost	0.99	0.99
Random Forest	0.80	0.84
Logistic Regression	0.80	0.84

5.3 Feature Importance

In the future importance diagram shown in the figure below ,it is clear that Diff Walk has significant amount of impact in the prediction of diagnosis of diabetes using XG boost. Further more stroke ,heavy alcohol consumption takes second and third place. Furthermore, the factors such as mental health,sex, BMI contribute very less to the prediction of diabetes. The figure 5.12 shows the feature important diagram

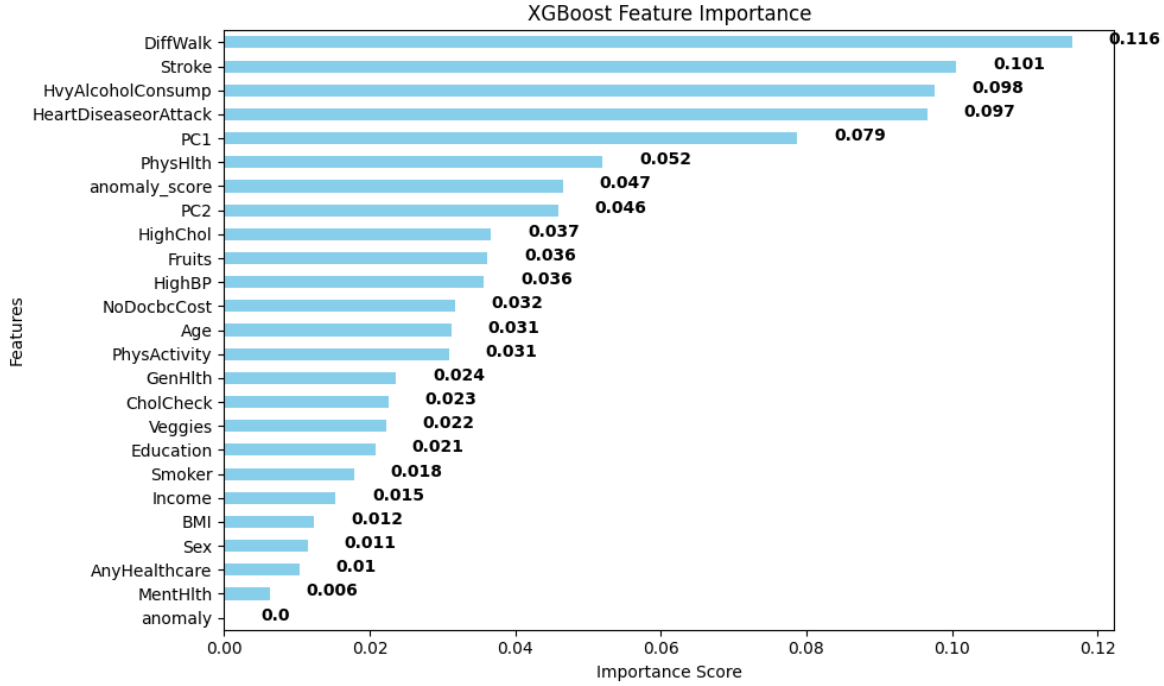


Figure 5.12

Conclusion

As we all diabetes is a lifestyle disease effecting millions of people regardless of the age, importance of disease prediction early on will help in managing the disease. Since it is a lifestyle disease. it is important for the patients to change their lifestyle to reduce the effects of issues caused by diabetes.

The results of this study have shown how critical it is to recognise diabetes early. The BRFSS 2015 data collection was used for the study. Five machine learning models, including the Naive Bayesian Classifier, Random Forest, Decision Tree, XG Boost, and Logistic Regression, were evaluated using this dataset. Anomalies are mined using isolation forest and removed . With accuracy score and F1 score of 0.99, XG Boost was the best model among them. Feature importance graph is plotted and it indicates that maintaining overall health is important to prevent diabetes

REFERENCES

1. Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, *165*, 292-299.
2. Soni, M., & Varma, S. (2020). Diabetes prediction using machine learning techniques. *International Journal of Engineering Research & Technology (IJERT)*, *9*(09), 2278-0181.
3. Rani, K. J. (2020). Diabetes prediction using machine learning. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *6*(4), 294-305.
4. Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, *8*, 76516-76531.
5. Xue, J., Min, F., & 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.
6. Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *Ict Express*, *7*(4), 432-439.
7. Ramesh, J., Aburukba, R., & Sagahyroon, A. (2021). A remote healthcare monitoring framework for diabetes prediction using machine learning. *Healthcare Technology Letters*, *8*(3), 45-57.
8. Ahmed, U., Issa, G. F., Khan, M. A., Aftab, S., Khan, M. F., Said, R. A., ... & Ahmad, M. (2022). Prediction of diabetes empowered with fused machine learning. *IEEE Access*, *10*, 8529-8538.
9. Febrian, M. E., Ferdinan, F. X., Sendani, G. P., Suryanigrum, K. M., &Yunanda, R. (2023). Diabetes prediction using supervised machine learning. *Procedia Computer Science*, *216*, 21-30.
10. Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare technology letters*, *10*(1-2), 1-10.
11. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset
12. https://www.geeksforgeeks.org/