IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Cutting-Edge Machine Learning Methods for Diabetes Forecasting

[1] V. Vishnupriya, [2]E. Pragatheeswari, [3]G. Nisanth, [4] D.Dhanushree, [5]P.Sivakumar

[1] Assistant Professor [2] Assistant Professor, [3]PG-MCA, [4]UG-BSc, [5]PG-MCA
[1] Department of Computer Science and Engineering,
[1] Kongu Engineering College
Perundurai, India.
vishnupriya5665@gmail.com

Abstract: Diabetes is a dangerous medical condition that can cause heart disease, renal problems, visual problems, and other complications. For these problems to be effectively managed and prevented, early diabetes prediction is essential. Using patient data, machine learning algorithms present a viable method for diabetes prediction. Support Vector Machine and K-Nearest Neighbor with Grid Search Optimization are two particularly effective methods that routinely yield very accurate predictions, properly detecting diabetes in about 99 out of 100 cases. Further useful techniques for diabetes prediction include Gradient Boosting Machines, Neural Networks, Principal Component Analysis, Logistic Regression, Singular Value Decomposition, K-Means Clustering, Hierarchical Clustering, and Gaussian Mixture Models. The Random Forest technique is another efficient approach that achieves comparable high accuracy rates of approximately 99% by combining numerous decision trees. By combining these algorithms with Random Forest, we can detect and treat diabetes early on, improving patient outcomes and prediction accuracy.

Keywords - Diabetes- KNN (K-Nearest Neighbor)- SVM (Support Vector Machine)- Hyperparameter- Grid Search- Machine Learning- Naive Bayes- Decision Tree- Random Forest- Gradient Boosting Machines (GBM)- Neural Networks- K-Means Clustering- Hierarchical Clustering- Principal Component Analysis (PCA)- Singular Value Decomposition (SVD)- Logistic Regression.

INTRODUCTION

Diabetes disease is notably the worst and the deadliest disease in the universe and has been studied in many papers they have discovered factors like obesity, inborn, age, and diet. Diabetes disease is caused by the high blood glucose level [14][17]. In the human body in the event that the pancreas doesn't create enough insulin and it increases in the blood glucose level so it causes diabetes and it develops step by step [13]. If we untreated diabetes it causes severe damage in human health conditions like eye damage, heart problems, kidney problems, urination problems, and blood pressure [3][5][16]. Currently, patients must visit a diagnostic center to receive their reports, which requires a significant investment of time and money. However, with the development of Machine Learning methods, advanced information processing systems are available that can be used to predict the risk of diabetes in a patient [10]. This analysis seeks to fabricate a system that can faultlessly predict the inception of a patient's disease. Additionally, information withdrawal has the potential to remove hidden data from a large amount of diabetes-related information [8]. Early diagnosis of disease is essential for the enhancement of health care services, and this will help individuals avoid hazardous health conditions before they become more complex. Diabetes Mellitus is a potentially lethal illness characterized by increased glucose levels due to insulin secretion defects. As a result, early diagnosis of the disease has become a major focus of recent research. Currently, there is a considerable amount of research being conducted on Machine Learning with a particular emphasis on medical applications [21].

According to World Health Organization (WHO) In 2012, diabetes was the leading cause of death worldwide, overall, 1.5 billion deaths. Most diabetes and complications can be prevented by eating healthy food, exercising regularly, and maintaining body weight. The prevalence of diabetes has increased significantly since 1980, rising from 108 million to 422 million individuals in 2014. This increase has been pronounced in developed and developing-income countries, as opposed to high-income nations. Furthermore, the rate of diabetes mortality by age increased by 3% between 2000 and 2019. In 2019, diabetes-related kidney disease was estimated to be responsible for 2 million deaths. Diabetes is found among various Countries like Egypt, China, the United States, and India, etc. Early detection of a diabetes disease can help to control and save a person's life [12]. Many cases of diabetes are observed on a daily basis. This chronic condition can take a lifetime to recover from, and the doctor can only save a patient in the earliest stages of the disease. If the disease is in the final stage, it is unlikely that the patient will be able to recover. This system is designed to provide the most effective solution for the early prediction of diabetes so that the doctor can detect the disease in an earlier stage. The primary objective of research in early diabetes is to further refine the detection system for diabetes in the model so that doctors can accurately predict the patient's condition. There are many research papers published in the public domain about

different types of diabetes to predict in various stages. The goal of this research is to figure out what the symptoms of the disease will be in the early stages so that it can be avoided in the future and the patient won't have to pay a lot of money in the future [11].

To achieve this paper examines the prediction of diabetes using various attributes associated with diabetes disease. To predict diabetes, we use the diabetes Datasets and various Machine Learning Algorithms. The machine learning method is to train the machines and computers. It is used to predict the best technique for earlier prediction [24].

LITERATURE REVIEW

Minakshi R. Rajput et.al., The main goal is to predict diabetes by assessing various human bodily parameters with five distinct machine learning methods. It calculates the link between the significant characteristics that cause diabetes and identifies them. The relative risk values of prediabetes and normal individuals for getting diabetes are also compared in this article.[1]

Minyechil Alehegn et.al., To protect human life, Data Mining Techniques (DMT) are very good at predicting medical datasets at a very early stage. Over time, several academics have employed various techniques to classify and forecast symptoms in medical data.[23] The PIDD (PIMA Indian Diabetes Datasets) data set has 768 records and can be accessed online. In this study, the majority of known prediction algorithms are employed. Various approaches are combined in Naïve Net, Proposed Ensemble Method (PEM), Support Vector Machine, and Decision Stump [2].

Sethupathi M et.al Diabetes prevalence is expected to double in the next decade, reaching 382,000 individuals per year. Several health problems, including myopia, renal impairment, amputation, and stroke, can be brought on by high blood sugar levels. A predictive model for diabetes is being created by machine learning and a variety of algorithms, such as K-means, Random, Logistic, Support Vector Machine, and Decision Tree [4].

Tejas N et.al., Diabetes is a chronic condition causing elevated blood glucose levels, affecting 382 million people globally by 2035. Early diagnosis is difficult due to complex interdependencies on various factors. A project aims to forecast diabetes using supervised machine learning techniques to propose an efficient technique for early detection [6].

Muhammad Azeem Sarwar et.al., This paper examines how big data in healthcare is habituated to predict diabetes by exploiting machine learning tactics. It examines six distinct algorithms and contrasts their accuracy and performance to determine which is most appropriate for diabetes prediction and how this can assist medical professionals in making prompt decisions regarding the health and course of treatment for their patients [7].

Muhammaad Exell Febrian et.al., Diabetes is a plague that can cause death, heart attacks, blindness, and damage to the kidneys. It is projected that by 2030 there will be 578 million diabetics worldwide, and by 2045 there will be 700 million. Technology for diabetes detection is needed to prevent serious issues. Diseases can be accurately and quickly diagnosed with machine learning [9].

Juncheng Ma The contemplation used six machine learning models to predict diabetes diagnosis including Support Vector Machine, Logistic regression, Boosting, Neural network, Decision tree, and Random forest. Results showed random forests, Boosting, and neural networks performed better. The neural network has the highest accuracy at 96% [15].

Aishwarya Mujumdar et.al This commentary proposes a Diabetes Prediction Model to improve diabetes classification in hospitals. It includes external factors like glucose, BMI, age, and insulin. The model improves accuracy by adding a new dataset and a Pipeline Model for Diabetes. This approach enhances the classification accuracy by incorporating new data and addressing the existing method's inaccuracies [18].

Mohammad Abu Tareq Rony et.al., The researchers used a variety of machine-learning approaches to develop a statistical sample for diabetes patients. Random Forest was chosen to be considerably factual (97.5%) for F1-measure and area under receiver employing distinct curve (99.80%) using data from two hospitals. Logistic Regression was the most accurate (77.7%) for evaluating information gain and correlation attributes [20].

Bello A. Bodinga, Examines the relationship between Diabetes and Machine Learning, with the primary emphasis on the early identification and management of the condition. Comparing the performance of RF, LR, and Random Forest algorithms using performance metrics Logistic regression achieved a 77% accuracy rate, Decision Tree a 0.77% accuracy rate, Random Forest a 0.64 accuracy rate, and Random Forest a recall rate of 0.58 [22].

PROPOSED WORK

The proposed approach is to employ a variety of widely-used Machine Learning strategies and estimate their performance on the Diabetic Data set to determine the most suitable machine learning algorithm to glimpse diabetes in individuals. The subsequent machine-learning algorithms are employed in this contemplation Support Vector Machine, K-Nearest Neighbor (Grid Search), Decision tree, Logistic Regression, Random Forest, Gradient-Boosting Machines (GBM), and Neural Networks and Bernoulli Naive Bayes. Figure 1 shows the overall evaluation process to forecast diabetes.

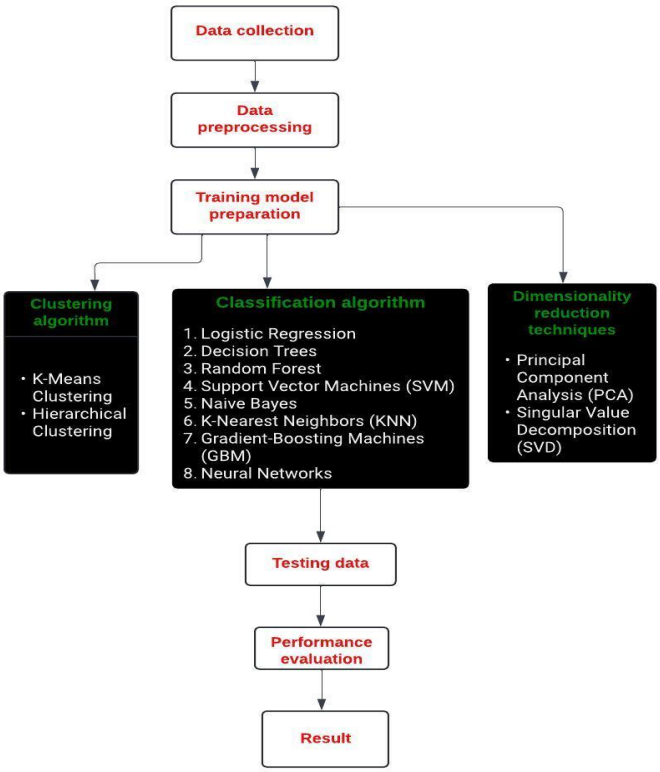

Figure 1. Flow diagram of Proposed System

A.Dataset Collection

The diabetes UCI dataset which were taken from the Kaggle. The datasets consist of 520 records of data which are divided into two class labels positive and negative classes and 17 attributes.

B.Data Preprocessing

The first and most vital phase is data preprocessing. Data preparation is the process of cleansing data so that it can be used by machine learning algorithms. A lot of medical-related data has noise and missing values. It may also be in formats that can't be processed directly by machine learning models. Cleaning the data also improves the model's accuracy and performance. First, we needed to find the missing value in the datasets available or not. In diabetes datasets, most of the attributes are categorical, so the categorical values have to be converted into numerical values. This allows us to predict good results using various machine-learning techniques.

C.Training Data and Test Data

The datasets are then broken down into training and test data. 80% operated for training and the testing data is the other 20%. Each model has the training data added to it, and then the predicted results are compared with the testing data to see how accurate they are. We also look at the confusion matrices and compare performance measures like precision, recall, accuracy, and F1-Score measures.

FEATURES	DESCRIPTION	VALUES	REPRESENTATION
AGE	THE AGES OF 20 TO 75	INTEGER	DAYS
GENDER	IT CAN BE MALE/FEMALE	CATEGORICAL VALUES	1-MALE 0-FEMALE
POLYURIA	WHETHER PATIENT URINATED OVERLY	CATEGORICAL VALUES	0-NO 1-YES
POLYDIPSIA	WHETHER THE PATIENT HAD A SEVERE THIRST	CATEGORICAL VALUES	0-NO 1-YES
ITCHING	THE PATIENT HAD ITCHINAESS.	CATEGORICAL VALUES	0-NO 1-YES
IRRITABILITY	THE PATIENT ENCOUNTERED A FIT OF IRRITATION.	CATEGORICAL VALUES	0-NO 1-YES
DELAYED HEALING	WHEN WOUNDED, AND THE PATIENT EXPERIENCED A SLOWED CONVALESCENCE PROCESS.	CATEGORICAL VALUES	0-NO 1-YES
PARTIAL PARESIS	THE PATIENT UNDERWENT MUSCLE FRAILTY.	CATEGORICAL VALUES	0-NO 1-YES
MUSCLE STIFFNESS	THE PATIENT UNDERWENT AN INSTANCE OF RIGIDITY IN THE MUSCLES.	CATEGORICAL VALUES	0-NO 1-YES
ALOPECIA	THE PATIENT HAD HAIR LOSS	CATEGORICAL VALUES	0-NO 1-YES
OBESITY	USING HIS BODY MASS INDEX, THE PATIENT CAN BE CLASSIFIED AS OBESE	CATEGORICAL VALUES	0-NO 1-YES
SUDDEN WEIGHT LOSS	THE PATIENT UNDERWENT A PERIOD OF RAPID WEIGHT LOSS	CATEGORICAL VALUES	0-NO 1-YES
WEAKNESS	THE PATIENT UNDERWENT A PERIOD OF WEAKNESS.	CATEGORICAL VALUES	0-NO 1-YES
POLYPHAGIA	THE PATIENT UNDERWENT A PERIOD OF FIERCE STARVATION.	CATEGORICAL VALUES	0-NO 1-YES
GENITAL THRUSH	THE PATIENT HAD A YEAST INFECTION OR NOT	CATEGORICAL VALUES	0-NO 1-YES
VISUAL BLURRING	THE PATIENT UNDERWENT A MOMENT OF EYESIGHT BLUR.	CATEGORICAL VALUES	0-NO 1-YES
CLASS	EXISTENCE OF DIABETES	CATEGORICAL VALUES	1-POSITIVE 0-NEGATIVE

TABLE 1 - DIABETES UCI DATASET

D. Methodology

By utilizing a sort of ML strategies and evaluating them against diabetic datasets, this proposed methodology makes it possible to determine which machine learning algorithm is most suited for detecting diabetes in an individual. The following algorithms are used in this methodology Decision Tree, Support Vector Machine,KNN with hyper-parameter gird search,Bernoulli Naive Bayes,Logistic Regression, Decision Trees,Random Forest,Gradient-Boosting Machines (GBM),Neural Networks,

CLASSIFICATION ALGORITHMS

1. Support Vector Machine

SVM is classified as supervised learning strategy. It can be further classified as a classification algorithm and a regression algorithm, however, SVM is primarily employed in classification algorithms.

SVM divides various target classes into hyperplanes in n-dimensional space or multidimensional space. The primary pursuit of the SVM strategies is to specify a perfect decision boundary separating two or more classes by a maximum margin. This allows the placement of new data points in the relevant class.

The data points in the hyperplane are called vectors. The kernel makes it super simple to work with large-scale data. There are excess SVM kernel functions that can utilized to turn non-linear data into linear data. There are different types of kernel functions are available

- Radial Basis Function (RBF)
- Sigmoid
- Polynomial
- Linear

Radial Basis Function Kernel (RBF)

It's one of the most well-known and used kernel functions in SVM. It's often used with non-linear data and helps with proper separation when you don't have any prior knowledge of the data.

Linear Kernel

Linear kernel functions, which are typically one-dimensional in nature and the most fundamental type of kernel. When the number of features is large, linear kernel functions tend to be faster than other functions.

Polynomial Kernel

Polynomial kernels, which are non-linear kernels employed in machine learning, essentially transform the data from one dimension to another employing a polynomial function.

Sigmoid Kernel

The kernel function is the most widely used type of neural network. It looks similar to a two-layer perceptron model inside a neural network and acts as the neuron activation function.

Algorithm For SVM

1) Plotting data points based on class labels.
2) Draw the N digit of possible hyperplanes.
3) choose the most satisfactory hyperplane that divides the class as maximum margin.
4) Figuring out how far away two planes are from each other is called the maximum margin or decision boundary.

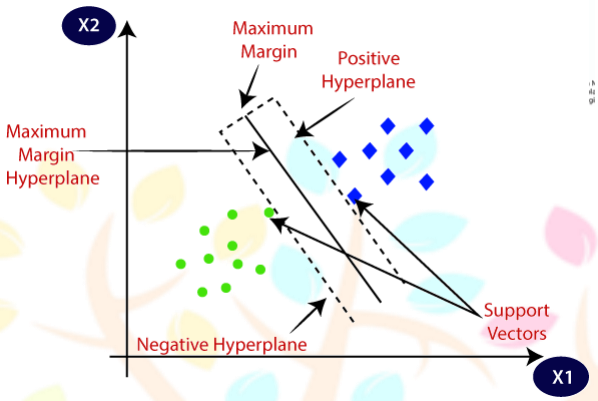

Figure.2 - Support Vector Machine

2. K-Nearest Neighbor

K-NN is a most popularized and fundamental ML strategy that uses supervised learning techniques. It detects similarities between freshly received instances/data and existing cases and classifies them accordingly. KNN saves all datasets and sorts new test values deployed to their closeness, allowing new data to be effortlessly sorted toward a well-fitted class. K-NN is a non-parametric strategy, which symbolizes that it creates no speculation regarding the data it is working with. It is also comprehended as a "lazy learner" since it does not immediately practice from a learning sample, instead holds data samples and enacts actions on them during the learning phase.

Algorithm For KNN

STEP 1 - Pick the k digit of nearest neighbor.

STEP 2- Computing the k-nearest neighbor by using Euclidean distance.

STEP 3- The K-nearest neighbors are stated by the Euclidean distance computed.

STEP 4- Estimate the amount of data samples within each class in the k neighbors.

STEP 5- Test provided data with training data and determine the majority of class labels.

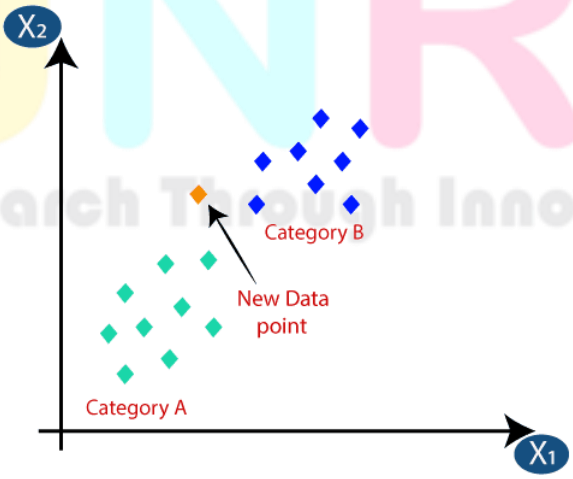

Figure.3 - K- Nearest Neighbor

K-NN (With Grid Search Optimization Techniques)

Grid search is one of the simplest algorithms you can use to get hyperparameters fine-tuned. We divide up the hyperparameter domain into different grids. Then, we try out all the possible combinations of parameters from each grid. Finally,

we cross-test some performance metrics against each grid to see which one gets the most from the average of the cross-validated parameters.

Grid search is an algorithm that goes through all the combinations and finds the best one in the domain, but the downside is that it's really slow. It takes a lot of time to check every combination in the space, which isn't always available and each point in the grid requires k-fold cross-validation, which means you need to train your model for k times. So, tuning your model's hyperparameters can be complicated and expensive. But if you want to find the best combination of hyper-parameter values, grid search is a good option.

3. Decision Tree

The Decision Tree is a widely used supervised learning strategy that is easy to implement and interpret. It tree format that consists of leaf nodes, branches, and internal nodes. In a classification or else the regression exceptions, leaf nodes represent the result or constrained variable, while inner nodes represent the unconstrained variables. The branches indicate the decision rules that will be followed as decision-making progresses through the tree. Decision-making initiates at the core of the tree and progresses through the tree before a decision being reached. Each non-coastal node comprises a selection procedure that is employed when correlating a feature from data samples with a tree. The nodes are assigned attributes Gini, information gain, and entropy.

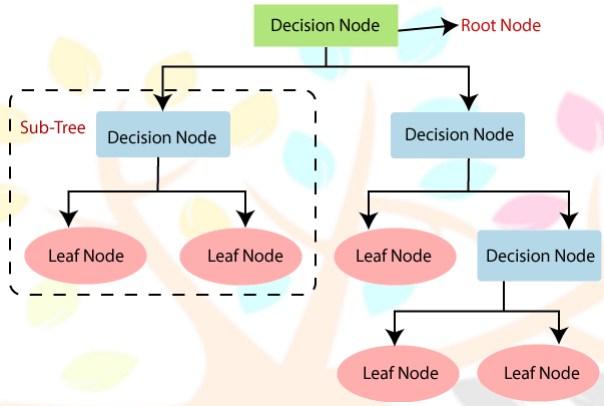

Figure.4 - Decision Tree

4. Bernoulli Naive Bayes

The Naive Bayesian Classification technique is a supervised learning strategy that originated from the Bayesian distribution theorem. It is designed to solve classification problems and is considered to be a precise and efficient Classification algorithm for machine-learning standards. It is competent in producing rapid projections and provides better results when datasets are of high dimension.

The Naive Bayesian variant of the Naive Bayesian theorem, developed by Bernoulli, is employed in machine learning. This variant is particularly useful when the dataset is a binary distribution in which the output label is either this or that. This algorithm has the advantage of only accepting binary features, such as

- 0/1
- YES/NO
- TRUE/FALSE
- POSITIVE/NEGATIVE

Bernoulli Naive Bayes Formula

$P(Ai / B) = P(i / B)\ Ai + (1 – P(i/ B))\ (1 - Ai)$

P (A | B) is the probability that Ai is going to happen if B has already happened. i is the event that x holds either 0 or 1.

E.Build Model for Proposed Methodology

STEP 1: Importing the libraries and datasets.

STEP 2: Data preprocessing is needed to remove the missing data.

STEP 3: The data needs to be divided into testing data and training data.

STEP 4: Training data incorporates with 80% of the total data and the testing data incorporates 20%.

STEP 5: Choose the algorithm that ought to be performed.

STEP 6: Utilize the test set to assess the Classifier model for the machine learning algorithm.

STEP 7: Comparison of the performance evaluation.

STEP 8: After analysis, select the algorithm that performs the best based on various parameters.

F.Performance Evaluation

There are a variety of machine learning models available that can be utilized to assess a model's performance. There is a range of evaluation metrics that can be applied to the model, including accuracy, recall, F1-score, and precision were calculated employing confusion metrics.

Accuracy Score

Accuracy is one of the simplest evaluation metrics to evaluate different types of machine learning algorithms.

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

Precision

This shows the rate of the total anticipated samples that were correctly predicted to be positive. It can be expressed as follows:

Precision = TP/(FP+TP)

Recall

Recall shows the rate of total diabetic samples that stood correctly classified as diabetic.

Recall = TP/(FN+TP)

F1-Score

F1-Measure is the total number of times the model performed correctly on the test data set according to the confusion matrix.

F1-Score = 2TP/(FN+FP+2TP)

Where,

TN =>True Negative
TP =>True Positive
FN =>False Negative
FP =>False Positive.

5. Logistic Regression

1.For binary classification tasks, such as those with a category target variable and two alternative outcomes (e.g., 0 or 1, yes or no), statistical regression is employed.

2.The logistic function, which converts any input to a number between 0 and 1, is used to model the likelihood that a given input belongs to a specific class.

3.The premise of logistic regression is that there is a linear relationship between the input variables and the outcome's log-odds.

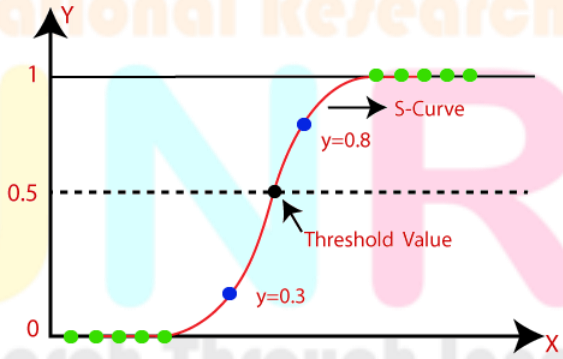

Figure.5 – Logistic Regression

6. Random Forest

• This ensemble learning technique is applied to both regression and classification problems.

• During training, it builds several decision trees, from which it produces the mode (classification) or average prediction (regression) of each tree.

• By training each tree using a bootstrap sample of the training data and taking into account a random subset of features at each split, Random Forest adds randomization to the system.

• When compared to individual decision trees, it reduces the likelihood of overfitting and increases prediction accuracy and generalization.

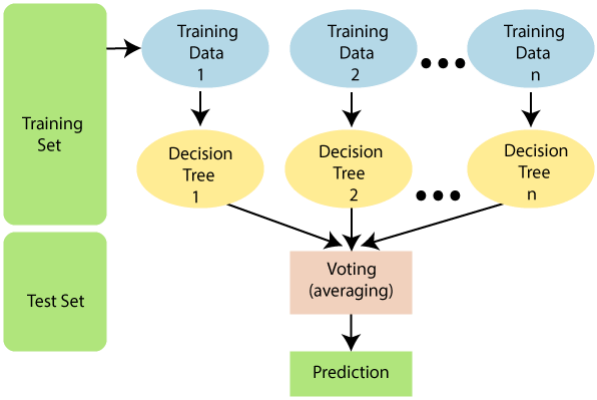

Figure.6 – Random Forest

7. Neural Networks

Artificial Neural Networks (ANNs), often known as neural networks, are a class of models that draw inspiration from the biological neural networks seen in the human brain.

• They are made up of interconnected layers of nodes, or neurons, each of which carries out a straightforward computation and then transfers its result to the one above it.

• Through the processes of forward propagation, which involves sending input data through the network to generate predictions, and backward propagation, which involves modifying the network's parameters in response to prediction failures, neural networks can discover intricate patterns in data.

•Their versatility allows them to be used for a wide range of tasks, including as image and speech recognition, regression, and classification.

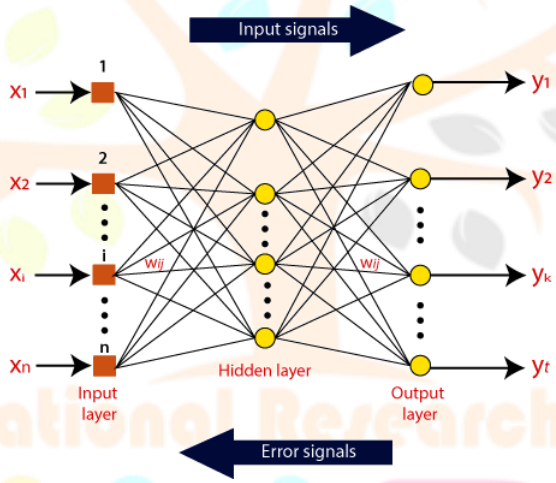

Figure.7 – Neural Networks

8. Gradient-Boosting Machines (GBM)

• This additional ensemble learning method is mainly applied to tasks related to classification and regression.

• GBM successively constructs an ensemble of weak learners, usually decision trees, whereby each new tree fixes the mistakes produced by its predecessor.

• It works by iteratively fitting new models to the residual errors of the prior models to minimize a loss function (such as the mean squared error for regression or the deviation for classification).

• GBM is a well-liked option in machine learning contests and practical applications because of its strong prediction accuracy and resilience to overfitting.

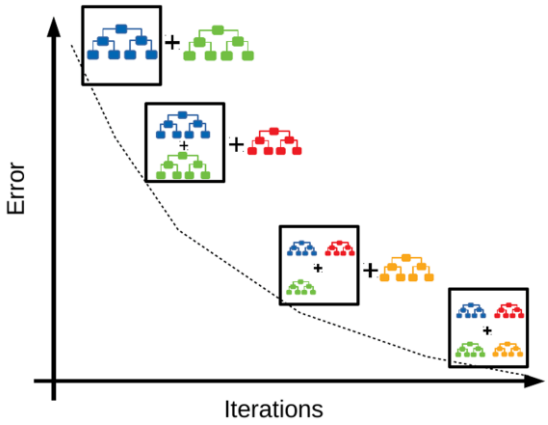

Figure.8 – Gradient Boosting Machines

CLUSTERING ALGORITHMS

K-Means Clustering

For applications involving grouping, one common unsupervised machine learning technique is clustering.

A dataset is divided into 'k' unique, non-overlapping clusters, where 'k' is an integer that the user specifies

.Using an iterative process, the algorithm places each data point in one of the 'k' clusters according to how similar their features are.

The goal is to reduce the within-cluster variance, which is typically calculated by adding the squared distances between each data point and the cluster centroid.

K-Means is widely utilized in many different applications, including picture compression, anomaly detection, and customer segmentation. It is also comparatively quick and scalable.

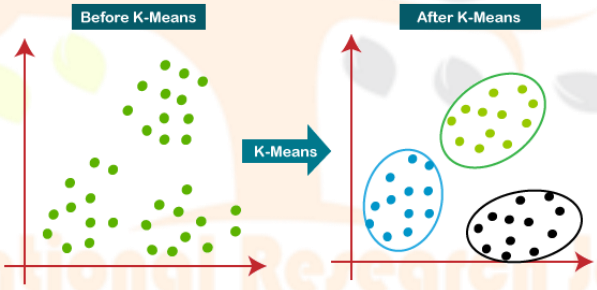

Figure.9 – K-Means Clustering

Hierarchical Clustering

Another unsupervised learning technique for assembling comparable data points into groups is hierarchical clustering.

Hierarchical Clustering does not require the user to predetermine the number of clusters, in contrast to K-Means.

Iteratively joining or dividing clusters according to the proximity of data points results in a hierarchical tree of clusters, or dendrogram.

Agglomerative (bottom-up) and divisive (top-down) hierarchical clustering are the two primary forms. Whereas divisive clustering starts with all the data points in a single cluster and splits them recursively, agglomerative clustering starts with each data point as a separate cluster and merges them iteratively.

Because of its adaptability and ability to be shown as a dendrogram, hierarchical clustering is a valuable tool for exploratory data analysis and comprehending the connections between different data points. Still, For big datasets, it could be computationally costly though.

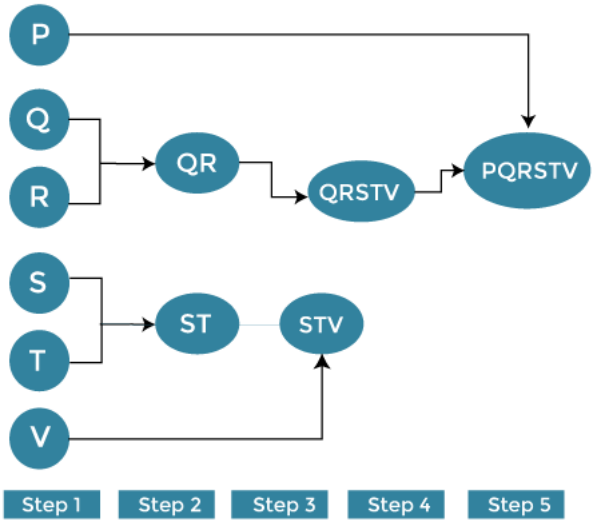

Figure.10 – Hierarchical Clustering

DIMENSIONALITY REDUCTION TECHNIQUES

Principal Component Analysis (PCA)

A dimensionality reduction method called principal component analysis is used to move high-dimensional data into a lower-dimensional space while keeping the most crucial information.

The original data is projected onto the directions (principal components) that show the greatest variation in the data, as determined using PCA.

The principle components rank according to how much of the variance in the data they can explain, and they are orthogonal to one another.

PCA lowers the dimensionality of the dataset by choosing a subset of the principle components that account for the majority of the variance in the data.

PCA is frequently used as a preprocessing step for several machine learning methods, as well as for exploratory data analysis and visualization.

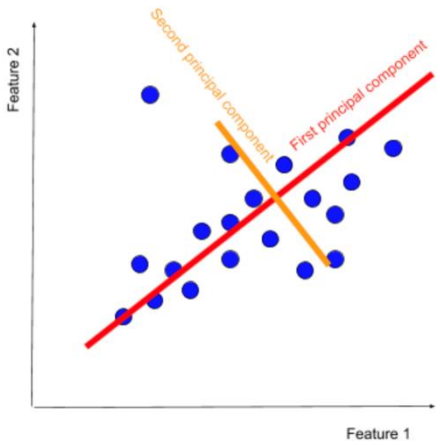

Figure.11 – Principal Component Analysis (PCA)

Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) is a matrix factorization technique that breaks down a matrix into three components:U, Σ,and $V^{T.}$

These components represent orthogonal matrices and a diagonal matrix of singular values, respectively. Widely used in various applications such as dimensionality reduction and image compression, SVD is akin to PCA but offers numerical stability and versatility, accommodating matrices of any size**.**

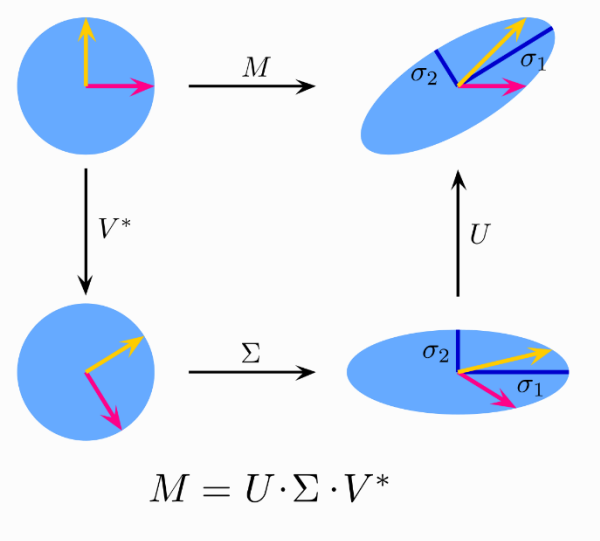

Figure.12 – Singular Value Decomposition (SVD)

RESULTS

In this Research work, the diabetes disease projection utilizes machine-learning algorithms to be implemented. In this work, we use Google Colab and Kaggle datasets to predict a patient has diabetes or not.

After performing statistical calculations, the program will gain a deeper understanding of the data. The data is split into an 80:20 portion, 80% of data is employed for training data, and 20 % of data is employed for testing data.

Eight Algorithms are used Decision Tree, Bernoulli Naive Bayes, Support Vector Machine,KNN (with Grid Search Optimization techniques),Logistic Regression, Random Forest, Neural Network and Gradient Boosting Machines are compared with accuracy, recall, precision, and f1-score which are calculated and compared.

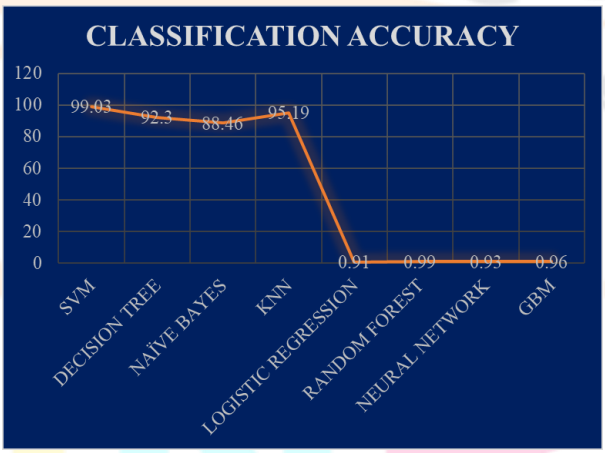

Figure.13 – Classification Accuracy

There are a bunch of different kernel tricks that can be used with SVM like linear, sigmoid, polynomial, and RBF. Using machine-learning algorithms, polynomial kernels are more accurate than other kernels, with 99.03% accuracy, linear kernels at 92%, sigmoid at 86%, and RBF at 98%.

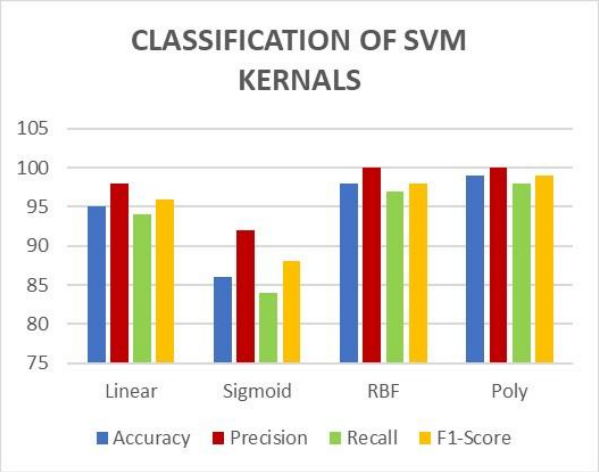

Figure.14 – Classification of SVM Kernal's

Algorithms	Accuracy	Precision	Recall	F1-Score
SVM	99.03	1.00	0.98	0.99
DECISION TREE	92.30	0.91	0.97	0.94
NAIVE BAYES	88.46	0.88	0.93	0.90
KNN	95.19	0.98	0.94	0.96
KNN (WITH GRID SEARCH)	99.03	1.00	0.99	0.99
LOGISTIC REGRESSION	0.91	0.95	0.91	0.93
RANDOM FOREST	0.99	1.00	0.98	0.99
NEURAL NETWORK	0.93	0.95	0.94	0.95
GRADIENT BOOSTING MACHINES	0.96	0.98	0.95	0.97

TABLE 2 – Evaluation Metrics Values

There are different evaluation metrics available like recall, precision, f1-score, and accuracy. Comparison of the metrics that are shown below Figure.11 - Evaluation Metrics

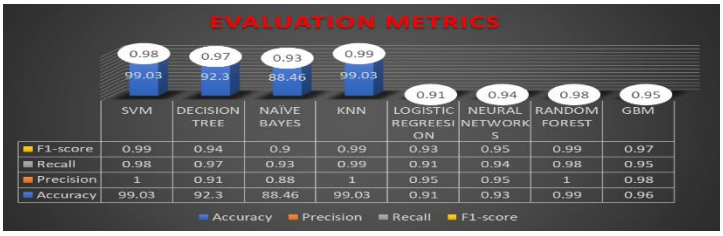

Figure.15- Evaluation Metrics

CONFUSION MATRIX FOR CLASSIFICATION ALGORITHMS

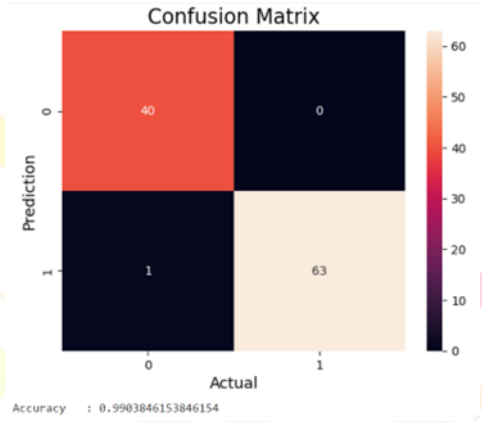

Figure.16 - Matrix for Support Vector Machine

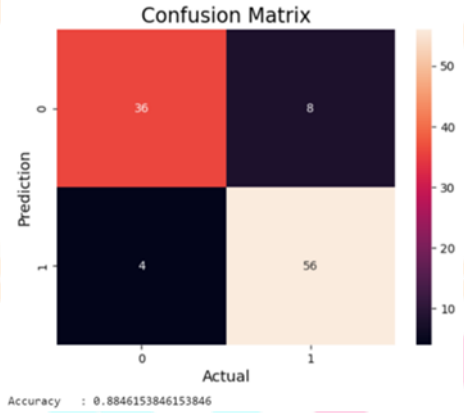

Figure.17 - Matrix for Naive Bayes

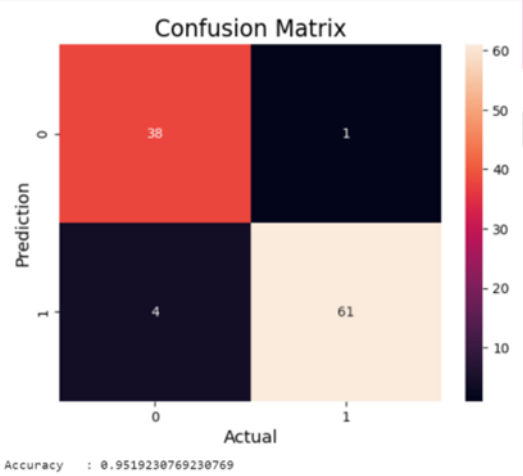

Figure.18 - Matrix for K-Nearest Neighbor

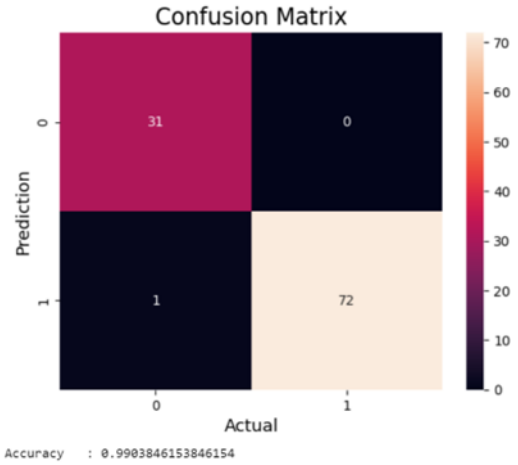

Figure.19 - Matrix for KNN (GRID SEARCH)

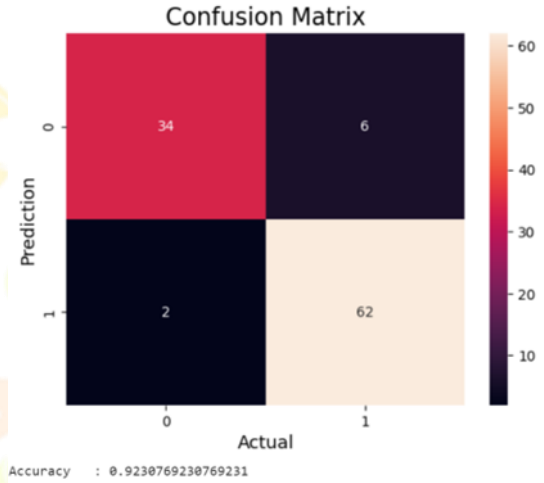

Figure.20 - Matrix for Decision tree

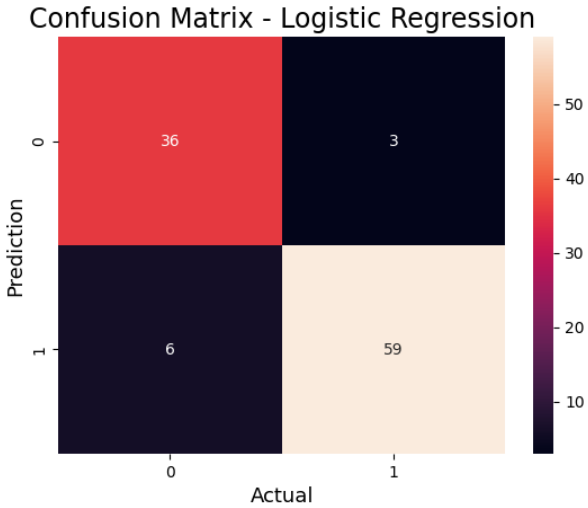

Figure.21 - Matrix for Logistic Regression

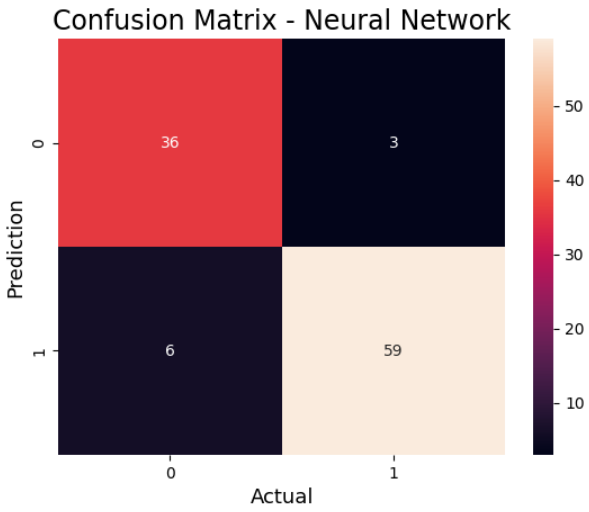

Figure.22 - Matrix for Neural Network

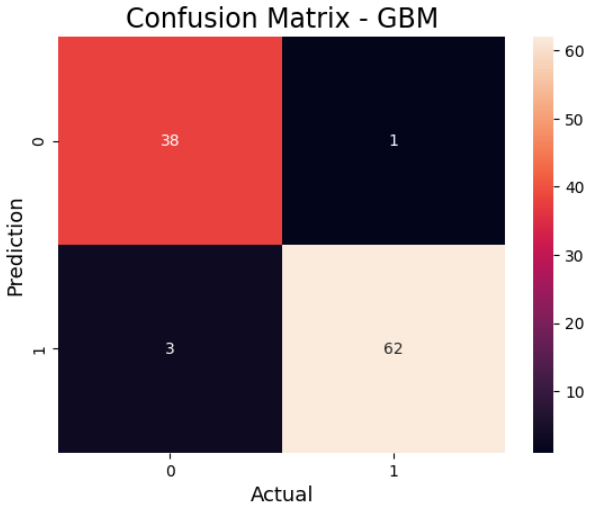

Figure.23 - Matrix for GBM

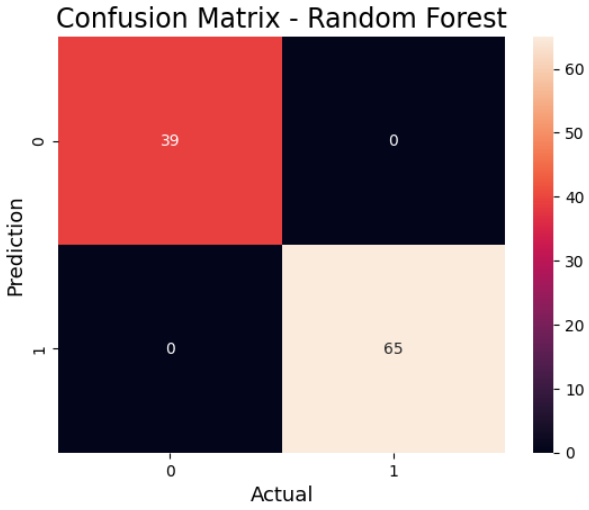

Figure.24 - Matrix for Random Forest

CLUSTERING ALGORITHMS OUTPUT

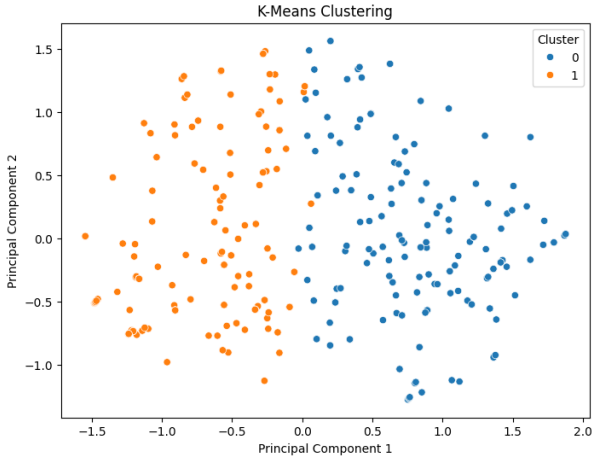

Figure.25 – K-Means Clustering

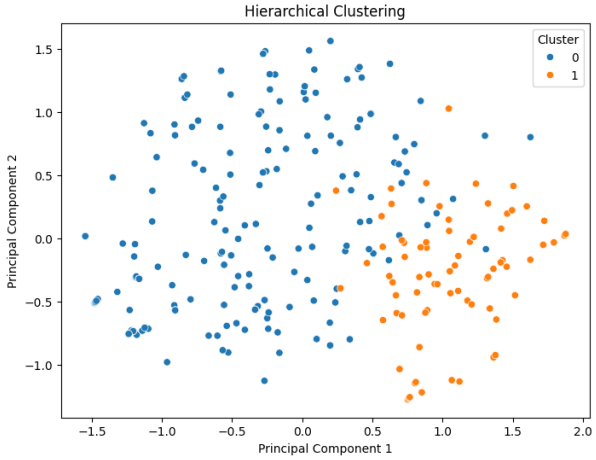

Figure.26 – Hierarchical Clustering

.

DIMENSIONALITY REDUCTION TECHNIQUES OUTPUT

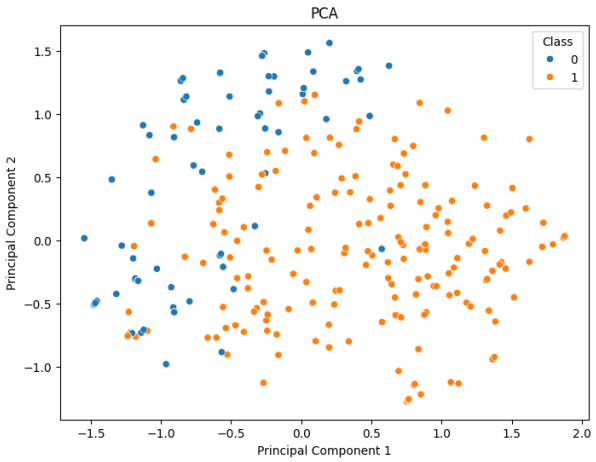

Figure.27 – Principal Component Analysis (PCA)

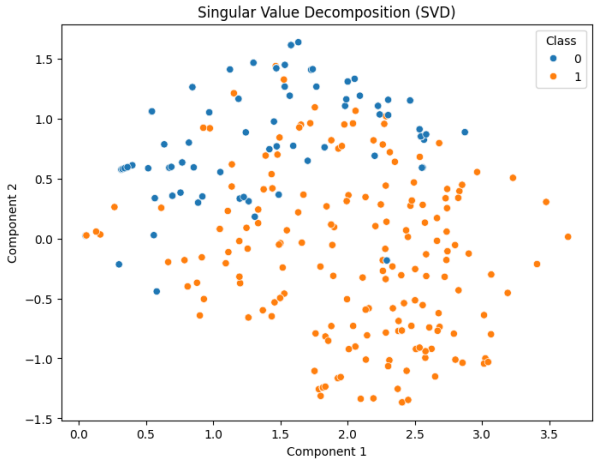

Figure.28 – Singular Value Decomposition (SVD)

CONCLUSION AND DISCUSSION

The classifiers with the highest accuracy, SVM and KNN, are 99.03%.The accuracy of Random Forest is likewise quite excellent, at 99.00%.The next three models have accuracies of 92.30%, 88.46%, and 96.00%, respectively: Decision Tree, Naïve Bayes, and GBM.The accuracy of neural networks and logistic regression is 91.00% and 93.00%, respectively.Other metrics such as precision, recall, and F1-score become important in comparing algorithms when their accuracy is equal or higher.While SVM and KNN both attain the maximum accuracy in this instance, KNN performs better than SVM in terms of precision, recall, and F1-score. Hence, KNN seems to be a preferable option for predicting diabetics because it offers excellent accuracy and balanced performance in terms of precision, recall, and F1-score measurements. Overall, Random Forest is a formidable competitor for diabetic prediction due to its somewhat lower accuracy combined with outstanding precision, recall, and F1-score.

REFERENCES

[1] Minakshi R. Rajput, Sushant S. Khedgikar, 2022, 07-January. “Diabetes Prediction and analysis using medical attributes”. Journal of Xi’an University of Architecture & Technology. ISSN.

[2] Minyechil Alehegn, Rahul Joshi and Preeti Mulay,2018. “Analysis and Prediction of Diabetes Mellitus using Machine Learning Algorithm”. International Journal of Pune and Applied Mathematics.

[3] Mitushi Soni and Sunita Varma ,2020, September. “Diabetes Prediction using Machine learning Techniques”. International Journal of Engineering Research & Technology (IJERT).

[4] Sethupathi M and Privietha P, 2023, April-June. “Diabetes Prediction Using Machine Learning”. ISSN.

[5] Kumar, A. Hemantha, and R. Swetha,2022. "Diabetes prediction using machine learning techniques." Journal of Engineering Sciences

[6] Tejas N. Joshi and Pramila M. Chawan ,2018, January. “Diabetes Prediction Using Machine Learning Techniques”. International Journal of Engineering Research and Application. I

[7] Muhammad Azeem Sarwar et.al.,2018,6-7 September. “Prediction of Diabetes Using Machine Learning Algorithms in Healthcare”. International Conference on Automation & Computing.

[8] Priyanka Sonar and k. JayaMalini, 2019. “DIABETES PREDICTION USING DIFFERENT MACHINE LEARNING APPROCHES” 3rd International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2019.

[9] Febrian, Muhammad Exell, et al.,2022. “Diabetes prediction using supervised machine learning”. International Conference on Computer Science and Computational Intelligence.

[10] Aaditi Ranganath Satam et.al., 2023, April. “Diabetes Prediction using Machine Learning”. International Journal of Modern Developments in Engineering and Science. ISSN.

[11] Ali Nawaz, et.al.,2022, November. “An Applied Artificial Intelligence Techniques For Early Prediction of Diabetes Disease.” IEEE.

[12] Minhaz Uddin Emon. et.al.,2021,13 April. “Primary Stage of Diabetes Prediction using Machine Learning Approaches.”IEEE.

[13] Shadman Sakib et.al, 2021, 05 June.“Performance Analysis of Machine Learning Approaches in Diabetes Prediction.”IEEE.

[14] Sarra Samet, Mohamed Ridda Laouar and Issam Bendib,2022, April.“Use of Machine Learning Techniques to Predict Diabetes at an Early Stage”.IEEE.

[15] Juncheng Ma, 2022. “Machine Learning in Predicting Diabetes in the Early Stage”. 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). IEEE.

[16] Jingyu Xue et.al, 2020. “Research on Diabetes Prediction Method Based on Machine Learning”.Journal of Physics: Conference Series. AINIT.

[17] Battula Prasanth Kumar,2022,05-May “DIABETES PREDICTION AND COMPARATIVE ANALYSIS USING MACHINE LEARNING ALGORITHMS”. International Research Journal of Modernization in Engineering Technology and Science.

[18] Aishwarya Mujumdar and Vaidehi V, 2019. “Diabetes Prediction using Machine Learning Algorithms”.International Conference on Recent Trends In Advanced Computing. ICRTAC 2019.

[19] Abdulhakim Salum Hassan, I. Malaserene and A. Anny Leema,2020, March. “Diabetes Mellitus Prediction using Classification Techniques”.International Journal of Innovative Technology and Exploring Engineering (IJITEE).

[20] Mohammad Abu Tareq Rony et.al,2021, February. “Mining Significant Features of Diabetes Mellitus Through Employing Various Classification Methods”.International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD).IEEE.

[21] Abe, Oluwafemi Samuel, Olumide O. Obe, Olutayo K. Boyinbode, and Olagbuji N. Biodun. "Classifier Algorithms and Ensemble Models for Diabetes Mellitus Prediction: A Review." International Journal 10, no. 1 (2021).

[22] Bello A.Bodinga, 2022, June“On The Analysis of Some Machine Learning Algorithms for the Prediction of Diabetes.International Journal Advanced Networking and Applications. ISSN.

[23] Shubham Sain et.al,2023, March. “Diabetes Prediction Using ML”.International Research Journal of Engineering and Technology (IRJET).

[24] Amisha Singla et.al.,2022, May.“DIABETES PREDICTION MODEL”.International Research Journal of Modernization in Engineering Technology and Science. ISSN