DIABETES MELLITUS PREDICTION – A comparative study

A PROJECT REPORT

Submitted by

Pranav Kannan
CB.SC.I5DAS19024

in partial fulfilment of the requirements for the award of thedegree of

INTEGRATED MASTER OF SCIENCE

IN
DATA SCIENCE

AMRITA SCHOOL OF

ENGINEERINGAMRITA VISHWA

VIDYAPEETHAM

COIMBATORE 641112

JUNE 2023

AMRITA VISHWA VIDYAPEETHAM

AMRITA SCHOOL OF ENGINEERING, COIMBATORE, 641112

BONAFIDE CERTIFICATE

This is to certify that the project report entitled **"Diabetes Mellitus Prediction – A comparative study"** submitted by **Pranav Kannan (Register No: CB.SC.I5DAS19024)** in partial fulfillment of the requirements for the award of the **Degree of Integrated Master of Science** in **DATA SCIENCE** is a Bonafide record of the work carried out at Amrita School of Engineering, Coimbatore.

Signature of the Class Advisor

Designation

Project Coordinator

Designation

Chairperson
Department of Mathematics
Dr. J. Ravichandran

The project was evaluated on:

Examiners

AMRITA SCHOOL OF ENGINEERING

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE - 641 112

DEPARTMENT OF MATHEMATICS

DECLARATION

I, **Mr. Pranav Kannan** (**Register Number-CB.SC.I5DAS19024**), hereby declare that this project report entitled **"Diabetes Mellitus Prediction – A comparative study"**, is the record of the original work done by me. To the best of knowledge this work has not formed the basis for the award of any degree/ diploma/ associateship/ fellowship/ or a similar award to any candidate in any University.

Place: Coimbatore **Signature of the Student**

Date:07-06-2023

COUNTERSIGNED

Dr. Murali Krishna P

Project Advisor

Department of Mathematics

ACKNOWLEDGEMENTS

I would like to express my appreciation and heartfelt gratitude to everyone who contributed to the successful completion of this project.

I would like to thank our respected chairperson **Dr. Ravichandran J**, Department of Mathematics, Amrita School of Engineering, for giving me the golden opportunity to do this wonderful project on the topic of Machine Learning which paved me a way to explore and discover more about the subject.

I would also like to thank my class advisor, **Dr. Murali Krishna P**, Department of Mathematics, Amrita School of Engineering, for being my constant support and guiding me through out.

I would like to acknowledge all my faculty members of Department of Mathematics for their valuable contributions in various aspects of the project and, I would like to appreciate the support and resources provided by our institution.

My heartfelt thanks to all the staff and mentors for their invaluable teachings over the years.

I also thank my parents, family and friends for their unwavering support, love, and encouragement in every step of my academics. I acknowledge their contribution that has been a source of inspiration throughout my studies and during this project work.

Coimbatore, Pranav Kannan

June 2023

CONTENTS

1. ABSTRACT

Diabetes Mellitus is a chronic metabolic disorder that affects millions of individuals worldwide and poses significant health challenges. Early detection and prediction of diabetes can play a crucial role in preventing its complications and improving patient outcomes. This project aims to develop different predictive models for Diabetes Mellitus using machine learning algorithms and clinical data. This project serves as a comparative study between different machine learning classification algorithms to find out which model gives the best accuracy for our dataset. Finally, hybrid models are proposed to increase the accuracy of our base models.

2. INTRODUCTION

Diabetes Mellitus is a chronic disease that affects millions of people worldwide. Early diagnosis and intervention are crucial to preventing complications and improving patient outcomes. In this presentation, we will explore different prediction models to predict the risk of diabetes mellitus in patients. Diabetes can largely be classified into two categories. Type 1 diabetes is an autoimmune disease in which the body's immune system mistakenly attacks and destroys the insulin-producing cells in the pancreas. This results in a lifelong dependence on insulin injections to regulate blood sugar levels. Type 1 diabetes usually develops in childhood or adolescence and requires careful management to prevent complications. Type 2 diabetes is a chronic metabolic disorder characterized by high blood sugar levels due to insulin resistance and inadequate insulin production. It is often associated with obesity, sedentary lifestyle, and genetic factors. Management involves lifestyle modifications, medication, and regular monitoring of blood glucose levels to prevent complications.

2.1 What is Diabetes Mellitus Prediction?

Diabetes Mellitus Prediction refers to the process of using various techniques, such as machine learning and statistical analysis, to develop models that can accurately forecast the risk or likelihood of an individual developing Diabetes.

2.2 Objective/ Problem Statement:

The primary objective of Diabetes Mellitus Prediction is to identify individuals who have a higher probability of developing diabetes in the future. By leveraging predictive modeling techniques, healthcare professionals can identify at-risk individuals at an early stage, allowing for targeted interventions, preventive measures, and personalized treatment plans. Early detection and timely intervention can significantly reduce the risk of complications associated with Diabetes, such as cardiovascular diseases, kidney problems, and nerve damage.

2.3 Concepts used in the project:

2.3.1 KNN-imputing technique:

K-Nearest Neighbors is a popular imputing technique used for handling null or missing values in datasets. It is a simple and intuitive method that relies on the concept of similarity between data points. KNN imputation is particularly useful when dealing with datasets that contain both numerical and categorical variables. The KNN imputation algorithm works by finding the K most similar data points (nearest neighbours) to the observation with the missing value. The missing value is then imputed using the average (for numerical variables) or mode (for categorical variables) of the values from its nearest neighbors.

2.3.2 Feature Engineering:

Feature engineering is the process of transforming raw data into meaningful features that can improve the performance of machine learning models. It involves selecting, creating, and manipulating variables to extract relevant information and patterns. By understanding the domain knowledge and characteristics of the data, feature engineering helps in capturing the most important aspects for model training. Techniques such as scaling, one-hot encoding, binning, and feature extraction can be applied to enhance the predictive power of the model. Skillful feature engineering plays a crucial role in improving model accuracy, reducing overfitting, and enhancing the interpretability of results, ultimately leading to more robust and reliable predictions.

2.3.3 Logistic Regression algorithm:

Logistic Regression is a popular algorithm used for classification tasks. It is a supervised learning method that predicts the probability of an outcome based on input variables. The algorithm assumes a linear relationship between the independent variables and the log-odds of the dependent variable. It applies the logistic function, also known as the sigmoid function, to map the predicted values into probabilities between 0 and 1. By setting a threshold, such as 0.5, predictions can be classified into binary classes. Logistic Regression is widely used due to its simplicity, interpretability, and efficiency for handling large datasets.

2.3.4 Decision Tree algorithm:

The Decision Tree algorithm is a popular machine learning technique used for classification tasks. It constructs a tree-like model by recursively splitting the dataset based on different features, with the goal of maximizing information gain or Gini impurity. Each internal node represents a feature, while each leaf node represents a class label. During training, the algorithm selects the best feature to split the data, creating branches that segregate the instances. In the end, the tree can be used to make predictions on unseen data by following the path from the root node to a leaf node. Decision Trees are interpretable, efficient, and can handle both categorical and numerical data.

2.3.5 Random Forest algorithm:

Random Forest is a powerful algorithm for classification tasks. It is an ensemble learning method that combines multiple decision trees to make predictions. Each tree in the forest is built using a random subset of features and data samples, reducing the risk of overfitting. During prediction, each tree in the forest independently classifies the input, and the final classification is determined by majority voting. Random Forest is effective for handling large datasets with high dimensionality, as it can capture complex relationships and handle missing values. It is known for its robustness, scalability, and ability to handle both numerical and categorical data.

2.3.6 Support Vector Classification algorithm:

The Support Vector Machine (SVM) algorithm is a powerful tool for classification tasks. It works by finding an optimal hyperplane that separates different classes in a dataset. The algorithm maximizes the margin between the hyperplane and the closest data points, known as support vectors, thus enhancing its generalization ability. SVM can handle both linear and non-linear classification problems by using kernel functions to map the data into a higher-dimensional space. This allows SVM to capture complex relationships between features. SVMs are known for their effectiveness in handling high-dimensional data and their robustness against overfitting.

2.3.7 K- Nearest Neighbour algorithm:

The K Nearest Neighbour (KNN) algorithm is a simple yet effective method used for classification tasks. It operates on the principle of proximity, where a data point is assigned, a label based on the majority vote of its nearest neighbors in the feature space. The value of K determines the number of neighbors considered. During the classification process, KNN calculates the distance between the target point and all other points in the training set and selects the K nearest neighbors. It then assigns the label that occurs most frequently among those neighbors to the target point. KNN is a non-parametric algorithm, making it suitable for complex data distributions.

2.3.8 Naïve Bayes classification algorithm:

Naive Bayes is a popular classification algorithm that utilizes Bayes' theorem to predict the probability of an instance belonging to a particular class. It assumes that features are independent and have equal influence on the outcome. Naive Bayes calculates the likelihood of a class given the feature values and the prior probability of the class. It then selects the class with the highest probability as the predicted outcome. This algorithm is efficient, even with large datasets. However, its assumption of feature independence can limit its accuracy when dealing with correlated features.

2.3.9 XG Boost classification algorithm:

XG Boost, short for Extreme Gradient Boosting, is a powerful algorithm widely used for classification tasks. It is an ensemble learning method that combines the predictions of multiple weak classifiers to create a strong classifier. XG Boost iteratively builds decision trees, minimizing the errors of previous trees by using gradients. This approach enhances the model's predictive performance and reduces bias. It incorporates regularization techniques to prevent overfitting and handles missing values efficiently. XG Boost is known for its scalability, speed, and ability to handle large datasets. Its popularity stems from its exceptional accuracy and robustness in various classification scenarios.

2.3.10 Feed Forward Neural Network:

A feed-forward neural network is a type of artificial neural network commonly used for classification tasks. It consists of multiple layers of interconnected nodes called neurons. Each neuron receives input from the previous layer and applies a non-linear activation function to produce an output. During training, the network adjusts its weights through a process called backpropagation, minimizing the difference between predicted and actual class labels. This allows the network to learn complex patterns and make accurate predictions on unseen data, making feed-forward neural networks a powerful tool for classification tasks.

2.3.11 Bagging:

Bagging (Bootstrap Aggregating) is a powerful ensemble learning technique in machine learning. It involves creating multiple subsets of the training data through bootstrapping, where each subset is used to train a separate base model. These base models are then combined by averaging or voting to make predictions. Bagging reduces the variance of the model by introducing diversity in the training data and averaging the predictions. This helps in improving the model's performance and reducing overfitting. By using bagging, the ensemble model becomes more robust and generalizes better to unseen data. It is widely used in various algorithms such as Random Forests and Bagging Meta-Estimators.

2.3.11 Ada-Boosting:

AdaBoost (Adaptive Boosting) is a machine learning technique that combines multiple weak learners into a strong learner. It sequentially trains weak models on different subsets of the training data, assigning higher weights to misclassified instances at each iteration. This ensures that subsequent weak models focus on the previously misclassified samples, improving overall accuracy. The final prediction is made by aggregating the predictions of all weak models, weighted by their individual performance. AdaBoost is robust against overfitting, achieves high accuracy, and is widely used in various applications such as classification and regression tasks. Its adaptive nature makes it a powerful ensemble method in machine learning.

3. LITERATURE REVIEW

The paper **"Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective" by Olisah, C.C., Smith, L., and Smith, M. (2022)** investigates the application of data preprocessing techniques and machine learning algorithms for diabetes mellitus classification. The study aims to improve the accuracy of prediction and diagnosis of diabetes using these methods. The authors evaluate different machine learning models and feature selection techniques on a dataset related to diabetes mellitus. Their findings suggest that the proposed approach yields promising results, with increased accuracy in diabetes prediction and diagnosis. The research highlights the potential of data preprocessing and machine learning techniques in improving the identification and management of diabetes mellitus. The algorithms proposed in this paper include **Random Forest, Support Vector Machine and Deep Neural Network**. [1]

The paper titled **"Voting classification-based diabetes mellitus prediction using hyper tuned machine-learning techniques" by Z Mushtaq, MF Ramzan, S Ali and S. Baseer** aims to explore the use of machine learning techniques for the classification of diabetes mellitus. The authors propose a voting classification approach that combines multiple machine learning algorithms and compares the performance of these models. They also apply hyperparameter tuning to enhance the predictive accuracy of the models. The algorithms proposed in this paper include **Support Vector Machine, Naïve Bayes, K-Nearest Neighbours, Logistic Regression, Random Forest, and Gradient Boost Classifier**. [2]

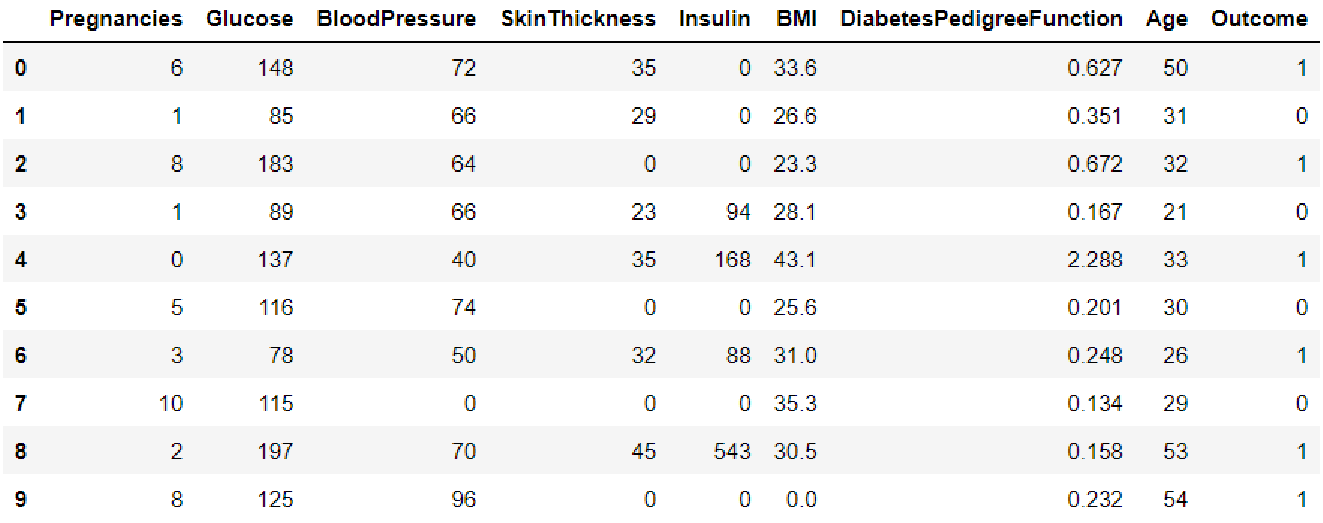
The paper **"A survey on diabetes mellitus prediction using machine learning algorithms" by R. Srivastava and R.K Dwivedi (2022)** provides an overview of various machine learning algorithms used for predicting diabetes mellitus. The study concludes that machine learning techniques, such as **decision trees, support vector machines, random forests, and neural networks**, have been extensively applied for diabetes mellitus classification. [3]

The paper **"Deep Convolutional Neural Network for Diabetes Mellitus Prediction" by Alex, S.A., Nayahi, J.J.V., Shine, H. and Gopirekha, V** investigates the application of a deep convolutional neural network (CNN) for the classification of diabetes mellitus. Through their experiments and evaluation, they conclude that the proposed CNN model achieves promising results in diabetes classification, with high accuracy and precision. The model demonstrates its effectiveness in distinguishing between diabetic and non-diabetic cases, providing a potential tool for early detection and prediction of diabetes mellitus. The findings suggest that deep learning techniques, specifically CNNs, can significantly contribute to improving the accuracy of diabetes diagnosis and management. [4]

4. DATASET

The dataset under consideration originates from the National Institute of Diabetes and Digestive and Kidney Diseases. Its primary purpose is to facilitate the diagnostic prediction of diabetes in patients based on specific diagnostic measurements contained within the dataset. A subset of instances was selected from a larger database, adhering to certain constraints. Notably, all patients included in this dataset are female and at least 21 years old and belong to the Pima Indian heritage.

The dataset contains **768 rows and 9 columns**. The dataset encompasses various medical predictor variables, alongside a target variable known as **"Outcome"** These predictor variables provide valuable information for the prediction task and include factors such as the number of pregnancies a patient has experienced, their body mass index (BMI), insulin level, age, and other relevant attributes.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1

Fig 4.1: Snapshot of the dataset

4.1 Pregnancies:

Pregnancy can increase the risk of developing gestational diabetes, a temporary condition that affects about 2-10% of pregnant women. However, it can also reveal underlying diabetes, such as type 1 or type 2, as hormonal changes during pregnancy can disrupt insulin production and utilization. Proper monitoring and management are crucial during pregnancy to mitigate the effects of diabetes.

4.2 Glucose:

Glucose plays a critical role in the development of diabetes. In individuals without diabetes, glucose levels are regulated effectively, maintaining a balance. However, in those with diabetes, either the body doesn't produce enough insulin (Type 1) or the cells become resistant to insulin (Type 2), resulting in elevated blood glucose levels.

4.3 Blood Pressure:

High blood pressure (hypertension) increases the risk of developing type 2 diabetes. The link between the two conditions is complex, involving shared risk factors such as obesity and insulin resistance. Managing blood pressure through lifestyle changes and medication can help reduce the chances of developing diabetes.

4.4 Skin Thickness:

Skin thickness does not directly impact the chances of having diabetes. However, individuals with thicker skin may have difficulties in accurate blood glucose monitoring through devices like glucose meters, as they may require longer or stronger needle penetration, potentially affecting diabetes management.

4.5 Insulin:

Insulin plays a crucial role in regulating blood sugar levels. In individuals with diabetes, either the body does not produce enough insulin, or the insulin produced is ineffective. By administering insulin externally, either through injections or pumps, it helps control blood glucose levels and reduces the chances of complications associated with diabetes.

4.6 BMI:

BMI (Body Mass Index) is a measure of body fat based on height and weight. Higher BMI levels are associated with an increased risk of developing diabetes. Excess body weight can lead to insulin resistance, a key factor in the development of type 2 diabetes, making maintaining a healthy BMI crucial for reducing diabetes risk.

4.7 Diabetes Pedigree Function:

The Diabetes Pedigree Function (DPF) is a measure that assesses the genetic predisposition for diabetes by analyzing the family history of an individual. A higher DPF score indicates a greater likelihood of developing diabetes, highlighting the impact of genetic factors on the chances of having the condition.

4.8 Age:

Age is a significant factor in the likelihood of developing diabetes. As individuals grow older, their risk of developing type 2 diabetes increases. This is primarily due to age-related changes in the body's metabolism and decreased insulin sensitivity, making older adults more susceptible to the disease.

5. METHADOLOGY

The approach to this project has been carried out in four phases.

5.1 Base Models implementation:

In the first phase of this project, the focus is on implementing fundamental machine learning classification models using a pre-processed dataset. Models such as Logistic Regression, Decision Tree, Random Forest, Support Vector Classifier, K-Nearest Neighbor, and Naïve Bayes are employed, and their accuracies are assessed to determine the most suitable algorithm for the dataset.

5.2 Feature Engineering:

The second phase of the project involves enhancing the existing dataset by employing aggregation techniques to introduce new features. The same models utilized in Phase 1 are then deployed, and their respective accuracies are compared for evaluation purposes.

5.3 Ensemble Techniques:

In the third phase of this project, Ensemble techniques, such as bagging and boosting, are employed on the base models acquired in Phase 1, and their accuracies are evaluated and compared. These methods combine multiple models to enhance predictive performance and provide valuable insights for decision-making and analysis.

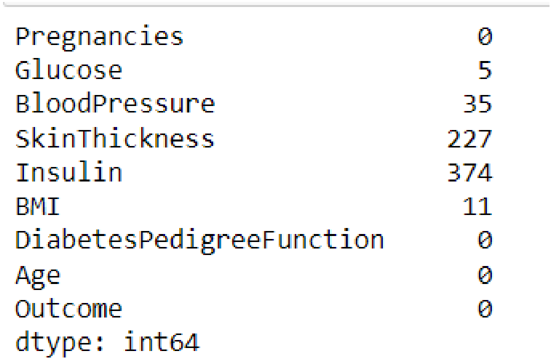
5.4 Hybrid Models Proposal:

During the fourth phase of the project, an analysis of the accuracies achieved in the previous three phases is conducted. Hybrid models, which combine two machine learning algorithms or one machine learning and one deep learning algorithm, are deployed and their accuracies are observed. Ultimately, the best hybrid model is selected as the outcome of the project.

6. IMPLEMENTATION

6.1 Pre-Processing:

During the initial examination of the dataset, no null values were apparent. However, upon further investigation, it became apparent that null values were disguised as 0s. To address this issue, the code replaces the 0 values in the 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', and 'BMI' columns with 'NaN'.



```
Pregnancies                   0
Glucose                       5
BloodPressure                35
SkinThickness               227
Insulin                     374
BMI                          11
DiabetesPedigreeFunction      0
Age                           0
Outcome                       0
dtype: int64
```

Fig 6.1: Snapshot of null values in the data

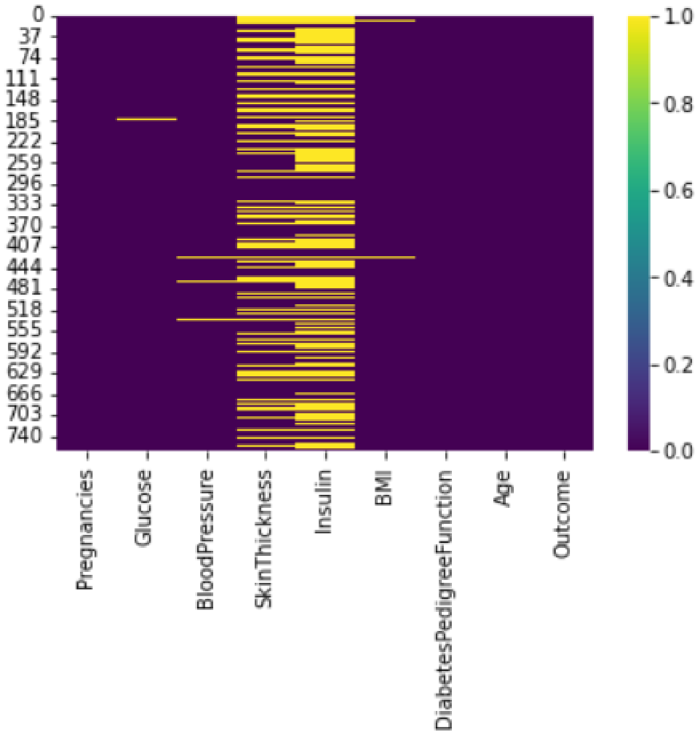


Fig 6.2: Heatmap representing null values in the data.

To handle the null values present in our data, two different approaches are employed.

6.1.1 Mean imputation:

For the columns – **'Glucose', 'BloodPressure', and 'BMI'**, mean imputation is used to handle the null values. It replaces the NaN values in the 'Glucose', 'BloodPressure', and 'BMI' columns with the respective mean values of those columns, providing a reasonable estimate for the missing data based on the available values.

6.1.2 KNN imputation:

The K-nearest neighbors (KNN) imputation method is used to fill the missing values in **the 'SkinThickness' and 'Insulin'** columns. KNN imputation replaces the missing values with estimates obtained from the values of the nearest neighbors, considering the five closest neighboring data points.

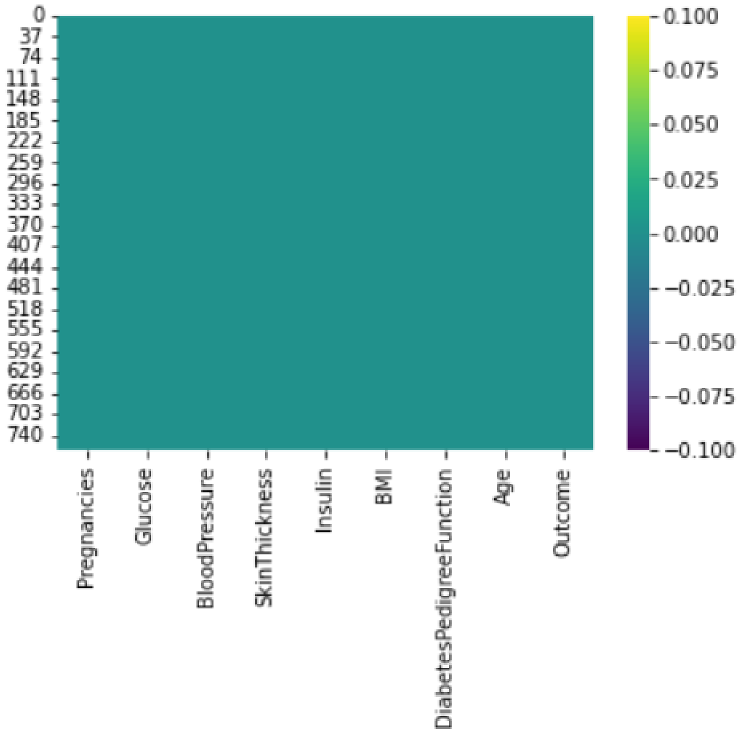


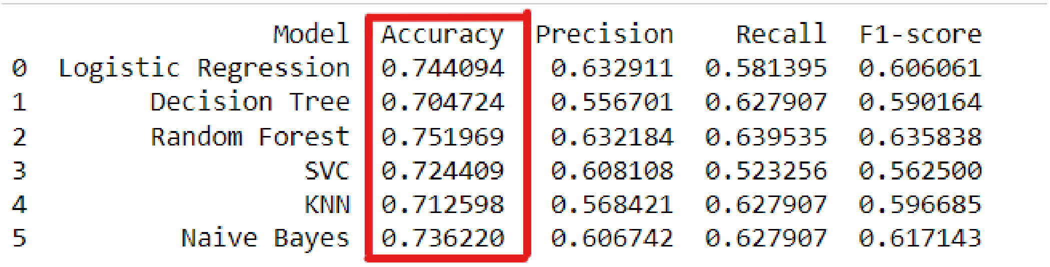
Fig 6.3: Heatmap showing no null values after pre-processing.

6.2 Train Test split:

To ensure consistency throughout our project's different phases, we have implemented a standardized train-test split. The dataset has been divided such that 80% of the data is allocated for training our models, while the remaining 20% is reserved for testing and evaluating their performance. This approach ensures reliable and comparable results across all experiments.

6.3 Base Model implementation:

In the context of our project, the implementation of base models serves as a fundamental step in building and evaluating predictive models. As a crucial initial step, we have implemented six base models: Logistic Regression, Decision Tree, Random Forest, Support Vector Classifier, K-Nearest Neighbor, and Naïve Bayes. These models were applied to pre-processed data, and their respective accuracies were measured, tabulated, and plotted for evaluation purposes.



```
                 Model  Accuracy  Precision    Recall  F1-score
0  Logistic Regression  0.744094   0.632911  0.581395  0.606061
1        Decision Tree  0.704724   0.556701  0.627907  0.590164
2        Random Forest  0.751969   0.632184  0.639535  0.635838
3                  SVC  0.724409   0.608108  0.523256  0.562500
4                  KNN  0.712598   0.568421  0.627907  0.596685
5          Naive Bayes  0.736220   0.606742  0.627907  0.617143
```

Fig 6.4: Accuracies of the base models

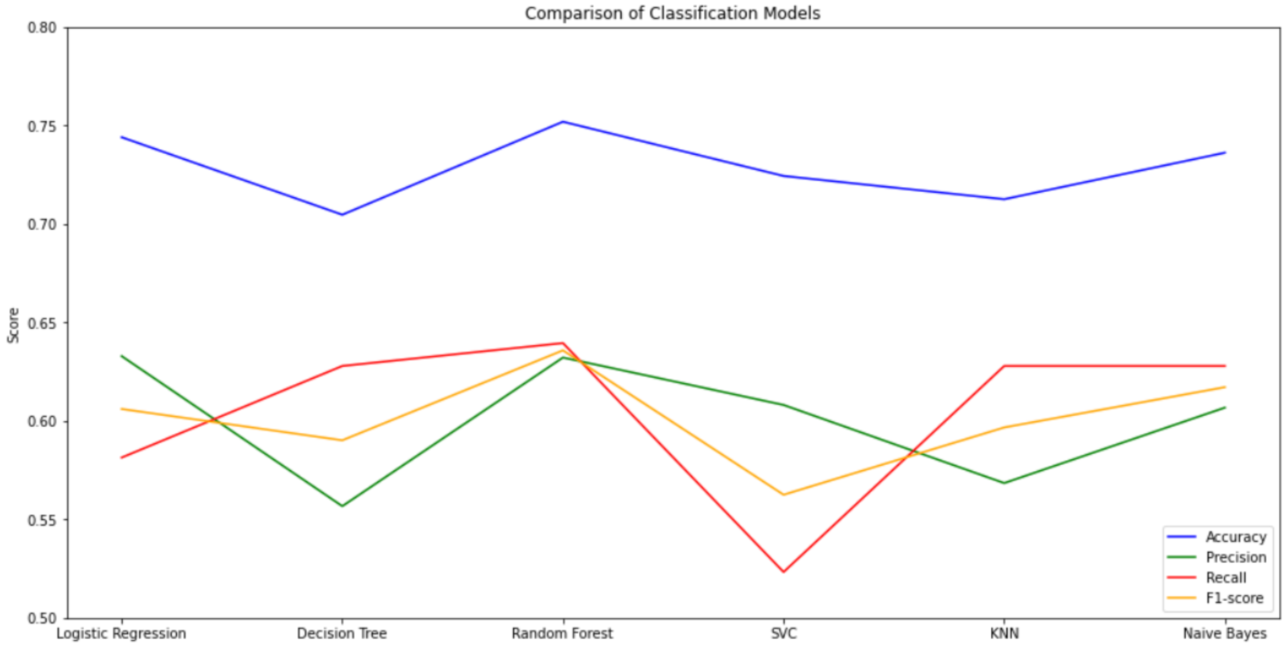


Fig 6.5: Line Graph showing the performance of different base models.

6.4 Feature Engineering:

In an attempt to increase the accuracy of the models, feature engineering is performed to the pre-processed dataset and new features are incorporated using aggregation. These new features are selected based on their correlation values with the target variable.

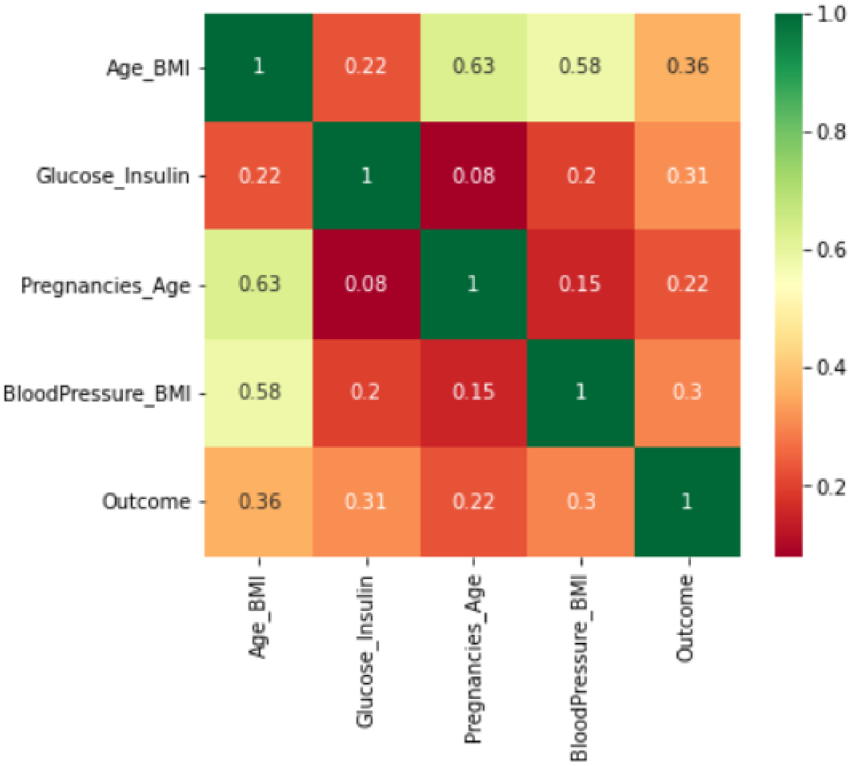


Fig 6.6: Correlation of the new features with respect to Outcome

6.4.1 Age_BMI:

Higher values of Age_BMI indicate increased age and higher body mass index. Thus, it can be hypothesized that the product of Age_BMI has a positive association with the likelihood of having diabetes.

6.4.2 Glucose_Insulin:

As higher values of the product of Glucose and Insulin indicate increased metabolic activity, it can be hypothesized that a higher Glucose_Insulin value may be associated with an increased chance of having diabetes, as it suggests a potential imbalance in glucose regulation and insulin resistance, both of which are key factors in the development of diabetes.

6.4.3 Pregnancies_Age:

Higher values of Pregnancies_Age may increase the likelihood of having diabetes. This assumption is based on the belief that a combination of higher pregnancy count and older age may contribute to metabolic changes and increased insulin resistance, potentially influencing the chances of developing diabetes.

6.4.4 BloodPressure_BMI:

Individuals with a higher 'BloodPressure_BMI' value may have an elevated chance of having diabetes compared to those with lower 'BloodPressure_BMI' values due to the potential correlation between hypertension, obesity (BMI), and diabetes.

6.4.5 Observations:

Although the new features engineered in our dataset may not exhibit strong correlations, they still have a modest impact on the outcome variable. Following feature engineering, we proceed to evaluate the performance of the same models used in section 6.2. By comparing the accuracies obtained in both phases, we can assess the effectiveness of the feature engineering process.

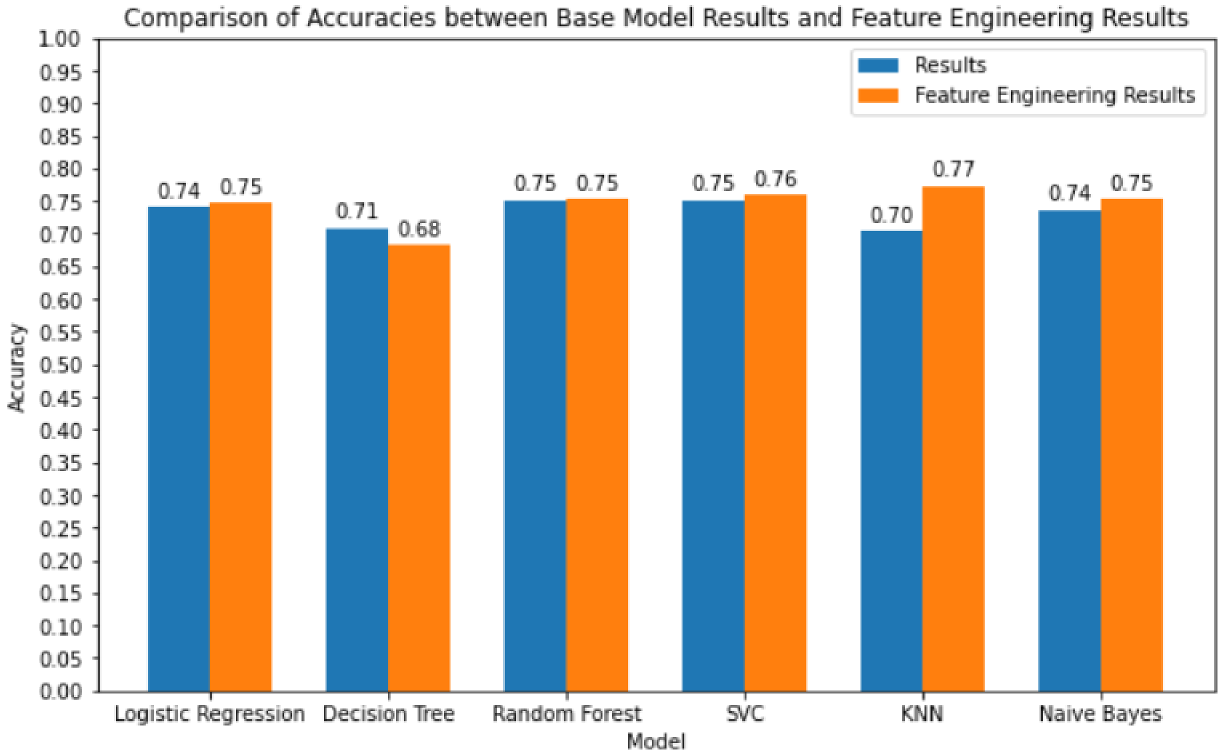


Fig 6.6: Figure showing Comparison of accuracies between Base Models and Feature Engineered Models

6.5 Ensemble Techniques:

The next phase of the project will leverage powerful ensemble techniques such as bagging and boosting to significantly enhance the accuracy of our base models. By integrating these advanced techniques, we aim to achieve even better performance and predictive capabilities.

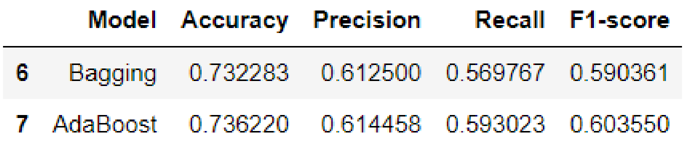
6.5.1 Bagging:

Bagging, short for Bootstrap Aggregating, is a powerful ensemble learning technique in machine learning. It aims to improve the predictive accuracy and stability of models by combining multiple individual models. Bagging works by creating multiple subsets of the training data through bootstrapping, where each subset is used to train a separate model. These models are then combined by averaging or voting to make predictions. By leveraging the diversity among the models, bagging reduces overfitting and increases generalization. It is widely used in various algorithms, such as random forests, where each tree is trained on a different subset of the data.

6.5.2 Ada Boost:

AdaBoost, short for Adaptive Boosting, is a powerful machine learning technique used for classification problems. It combines multiple weak classifiers to create a strong classifier. Each weak classifier is trained on a subset of the training data, with weights assigned to each instance. During training, AdaBoost adjusts the weights of misclassified instances, giving more importance to those that were classified incorrectly. The final classification is determined by the weighted combination of the weak classifiers. AdaBoost is effective in handling complex datasets and achieving high accuracy. Its adaptability and ability to focus on difficult instances make it a popular choice in machine learning applications.

6.5.3 Observations:



	Model	Accuracy	Precision	Recall	F1-score
6	Bagging	0.732283	0.612500	0.569767	0.590361
7	AdaBoost	0.736220	0.614458	0.593023	0.603550

Fig 6.7: Accuracies after bagging and boosting

6.6 Hybrid Model Proposals:

As the final phase of the project, three different hybrid models have been developed and tested on the dataset.

6.6.1 XG Boost and Random Forest:

The XG Boost and Random Forest hybrid model combines the strengths of both algorithms to improve predictive performance. In this ensemble approach, the Random Forest algorithm serves as the base model, generating multiple decision trees that vote on the final prediction. The XG Boost algorithm acts as the meta-model, learning from the errors of the base model and optimizing the ensemble's predictions. The XG Boost algorithm assigns weights to the base model's predictions based on their accuracy, thereby improving the overall model's accuracy. By combining the robustness of Random Forest with the boosting capabilities of XG Boost, this hybrid model can achieve enhanced predictive power and generalization.

6.6.2 Logistic Regression and Decision Trees:

Logistic Regression and Decision Trees are two popular machine learning algorithms used for classification tasks. In a hybrid model, these algorithms can be combined to create a more powerful and accurate model. The hybrid model works by training a base model, such as logistic regression, on the input data to make predictions. The predictions from the base model are then used as features for a meta model, which is often a decision tree. The meta model takes the base model's predictions as inputs along with other features and learns how to make the final classification decision. This hybrid approach leverages the strengths of both algorithms. Logistic regression is good at modeling linear relationships, while decision trees can capture non-linear and complex patterns in the data. By combining them, the hybrid model can handle a wide range of classification problems and potentially achieve better performance than using either algorithm alone.

6.6.3 Random Forest and Feed Forward Neural Network:

A hybrid model combining Random Forest (RF) and Feed Forward Neural Network (FFNN) leverages the strengths of both algorithms to enhance prediction accuracy. The RF serves as the base model and constructs an ensemble of decision trees using bootstrapped data samples and random feature subsets. Each tree independently predicts the outcome, and the final prediction is determined through voting or averaging. The FFNN acts as the meta model and takes the predictions of the RF as inputs. It is trained to learn the relationships between the RF predictions and the target variable, capturing complex nonlinear patterns. The hybrid model thus combines the RF's interpretability and robustness with the FFNN's ability to capture intricate patterns, resulting in improved prediction performance.

6.6.4 Observations:

After conducting an extensive analysis on the performance of various hybrid models, it becomes evident that the combination of Random Forest classifier and XG Boost yields a significantly higher accuracy than the other hybrid models, as shown in the plot given below.

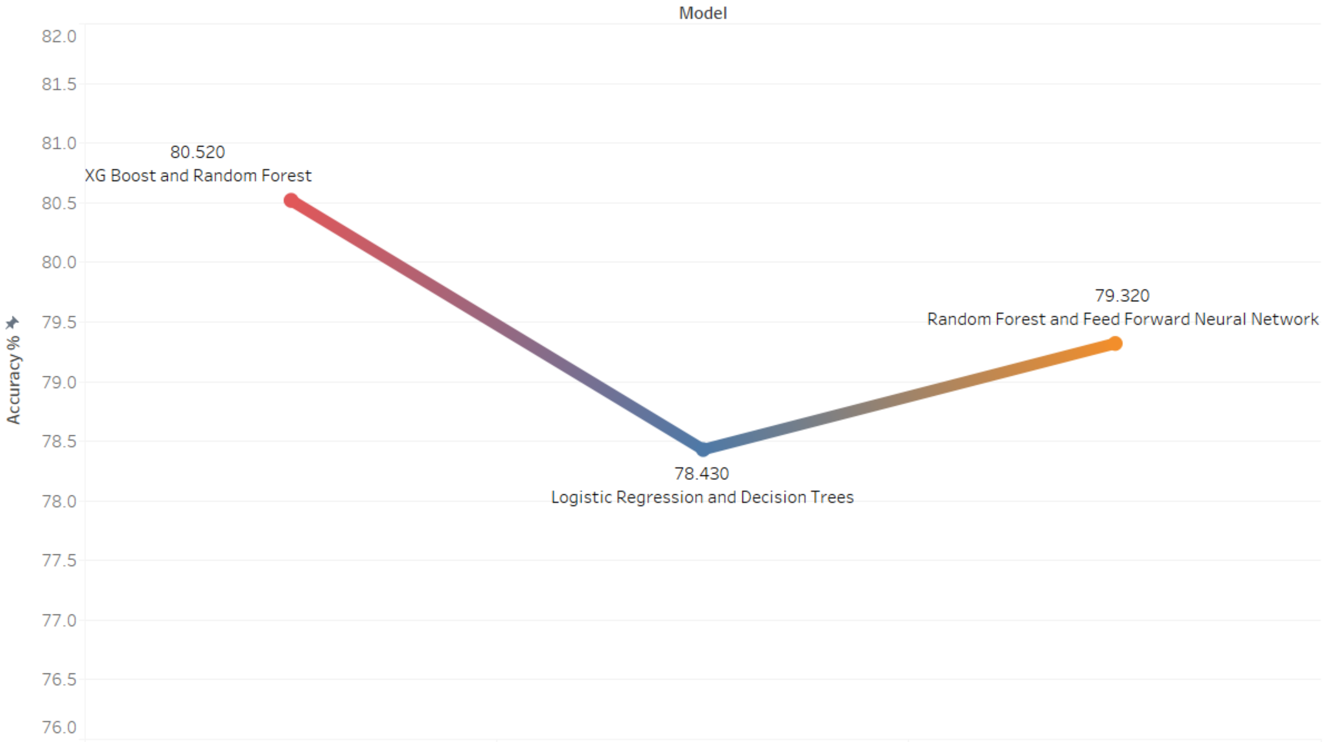


Fig 6.8: Figure showing Comparison of accuracies between the different hybrid models.

7. RESULTS

The objective of this project was to propose a suitable hybrid architecture that will help in Diabetes Mellitus Prediction. After the experiments conducted, it is evident that a hybrid architecture of **XG Boost and Random Forest** gives us the best accuracy (80.52%) to classify the dataset.

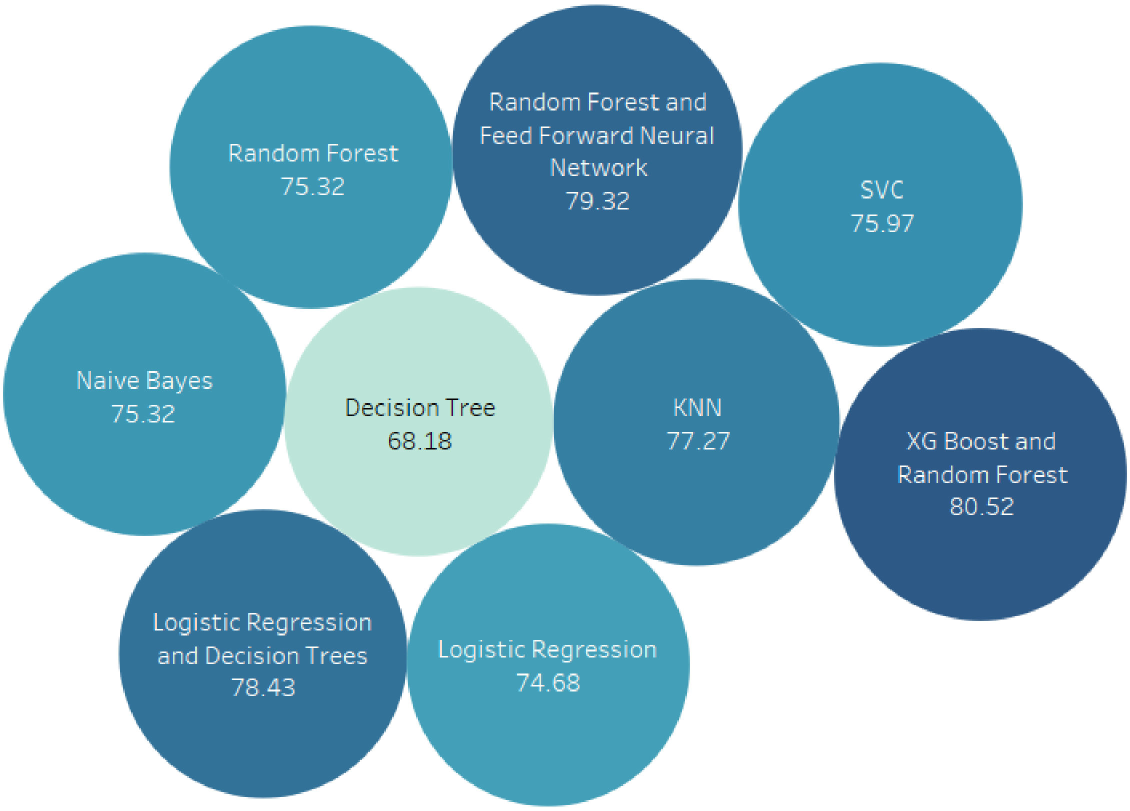


Fig 7.1: Hybrid model gives a significantly higher accuracy than the previous models.

8. CONCLUSION

In conclusion, predicting and diagnosing diabetes mellitus is a complex task that requires a multidimensional approach. Various risk factors, including genetics, lifestyle, and medical history, play significant roles in determining an individual's susceptibility to the disease. Machine learning algorithms and predictive models have shown promise in assisting healthcare professionals by analyzing large datasets and identifying patterns that may indicate the likelihood of developing diabetes.

By leveraging features such as age, body mass index (BMI), glucose levels, blood pressure, and cholesterol levels, predictive models can estimate an individual's risk of developing diabetes mellitus. These models can help healthcare providers prioritize preventive measures, lifestyle modifications, and early interventions for high-risk individuals. Additionally, they can assist in allocating healthcare resources effectively and developing personalized treatment plans.

However, it is important to note that while predictive models can be valuable tools, they should not be viewed as definitive diagnostic tools. The accuracy and reliability of these models depend on the quality and representativeness of the data used for training. Furthermore, diabetes mellitus is a complex and multifactorial disease, and there may be factors that are not captured in the models, limiting their predictive capabilities.

Therefore, it is crucial to combine predictive models with clinical judgment and additional diagnostic tests for an accurate assessment of an individual's diabetes risk. Regular screenings, healthy lifestyle choices, and ongoing monitoring remain essential in managing and preventing diabetes mellitus. Continued research and advancements in machine learning techniques can further enhance our ability to predict and prevent this prevalent and chronic condition.

9. REFERENCES

[1]
Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective
Olisah, C.C., Smith, L., and Smith, M
https://www.sciencedirect.com/science/article/pii/S0169260722001596
Year: 2022

[2]
Voting classification-based diabetes mellitus prediction using hyper tuned machine-learning techniques
Z Mushtaq, MF Ramzan, S Ali and S. Baseer
https://www.hindawi.com/journals/misy/2022/6521532/
Year: 2022

[3]
A survey on diabetes mellitus prediction using machine learning algorithms.
R. Srivastava, R. K Dwivedi
https://link.springer.com/chapter/10.1007/978-981-16-5987-4_48
Year: 2022

[4]
A survey on diabetes mellitus prediction using machine learning algorithms.
Alex, S.A., Nayahi, J.J.V., Shine, H. and Gopirekha, V
https://idp.springer.com/authorize/casa?redirect_uri=https://link.springer.com/article/10.1007/s00521-021-06431-7&casa_token=vPfZuexZhi4AAAAA:b0PjK7s9JT3nPcsMWozYyPffeI8MyDOyWwtOFwjdh3qJx2R4SAtrN2qvZ-mwIWyj7pTFDTq_4Xlgq8103wE
Year: 2022