"Machine Learning Approach For Early Detection Of Diabetes"

BY

IFTYAR ALAM

ID: 193-15-3001

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the **Degree of Bachelor of Science in Computer Science and Engineering**

Supervised By

Md Sazzadur Ahmed
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Mr Abdus Sattar
Associate Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
17 SEPTEMBER, 2025

APPROVAL

This Project titled "Machine Learning Approach For Early Detection Of Diabetes", Submitted by "Iftyar Alam" to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 17 SEPT 2025.

BOARD OF EXAMINERS

Dr. Sheak Rashed Haider Noori
Professor and Head
Chairman, Department of CSE, DIU
Daffodil International University

Chairman

Dr. Md. Zahid Hasan
Associate Professor
Department of Computer Science and Engineering
Daffodil International University

Internal Examiner

Samia Nawshin
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Internal Examiner

Dr. Md. Arshad Ali
Professor
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science & Technology University

External Examiner

DECLARATION

I hereby declare that, this project has been done by us under the supervision of Md Sazzadur Ahmed, Assistant professor, Department of Computer Science and Engineering, Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Md. Sazzadur Ahmed
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Dr. Abdus Sattar
Associate Professor
Department of CSE
Daffodil International University

Submitted by:

Iftyar Alam

ID: 193-15-3001
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

Firstly, I would want to thank Allah Ta'ala for His great grace, which enabled me to successfully finish my final year project and internship.

The **Sazzadur Ahmed, Assistant Professor,** Department of CSE Daffodil International University, Dhaka has my sincere gratitude and our deepest gratitude. Our supervisor has extensive knowledge of and a strong interest in "machine learning" in order to complete this job. This endeavor has been made possible by his unending patience, academic assistance, unceasing encouragement, vigorous and continuous supervision, constructive criticism, insightful counsel, and his reading of several subpar versions and their revision at every level.

Our sincere thanks go out to **Sazzadur Ahmed Sir** and **Mr Abdus Sattar Sir** of the CSE Department for their kind assistance in completing my project, as well as to the other instructors and personnel of Daffodil International University's CSE department.

I want to express my gratitude to every student at Daffodil International University who participated in this discussion while finishing the course assignment.

Lastly,I must respectfully thank our parents for their unwavering patience and support. The study therefore demonstrates significant potential for the integration of computer science and medicine to enable early identification of hazardous conditions. Diabetes is one of the most widespread chronic diseases in the world today. Millions of people are affected by it,and the numbers continue to grow each year. What makes diabetes especially dangerous is that many people do not know they have it until serious health problems arise. Early detection is therefore extremely important, as it gives patients a better chance to manage or even prevent severe complications.

ABSTRACT

Technological advancements have led to a closer relationship between the medical industry and machine learning. This study employs machine learning to predict the occurrence of diabetes, a global disease. The aim is to identify the disease in its initial stages to facilitate easier treatment or management of the condition. I have utilized a dataset containing nine features and 100,000 instances. This dataset contains information on hypertension, blood glucose level, BMI, age, smoking history, heart disease, gender, and HbA1c level. These are the main indicators of diabetes. I utilized the Random Forest Classifier to predict the illness. The results of my research have been compared to those from other machine learning methods, including Decision Tree Classifier, Logistic Regression, and KNN. Among these techniques, the Random Forest Classifier exhibited the highest accuracy (95.67%) and AUC score (0.97), confirming its robustness for diabetes prediction. While the findings are promising, the research also identifies important gaps: limited dataset diversity, lack of interpretability of models, minimal clinical integration, and low accessibility for end-users. Recognizing these gaps highlights opportunities for future work, such as incorporating more inclusive datasets, developing explainable AI approaches, and building user-friendly mobile or clinical applications.

The study therefore demonstrates significant potential for the integration of computer science and medicine to enable early identification of hazardous conditions. Diabetes is one of the most widespread chronic diseases in the world today. Millions of people are affected by it, and the numbers continue to grow each year. What makes diabetes especially dangerous is that many people do not know they have it until serious health problems arise. Early detection is therefore extremely important, as it gives patients a better chance to manage or even prevent severe complications.

Recent advances in technology, especially in the field of **machine learning (ML)**, have opened new opportunities for healthcare. By analyzing large amounts of medical data, machine learning algorithms can identify patterns that are not always visible to doctors during routine checkups. This study focuses on using machine learning to predict the occurrence of diabetes, with the hope of assisting healthcare providers and patients in catching the disease earlier.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Introduction

Diabetes has become a global health problem. Millions of lives have been affected and it is now a major concern for the health system.. Reports suggest that there are at least 500 million people with diabetes worldwide. By 2030, this figure could reach 25 percent and 51 percent, respectively. According to WHO data from World Diabetes Day in 2016, 422 million people worldwide suffer from diabetes, with 1.6 million reported deaths [3]. There are three main types of diabetes: gestational, type 1, and type 2 [4]. As for type 1 diabetes, this condition often begins as an inflammatory disease before the age of 20. This type of diabetes is characterized by damage to the cells of the pancreas that produce insulin. Insulin resistance develops in the body's organs when you have type 2 diabetes, increasing the need for insulin. Right now, the pancreas isn't producing enough insulin. Pregnant women are more likely to develop gestational diabetes due to insufficient insulin production by the pancreas. Each of these different types of diabetes requires treatment.[5]. The precocious detection in this state is essential for improving the patient's situation and succeeding in treating the disease. Machine learning is one of several sub-fields of artificial intelligence that has demonstrated tremendous impact and promise in many fields, including healthcare. This project delves deeply into the predictive capabilities of machine learning for diabetes. Our aim is to assist patients and enhance early diagnosis of this disease. The convergence of computer technology and medical informatics opens up new opportunities in healthcare. Machine learning can be used to gain deeper insights into large, complex data sets. The perfect combination of these two elements promises a better future for early diagnosis of diabetes. Previous studies have shown that fat cell size and anthropometric measurements can be used to predict diabetes risk using standard statistical methods. [6] [7]

[8][9]. Research has shown that data mining can be an essential component in understanding the unknowns and possibilities of models developed for the medical field.[10][11][12] This study examines the current status of diabetes prediction using a method related to automatic learning. We examined related research to fully understand the latest methods, algorithms, and functions used in the same research. He also explained the approach to data collection, preparation and creation of models. The results of this study could affect the decisions of manufacturers, researchers, and medical experts that work to improve diabetes and preventive management initiatives.

1.2 Motivation

Urgency of Innovation:

1. The increasing prevalence of diabetes necessitates urgent development of innovative methods for early identification and treatment.

2. Traditional diagnostic methods may not fully capture the complex interactions among diabetes risk factors, underscoring the need for advanced models.

Specific Focus on Bangladesh:

*Awareness of the impact of diabetes is low in Bangladesh and the risk is increased due to lack of understanding among the people.

*There is an urgent need to introduce effective early diagnosis methods tailored to local problems and risk factors.

Role of Machine Learning:

- Machine training offers revolutionary potential in improving diabetes prediction.
- The study is aimed at solving the complexity of the risk factors of diabetes (biomarkers, medical history, demography) and the provision of personalized treatment options.

Utilization of Machine Learning Techniques:

- Techniques such as random forest and logistic regression are used to develop accurate predictive models.
- The goal is to provide reliable predictions that will facilitate earlier treatment and significantly improve patient outcomes.

Impact on Healthcare Delivery:

- The research aims to advance diabetes predictions and revolutionize healthcare delivery by leveraging the accuracy of machine learning algorithms.
- The approach is proactive, personalized, and based on the latest machine learning techniques, and promises to pave the way for transformative changes in diabetes management.

1.3 Objective

- **Focus on Machine Learning:** The research aims to use machine learning techniques for accurate and early prediction of diabetes.
- **Developing a robust model:** By exploring different algorithms and analyzing diverse patient datasets, the goal is to develop a robust predictive model.
- **Timely Information for Healthcare:** These templates are designed to provide healthcare professionals with timely information to drive proactive interventions in diabetes management.
- **Improved Diagnosis and Treatment:** The ultimate goal is to improve the accuracy of diabetes diagnosis, leading to more personalized and effective treatment plans depending on the widespread prevalence and importance of diabetes.

Research Questions

- When I began this research project, I had many questions. There weren't many problems, but some questions remained unanswered. For example:
- Which machine learning method is most effective at predicting diabetes: logistic regression, random forest, or k-nearest neighbor
- Which specific information such as age, medical history, and other factors help accurately predict the likelihood of diabetes
- How can we make the results of machine learning models easier to understand and help doctors prevent diabetes in their patients

Expected Outcome

- For better detection of diabetes.
- Select an algorithm for this dataset.
- The premise of the data is set to discover the best results as much as possible.
- Ablation research on the best performance model.
- Performance analysis and statistical analysis based on confusion matrix.

CHAPTER 2

BACKGROUND

2.0 Introduction

Diabetes mellitus is a major global health issue, and early prediction is essential for prevention and management. Recent studies have applied machine learning and data mining techniques such as Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, and Neural Networks to predict diabetes using clinical and lifestyle data. Datasets like the Pima Indians Diabetes Database are widely used in these works.
The literature shows that predictive models can support healthcare professionals by identifying individuals at risk before severe complications arise. However, challenges remain, including data imbalance, limited generalization across populations, and the need for better feature selection. Recent advances in ensemble methods and deep learning have improved accuracy and reliability. Overall, research highlights the strong potential of predictive models to complement traditional diagnosis and improve early detection of diabetes.

2.1 Literature Review

The name of my article is the prediction of diabetes that uses the path to early intervention. The purpose is to detect the disease in an early stage according to a specific situation, and taking appropriate drugs can help prevent illness. Additionally, we have collected some articles on this topic that reflect works on the same topic, summaries of which have been added below:
Research conducted by Veena Vijayan V. Anjali S [13] indicates that diabetes occurs due to hyperglycemia. Various computer information systems have been developed to predict and diagnose diabetes using different classifiers. The choice of the right classifier considerably increases the accuracy and efficiency of the system. In this context, a decision assistance system is proposed to use the Ad boost algorithm with a decision strain as a basic classification In addition, Support Vector Machine, Naive Bayes, and Decision Tree are also used as base classifiers in the AdaBoost algorithm to check the accuracy. The AdaBoost algorithm using decision stumps as the base classifier achieved an accuracy of 80.72%, outperforming the accuracy of Support Vector Machine, Naive Bayes, and Decision Tree. This suggests that the proposed system, particularly with AdaBoost, can effectively contribute to the accurate prediction and diagnosis of diabetes.
Another research done by P. Sonar and K. Jaya Malini [14] stated that diabetes, being a life-threatening disease, can lead to various complications such as heart failure, blurred vision, kidney disease, etc. Diabetes care often involves frequent visits to diagnostic centers for consultations and reports, which requires both time and financial investment from patients. However, the advent of machine learning offers a promising solution. They discovered that by utilizing cutting edge data processing systems and techniques, it is possible to predict whether a person is at risk of developing diabetes, which is essential for early intervention so that timely preventive measures can be taken. The study focuses on developing an advanced system that can accurately assess a patient's

diabetes risk level.
Another paper by AiswaryaIy er, S. Jeyalatha and Ropak Somali on diagnosing diabetes using classification analysis techniques [5] reflects the devastating outcome of the disease. In their study, they used decision trees and naive Bayes algorithms to predict and diagnose diabetes. They used different methods such as validation and percentage separation to achieve different levels of accuracy. Specifically, they obtained an accuracy of 74.86% using the decision tree, 76.956% using the validation technique, and 79.56% using the naive Bayes algorithm.
The experimental results demonstrate that KNN exhibits the highest accuracy among these techniques which is 89%. This underscores the potential of data mining, particularly through KNN, in providing accurate and early predictions for diabetes, thereby contributing to more effective healthcare practices.
Another paper published by Ionic Kavakiotis et al. [16] attempted to synchronize healthcare and biotechnology. It is stated that significant advances in the fields of biotechnology and medicine have led to the generation of vast amounts of data, including high-throughput genetic data and clinical information from electronic health records (EHRs). In response, the application of machine learning and data mining techniques in life sciences has become critical to intelligently transform this wealth of information into valuable knowledge. Diabetes mellitus (DM) is a group of metabolic diseases with global impact on human health and has been extensively studied in various aspects.
Mercado, F., Nardone, V., and Santana, A. suggested that feature selection is a good choice to improve accuracy. They used Multilayer Perception, Baye Net, Random Forest, Hefting Tree, Decision Tree, and Grip in their study. They showed that Hefting Tree was the most suitable for their study [17].

Another study conducted by Osman A.H. et al [18] combined K-means clustering and support vector machine to predict diseases. The research reveals that Diabetes has become a leading cause of death worldwide. The abundance of medical information has spurred the search for effective tools to aid physicians in accurately diagnosing diabetes. This study focuses on improving diagnostic accuracy and reducing misclassification by identifying important features associated with diabetes. The selection of these features is critical to the performance of a classifier that was developed using knowledge discovery techniques to solve the classification problem of diabetic patient data. The study proposes a comprehensive approach combining SVM (Support Vector Machine) technique with K-means clustering algorithms for diabetes diagnosis. The experimental results indicate high accuracy in distinguishing patterns between diabetic and non-diabetic patients, surpassing current diagnostic methods in terms of performance indicators. The T test statistical method showed significant improvement results, especially when applied to the UCI Pima Indian standard dataset using the Kernelized SVM (KSVM) technique. This suggests a promising avenue for more accurate and reliable diabetes diagnosis, thereby contributing to advancements in healthcare practices.

2.2 Overview of Existing Works

TABLE 2.1: EXISTING WORKS

Author	Algorithm	Performance Measure
Veena Vijayan V.and AnjaliC [2]	AdaBoost	80.72%
P. Sonar and K. Jaya Malini[3]	Decision Tree Classifier	85%

Aiswarya Iyer ,S. Jeyalatha and Ropak Somberly [4]	Naïve Bayes	79.56%
M Committal.[8]	KNN	89%

Many works have been carried out worldwide on this particular subject in order to study and study more me thods to diagnose the disease. Various criteri a were supported for research .The previous works mentioned in the table have used different algorithms for their own work .But as far as I have read Random Forest Classifier haven ot been used or found as the best algorithm to be followed.

2.3 Challenges

- **Dataset:** The most difficult part of this research is choosing the right dataset.
- **Algorithm Selection:** The object detection model can be trained using a wide variety of algorithms. Therefore, choosing the best algorithm was a difficult task
- **Preparing data:** A challenge was getting the image data ready for training. As a result, preparing the entire dataset was a difficult task.
- **Implementing:** The model's implementation is another difficult task. Due to the high cost of the testing equipment, I used local computers to measure performance.

2.4 Gap Analysis

- Despite the promising results of this study, several research and practical gaps remain in the field of diabetes prediction using machine learning.

- Data Gaps
  - Most available datasets (including the one used in this study) are secondary and limited in demographic diversity.
  - Features such as lifestyle, genetics, and long-term patient history are underrepresented.
  - Imbalance between diabetic and non-diabetic cases still limits real-world applicability, even after using SMOTE.

- Methodological Gaps
  - Many models, including Random Forest, provide high accuracy but lack explain ability, making clinical adoption difficult.
  - Overfitting risks remain when training models on relatively small or biased datasets.
  - Limited comparison with deep learning models (e.g., CNNs, LSTMs) that may capture complex patterns more effectively.

- Clinical Integration Gaps- Current prediction models operate in an offline, research-oriented environment and are not yet integrated into real-time clinical workflows.
  - Lack of interoperability with Electronic Health Records (EHRs) and wearable devices restricts practical deployment.
  - Healthcare professionals need interpretable outputs (risk factors, explanations), which current models rarely provide.

- User-Centered Gaps
  - Existing models are designed mainly for researchers, not end-users (patients or doctors).
  - Limited availability of mobile/web applications for easy diabetes risk assessment.
  - Low awareness among patients in countries like Bangladesh about predictive tools.

Summary:

- This analysis shows that while this study achieved high accuracy (95.67% with Random Forest), there remain gaps in dataset diversity, interpretability, real-time usability, and accessibility. Addressing these will make diabetes prediction models more reliable, equitable, and clinically useful.

GAP ANALYSIS TABLE REPORT

TABLE 2.4.1 Analysis Table

Features / Gaps	Current Research Models	Proposed System
Diverse dataset with demographic coverage	No	Yes
Inclusion of lifestyle, genetics, and long-term history	No	Yes
Handling of class imbalance effectively	Partial	Yes
Model explainability (clinician-friendly)	No	Yes
Risk of overfitting minimized	No	Yes
Use of deep learning (CNN, LSTM, hybrid)	Limited	Yes
Real-time clinical workflow integration	No	Yes
Interoperability with EHRs & wearable devices	No	Yes
Interpretability of outputs for doctors	No	Yes
Mobile/web accessibility for patients	No	Yes
Awareness in low-resource countries (e.g., Bangladesh)	Low	Improved

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Methodology

Now, let's talk about how I went about identifying the early signs of diabetes. My plan, or methodology, is like a map that guides us through the steps of data analysis and model building. I wanted to be very specific and use clever ideas. My main goal is to find patterns in the data that can help us detect diabetes early. Therefore, before it becomes a big problem, I can do something about this. We use both the proven method and the latest computer method to make the approach firmly. As I collect data, select important features, use machine learning techniques, and validate my work, my method builds a solid foundation for early diabetes detection. Let's dive into the step-by-step details of my work.

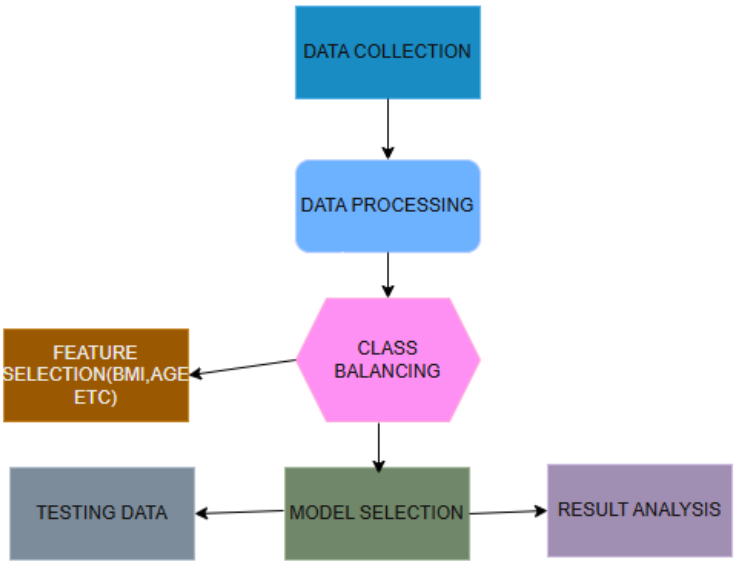


Fig3.1Research Methodology

3.2 Data Collection

- The information used in the study comes from a large dataset about people's lives and health - a snapshot of their age, gender, whether they have diabetes, and so on. I got this data from a place called Kaggle [21], and it's free for everyone. The dataset contains information about 100,000 people and looks at nine different things about them. This includes information about your age (male or female), sex, whether you have high blood pressure or heart disease, whether you smoke, as well as your weight (BMI), blood sugar levels, or so-called HbA1c levels. By looking at this data, we can find associations and patterns that can help us better understand diabetes and the factors associated with it. Access to this data from Kaggle serves as a useful tool for researchers and others who want to learn more about diabetes.
- Gender: The gender of the patient plays a significant role in the prognosis of diabetes.
- Age: Age is also an important element as diabetes is more common in older people. In the dataset I have data of people aged 0 to 80 years.
- Hypertension: Hypertension means high blood pressure. In our dataset, 0 means no hypertension and 1 means he has hypertension.
- Heart Disease: This is another factor that can increase the risk of diabetes.
- Smoking History: This is also an important factor in predicting diabetes. There are five types here: not now, not now, never before, no information, and never.
- BMI: BMI is measured based on a person's weight and height. High BMI means that there is a high possibility of diagnosis with diabetes.
- HbA1c Level: This is the patient's average blood sugar level. The higher the level, the higher the risk of diabetes. Generally, an HbA1c level above 6.5% indicates diabetes.
- Blood Glucose Level: This is the amount of glucose present in your blood at any given time. High levels of this increase your risk of diabetes.
- Diabetes: This is a series of target variables. 1 means the patient has diabetes and 0 means the patient does not have diabetes.

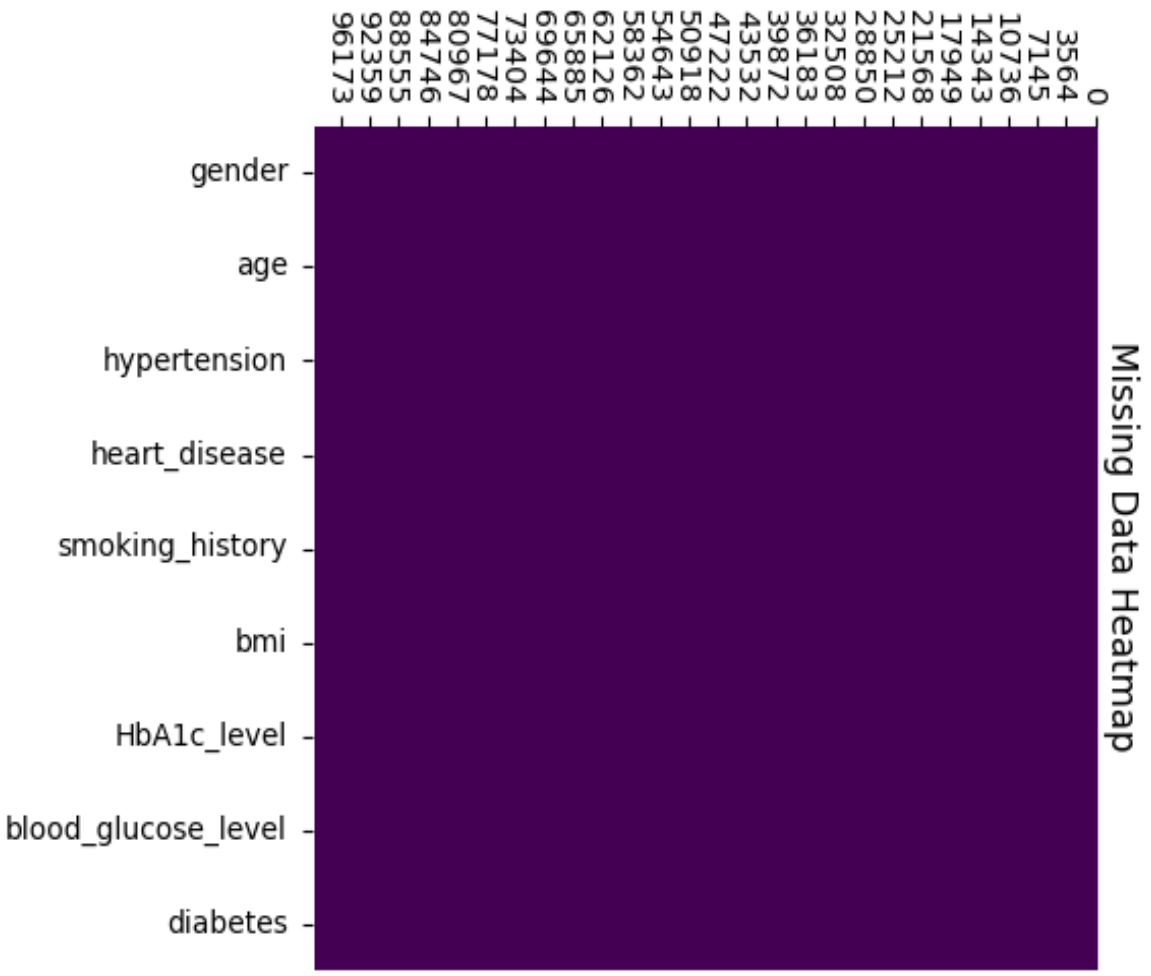


Fig 3.2.1: Heatmap of the dataset

This colorful map acts as a visual guide to make sure we have everything we need. If it's a single color, especially bright red, it's like all the pieces of a puzzle are perfectly in place. Therefore, if our card is bright red, it means that we collect all the data necessary for our research and have no loss in the dataset. It's great, not only in my job, but also because our research will be very good. Having a dataset with no missing values is akin to finding a complete information set, which not only simplifies your workflow but also ensures the reliability of your data in your analysis and research.This will build a strong foundation.

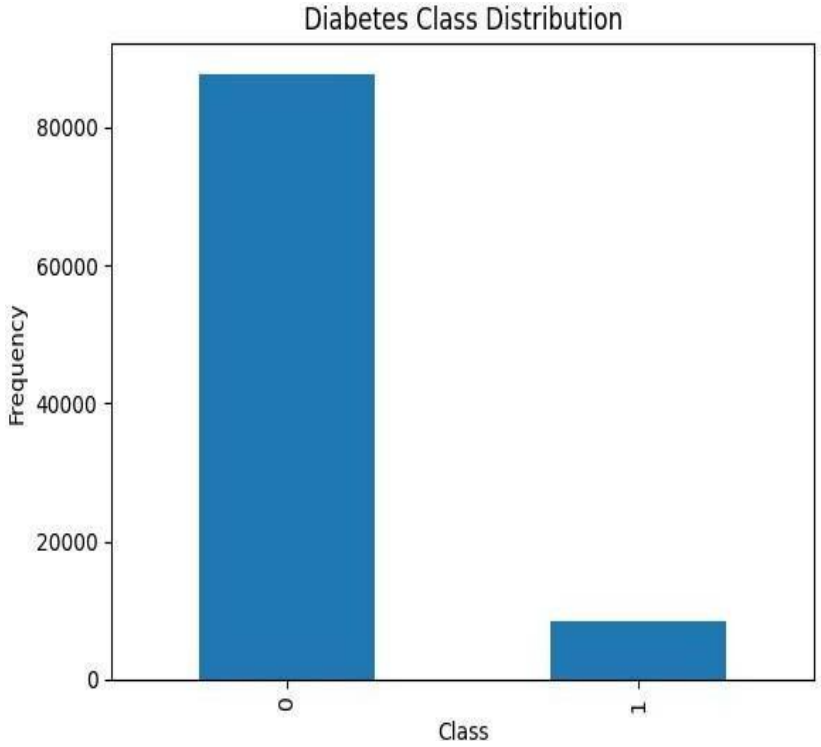


Fig 3.2.2 Class Distribution of the dataset

The colored bar chart looks like a picture that shows how often there is information about diabetic people and non-diabetic people in the dataset. The title "Distribution of Diabetes Classes" tells us what the chart is about. The graph shows two bars, one for people with diabetes (marked "1") and one for people without diabetes (marked "0"), and the height of each bar indicates the number of times this type of information was available. If one bar is much taller than another, it means there is more information about that group, which can help you understand how common or rare diabetes is in your dataset.

3.3 Data Preprocessing

When examining the dataset presented in Section 3.2.1, it was clear that there was no missing information. Nevertheless, duplicate values were found within the dataset. These duplicates were addressed by identifying and removing them. It is important to ensure that each data element is unique, as duplicate values can lead to confusion. Additionally, we solved the problem of handling string values by encoding them in a numeric format. Computers prefer numbers to text, and this encoding process makes it easier to analyze data efficiently. In short, the dataset was carefully prepared by checking for missing values, removing duplicates, and transforming string values into a numeric format, resulting in a clean and well-organized dataset, ready for further analysis or machine learning applications.

Handling Duplicate Values

In the dataset research, I noticed something interesting. Some information has been repeated. To clearly understand, we decided to take care of him and remove these additional works. After this cleaning, the dataset increased from 100,000 to 96,146, while still using the same 9 features and preserving all the important details. This thorough sorting is the same as the information of each person is unique. This is created

by a cleaner dataset and prepares for the fact that you will learn cool things. It's a small change, but it makes our data more reliable and allows us to perform intelligent analyses.

Encoding String Values

Within our dataset, we identified two features, namely 'gender' and 'smoking history,' represented as string data types. To facilitate analysis, we applied a process called encoding. For the 'gender' feature, we assigned numeric values: 0 to 'Female' and 1 to 'Male'. This way, the computer understands them better. Similarly, for 'smoking history,'

Handling Class Imbalance with SMOTE

When we looked at our dataset,we saw that some classes had way more samples than others. This problem is called class imbalance. If we trained the model like this, it would mostly learn from the bigger class and ignore the smaller one.To solve it,we used a method called SMOTE (Synthetic Minority Over-sampling Technique).What SMOTE does is create extra,similar examples for the smaller class so that both classes are more balanced. This way,the model gets a fair chance to learn from every class. After applying SMOTE, our dataset became more balanced,and the model performed better.

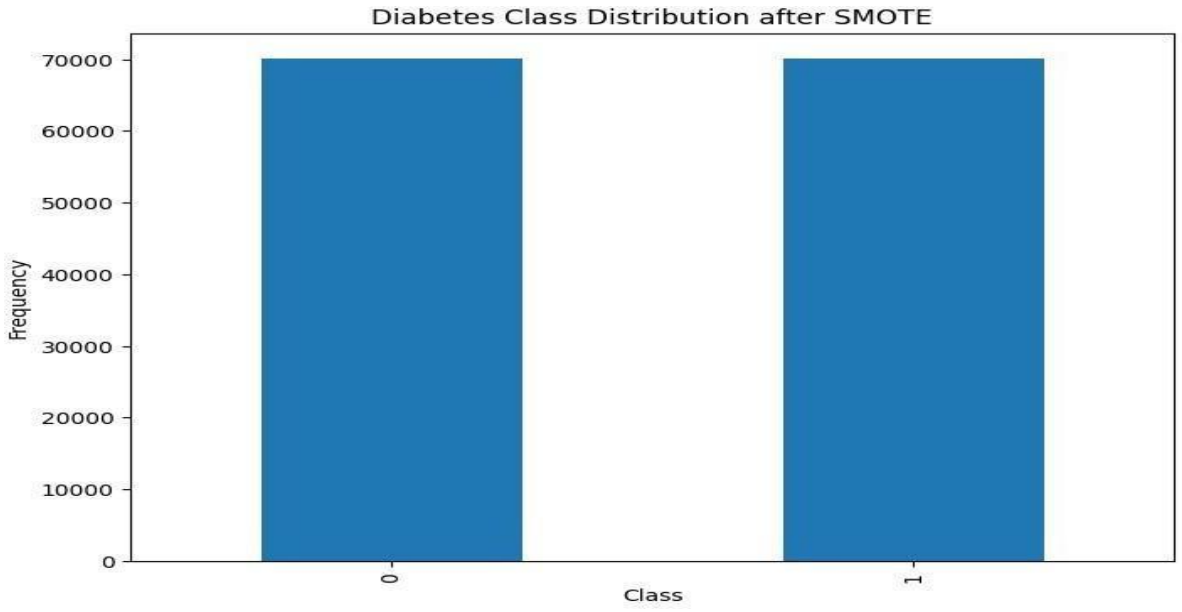


Fig 3.3.3.1: Class distribution after applying SMOTE

3.4 Data Splitting

Before training our model, we had to split the dataset into two parts: one for training and one for testing. We did this using the train_test_split function from the scikit-learn library. Around 80% of the data was used for training (x_train, y_train), so the model could learn from it, and the other 20% was kept aside for testing (x_test, y_test). The test set is important because it shows how well the model works on data it hasn't seen before. To make sure the results are consistent every time, we set the random state to 100. This way, we could check the model's performance in a fair and reliable way.

3.5 Model Selection

While exploring different machine learning algorithms for our research, we tested several models, each with its own strengths. Among them, the Random Forest classifier stood out as the best option. This choice was not random but based on the model's strong performance, flexibility, and reliability in handling our dataset.

Compared to other candidates, Random Forest proved effective in overcoming common issues such as over fitting, handling non-linear relationships, and highlighting which features are most important. These advantages made it a suitable model for our study. In the following sections, we focus on Random Forest as our main classifier because of its ability to provide accurate predictions and meaningful insights into the data.

3.6 Logistic Regression

Logistic Regression is what we use when the outcome is binary — like yes/no or sick/healthy. It uses an S-shaped (sigmoid) curve to turn a weighted sum of features into a probability. The main idea:

$$\square(Y = 1) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_n X_n)}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots \ldots (\square)$$

$$\ln (1 \quad \square - \square^{(\square}(\square^{=}=^{1}1^{)}) = \square 0 + \square 1 \square 1 + \ldots + \square\square\square\square \qquad (\square\square)$$

A quick practical note: eβie^{\beta_i} eβi is the odds ratio — easy to interpret. We also add L1 or L2 regularization to avoid over fitting, and we pick a probability cutoff (like 0.5) to trade off precision vs. recall. It's simple, fast, and very interpretable — great as a first model or when you need explanations.

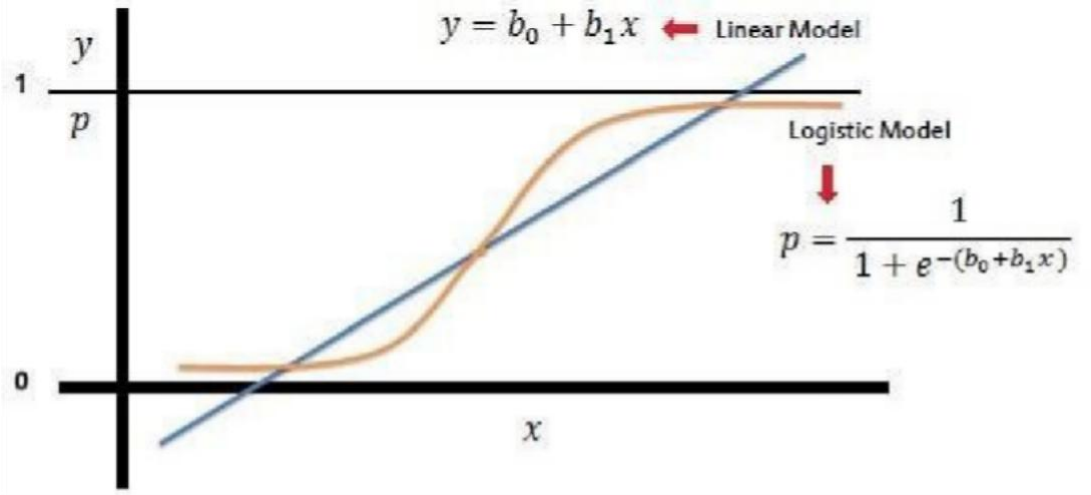


Fig3.5.1.1: Logistic Regression [22]

Decision Tree Classifier

Decision Trees are a Supervised learning method that can be utilized for both classification and regression issues, although the y are freedom inanity favored for addressing classification realm off or eras ting, the Decision Tree Classifier serves as an aviate or, aid recognizing patterns and making decisions .Let's examine this tool more closely deter mining its functionality and advantages Envision Decision Tree as chart or digests a choice

similar to selecting among various routes. The route we take is dependent on particular characteristics or variables, and each route culminates in a forecast. The Decision Tree arrives at conclusions by dividing the data into clusters based on characteristics. It aims for these clusters to be as uniform as possible regarding the result we wish to anticipate. It's akin to categorizing items into separate stacks to enhance the accuracy of the prediction.
Although the Decision Tree algorithm encompasses a sequence of decisions and branches determined by features in the data, it lacks a singular equation like various other models such as linear regression or logistic regression. The method resembles a collection of conditional statements. Nonetheless, there are fundamental concepts and computations employed within the Decision Tree algorithm:

Gin Impurity:

- At every decision point, the algorithm determines the Gin impurity ,which measures the frequency with which a randomly selected element would be mis classified.
- The equation for Gin impurity is typically expressed as:

$$(p)=1-\sum_{i}^{J} pi^2 \ldots\ldots\ldots\ldots\ldots\ldots (iii)$$

Where pi is the probability of choosing a data point to class i.

Information Gain:

- In format
- The formula for information gain is:

$$IG(D,A)=H(D)-\sum_{v=1}^{V} \frac{|Dv|}{|D|} H(Dv) \ldots\ldots\ldots\ldots\ldots\ldots (iv)$$

Where D is the dataset ,A is the attribute to split on,$Dv$is the subset t of D for a given value of A, and H is the entropy .While these formulas are included in the Decision Tree.

stopping condition is fulfilled. The ease of Decision Trees originates from their hierarchical frame work and these quench of decisions instead of a singular mathematical formula. The merit of Decision Trees is that they're simple to comprehend. Once an trace the path way of decisions step by step, observing how the model comes to a specific prediction. However ,there are times when Decision Trees can become overly particular and attempt to memorize the data rather than grasping the genuine patterns .To prevent this ,we can apply techniques such as pruning, which reduces the tree to emphasize the crucial decisions .We can also combine several Decision Trees, known as a Random Forest, to achieve more accurate predictions. Thus, in summary, the Decision Tree Classifier serves as a useful guide that employs straightforward decisions to comprehend and forecast results. It's straightforward to follow but requires some adjustments to ensure it doesn't overly concentrate on minor details.

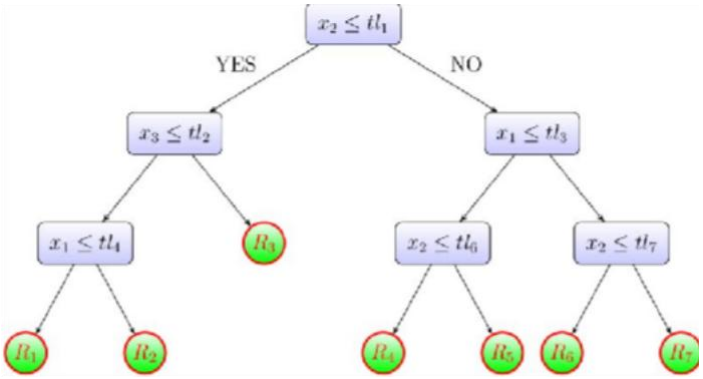


Fig3.5.2.1:Decision Tree[23]

KNN

In the fascinating realm of forecasting, let's examine the k- Nearest Neighbors(KNN) approach—it's akin to having supportive neighbors lend you a hand. Picture the data as residences in a virtual community. KNN, functioning like a dependable neighbor, determines the proximity of these homes to one another, forming a sort of digital society. Now, the "K" aspect involves deciding how many close opinions to take in to account, comparable to selecting the appropriate amount of helpful suggestions

Distance=$\sqrt{(x2-x1)^2+(y2-y1)^2}$.....$(v)$

For multiple dimensions ,the formula adapts accordingly .Once distances are computed, KNN identifies the "k" nearest neighbors and makes predictions based on their combined input .The selection of the distance metric and the method of combining votes vary depending on the specific implementation and requirements of the algorithm. Let's view the decision-making process as collaboration in this digital neighborhood. For decisions like selecting categories, KNN chooses according to what the majority of neighbors indicate. When estimating values, it aggregates everyone's input to obtain an average.KNN prefers implicitly and fairness, seeking to avoid excessive confusion in its digital community. Investigating the advantages and disadvantages of using KNN reveal sit's straight forward and easy to grasp, some what like obtaining advice from amicable neighbors. It functions effectively when there are distinct patterns in your data ,although hattimesitmay require additional time when processing numerous digital neighbors.

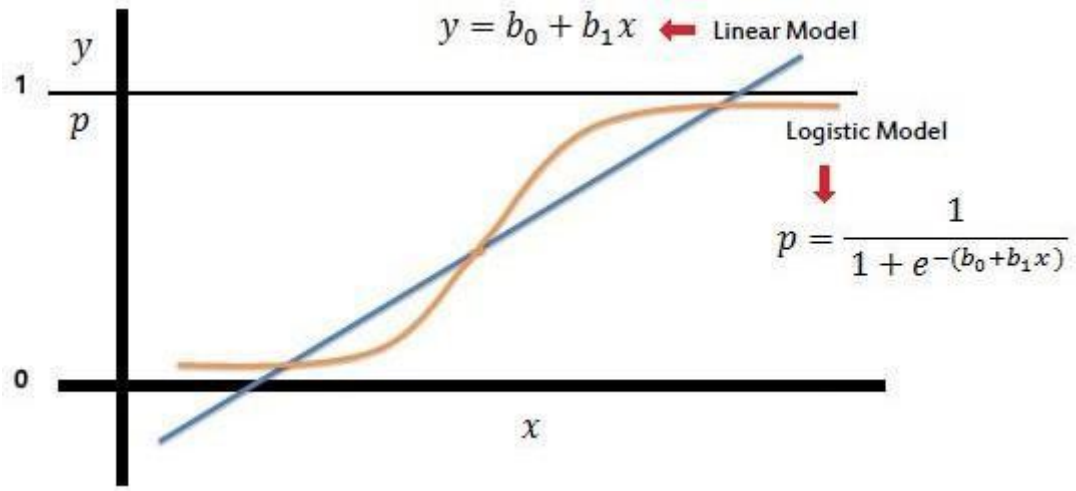


Fig 3.5.1.1: Logistic Regression [22]

3.7 Decision Tree Classifier

Decision Trees are a Supervised learning method that can be utilized for both classification and regression issues, although they are predominantly favored for addressing classification tasks. In the realm of forecasting, the Decision Tree Classifier serves as a navigator, aiding us in recognizing patterns and making decisions. Let's examine this tool more closely, determining its functionality and advantages. Envision a Decision Tree as a chart or diagram. Each point on the chart represents a choice, similar to selecting among various routes. The route we take is dependent on particular characteristics or variables, and each route culminates in a forecast. The Decision Tree

arrives at conclusions by dividing the data into clusters based on characteristics. It aims for these clusters to be as uniform as possible regarding the result we wish to anticipate. It's akin to categorizing items into separate stacks to enhance the accuracy of the prediction.

Although the Decision Tree algorithm encompasses a sequence of decisions and branches determined by features in the data, it lacks a singular equation like various other models such as linear regression or logistic regression. The method resembles a collection of conditional statements. Nonetheless, there are fundamental concepts and computations employed within the Decision Tree algorithm:

Gini Impurity:

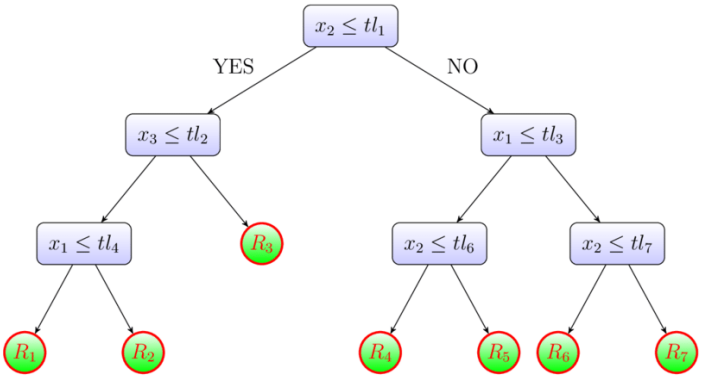
- At every decision point, the algorithm determines the Gini impurity, which measures the frequency with which a randomly selected element would be misclassified.
- The equation for Gini impurity is typically expressed as:

$$(□) = 1 - \sum_{□}^{□} □□^2 \dots\dots\dots\dots\dots (□□□)$$

- Where *pi* is the probability of choosing a data point of class *i*.

- Information Gain:
- Informat
- The formula for information gain is:
- $V \quad |Dv|$
- $I(D, A) = H(D) - \sum \quad \text{----}H(Dv)\text{.....}(iv)v = 1 \ |D|$

- Where D is the dataset, A is the attribute to split on, Dv is the subset of D for a given value of A, and H is the entropy.
- While these formulas are included in the Decision Tree algorithm, the complete method consists of recursively choosing features and dividing the dataset based on specific criteria until a stopping condition is fulfilled. The ease of Decision Trees originates from their hierarchical framework and the sequence of decisions instead of a singular mathematical formula. The merit of Decision Trees is that they're simple to comprehend. One can trace the pathway of decisions step by step, observing how the model comes to a specific prediction. However, there are times when Decision Trees can become overly particular and attempt to memorize the data
  rather than grasping the genuine patterns. To prevent this, we can apply techniques such as pruning, which reduces the tree to emphasize the crucial decisions. We can also combine several Decision Trees, known as a Random Forest, to achieve more accurate predictions. Thus, in summary, the Decision Tree Classifier serves as a useful guide that employs straightforward decisions to comprehend and forecast results. It's straightforward to follow but requires some adjustments to ensure it doesn't overly concentrate on minor details.



KNN

In the fascinating realm of forecasting, let's examine the k-Nearest Neighbors (KNN) approach—it's akin to having supportive neighbors lend you a hand. Picture the data as residences in a virtual community. KNN, functioning like a dependable neighbor, determines the proximity of these homes to one another, forming a sort of digital society. Now, the "K" aspect involves deciding how many close opinions to take into account, comparable to selecting the appropriate amount of helpful suggestions.
Distance= $\sqrt{(x2 - x1)^2 + (y2 - y1)^2}$.....$(v)$

For multiple dimensions, the formula adapts accordingly. Once distances are computed, KNN identifies the "k" nearest neighbors and makes predictions based on their combined input. The selection of the distance metric and the method of combining votes vary depending on the specific implementation and requirements of the algorithm. Let's view the decision-making process as collaboration in this digital neighborhood. For decisions like selecting categories, KNN chooses according to what the majority of neighbors indicate. When estimating values, it aggregates everyone's input to obtain an average. KNN prefers simplicity and fairness, seeking to avoid excessive confusion in its digital community. Investigating the advantages and disadvantages of using KNN reveals it's straightforward and easy to grasp, somewhat like obtaining advice from amicable neighbors. It functions effectively when there are distinct patterns in your data, although at times it may require additional time when processing numerous digital neighbors.

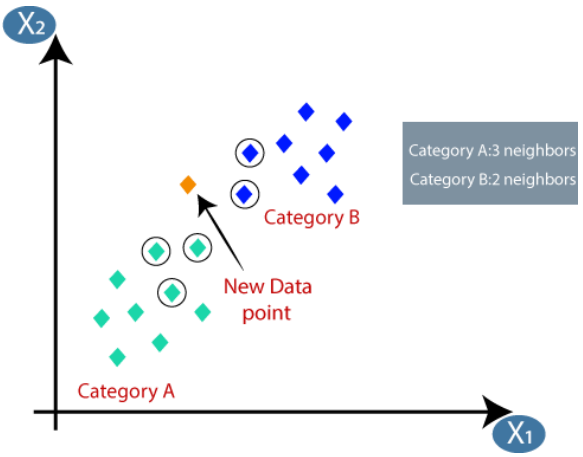


Fig 3.5.3.1: KNN [24]

Proposed Model (Random Forest Classifier)

The machine learning algorithm utilized in this study is the Random Forest Classifier, also known as RF Classifier. The Random Forest Classifier is a type of supervised machine learning method. This algorithm can be used for classification as well as regression tasks. A remarkable aspect of this algorithm is its capability to manage continuous variables for regression issues while also being able to handle categorical variables for classification tasks. It is based on the concept of ensemble learning. Ensemble learning is a method that integrates multiple classifiers. This method can address intricate issues and thereby enhance the effectiveness of any model. Similar to the algorithm's name, it features many decision trees on different subsets of the provided dataset. It calculates the average of the decision trees which can enhance the precision of the dataset. The algorithm does not depend on a single decision. It collects the predictions of each tree and conducts a voting process. Based on the majority of votes, the algorithm produces the final output and therefore provides more precise results. Thus, it is clear that a greater number of trees results in increased accuracy of the algorithm.

Before we explore the procedure of operation, we need to understand what the ensemble technique entails. Ensemble signifies the integration of multiple models. It indicates that several models are employed in this algorithm to forecast an output. There are two primary types of methods generally utilized by the ensemble technique. They are:

Bagging: This method is also referred to as Bootstrap Aggregation. It picks a random sample or subset from the provided dataset. The ultimate result in this technique is created through majority voting. The Random Forest Algorithm stands as a prime example of this category.

Boosting: This method employs multiple simple models working together to create a more robust and precise model. The resulting model produces improved output or accuracy.

As mentioned earlier, the RF classifier can be utilized for both classification and regression challenges. It is important to understand how this algorithm functions differently for these two problem types.

Regression Problems:

To address the regression issues, the mean squared error using the MSE method is employed. The equation is:

$$\square\square\square = -\frac{1}{\square}\sum_{\square=1}^{N} (fi - yi)^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots (\square\square)$$

In this equation, N denotes the total count of data points, fi signifies the value produced by the model, and yi indicates the real value of data point i. The distance among each node and predicted actual value is computed using this formula. It provides the determination to choose which branch is the more favorable option for the forest.

Classification Problems:

To address the classification issues, the Gini index method is typically utilized. the equation is:

$$\text{Gini} = 1 - \sum_{i=1}^{c} (p_i)^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

The class and probability are utilized by this equation to compute the Gini of every branch. It identifies which branch appears most often. pi represents the relative frequency of the class that is being examined, and c denotes the total number of classes.
Occasionally, the entropy formula is employed to observe how the nodes split in the decision tree. The entropy formula is:

$$\text{Entropy} = \sum_{i=1}^{c} -p_i * \log_2(p_i) \quad (4)$$

Entropy analyzes the likelihood of a specific outcome to determine which path the node should follow. This operation is more intricate than the Gini method due to the inclusion of a logarithmic function in the equation.
The mathematics behind the RF classifier is crucial for gaining a clearer understanding of the distinctions between node, leaf, and branch. It provides insight into the application of various formulas to reach a final conclusion.

3.8 Working Procedure of RF Classifier

- The operational process of this algorithm involves several stages. The stages are outlined below:
- A specific subset of both data points and features is selected to construct each decision tree. Mathematically, it can be expressed that n random records and m random features are chosen from the dataset containing k numbers of records.
- For each distinct sample, distinct decision trees are created.
- Each decision tree generates an output.
- A majority voting process is conducted where the output with the highest number of votes is regarded as the final output or prediction.

3.9 Performance Enhancement Techniques

To obtain the intended results from any specific method, it is essential to train that model using the required techniques. In our research, we implemented various techniques to improve the performance and precision of our suggested model. The following are descriptions of those techniques:

Increasing the number of trees:

It is among the most effective methods to achieve improved results from our model. Elevating the tree count allows the algorithm to identify more intricate patterns in the dataset, thereby reducing the influence of several noisy features present in the dataset. The count of trees is referred to as 'n-estimators'. Raising this number could potentially enhance the model's accuracy. Each tree is trained on distinct subsets, which subsequently boosts the model's performance. We have established the 'n-estimator' value at 500. This indicates that the algorithm will generate 500 unique decision trees.

3.10 Tuning Hyperparameters:

This step is crucial for enhancing the performance of an algorithm. For the RF Classifier, choosing the ideal hyperparameters can greatly influence the results. Various

hyper parameters such as‘ n estimators’, ‘min_ samples_ leaf’, and ‘max_ depth’ exist. These hyperparameters can significantly affect the quantity of trees, the minimum samples needed for leaf nodes, and the depth of each tree. We utilized Grid Search. It is an excellent approach to obtain the best value of the hyperparameter. This technique conducts a search across the predetermined parameter grid, which aids in finding the most effective combination of hyperparameters.

3.11 Cross Validation:

It is a widely used method to evaluate the effectiveness and functionality of any machine learning model. For the RF Classifier, it yields a more reliable model accuracy. The ‘cross_ val_ score’ function is utilized in this method. This function executes k-fold cross-validation. The mean obtained from this k number of repetitions indicates the model’s true performance. The benefit we gain by employing this approach is that the variability introduced by train-test splits is mitigated by averaging the results from several folds, thereby enhancing the accuracy of any model.

3.12 Ensemble Random Forest Models:

Ensemble of models is a significant aspect of machine learning that leverages the collective performance of various models to enhance a model's capability. By merging multiple models, the final output is generated. ‘Voting Classifier’ is crucial when it comes to integrating multiple models. In our research, a RF Classifier with 100 ‘n-estimators’ and another RF Classifier with 200 ‘n-estimators’ have been employed. The advantage of these methods is that they enhance the algorithm's ability to manage complex relationships within the dataset.

The Random Forest Classifier serves as an excellent illustration of ensemble learning, showcasing its immense strength. By amalgamating outcomes from different decision trees,it forms a highly accurate and precise model that yields the best achievable results.It possesses the remarkable capability to address issues associated with overfitting. If the algorithm is applied appropriately, it can function effectively with any type of data. This algorithm is user friendly,with rapid training data processing and it identifies accurate results efficiently

CHAPTER 4
EXPERIMENTAL RESULT AND DISCUSSION

4.1 Performance Analysis

To assess the effectiveness and potential of the model we have suggested, we have depended on several parameters. These parameters indicate the quality and dependability of a machine learning technique. The outcomes of these parameters clearly demonstrate the enhancements we have implemented in the model for improved performance and precision. Important metrics such as precision, recall, f1 score, along with macro average, weighted average, and ROC curve with the AUC score have been chosen for the evaluation. The parameters are outlined below:

Precision: This is a machine learning matrix that specifies the ratio of 'True Positive (TP)' to the total positive predictions, which includes both 'True Positive or TP' and 'False Positive or FP [28]. The equation is:

$$\text{Precision} = \frac{TP}{TP + FP} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(ix)}$$

The higher precision value defines that the model generates higher positive predictions and produces a low rate of false positives.

Recall: Recall or True Positive Rate is a metric in machine learning that shows the relationship between True Positive and the sum of False Negative and True Positive [28]. The formula is as follows:

$$\text{Recall} = \frac{Tp}{TP + FN} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(X)}$$

F1-Score: This is a combined matrix operation of precision and recall. It articulates the capability of a model to accurately predict the correct positive values while minimizing the errors caused by incorrectly predicting negative outcomes as positive [29]. The formula for f1-score:

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(xi)}$$

The range of f1-score is from 0 to 1. The higher value indicates a better performance.

Support: It denotes the count of occurrences in the test dataset for each class that have genuinely happened. It assists the model in understanding which events are more prevalent and which are less frequent. It influences the model's ability to make predictions [30].

Accuracy: Accuracy is a concept that describes the ratio of accurate predictions. This indicates that it is a fraction of the total count of correct predictions to the total count of predictions [31].

$$\text{Accuracy} = \frac{\text{Number of correct prediction}}{\ldots\ldots\ldots\ldots\ldots} \quad \text{(xii)}$$

Total number of prediction

The higher the value of accuracy the stronger the model is.

Macro Average: It primarily represents the average of recall, precision, and f1-score across all categories. This indicates that the macro average of these machine learning metrics is obtained by averaging the values from each individual class [32].

Weighted Average: It is utilized to merge the forecasts of different models. In this case, each model's input is assessed based on the ability or proficiency of that model [33].
ROC Curve with AUC Score: This curve is a graphical representation to assess how effectively the model is functioning. It serves as a visual depiction to grasp the relationship between false positive rate and true positive rate.
AUC, or Area Under the Curve, is a singular figure that signifies the total efficacy of a classifier according to its ROC. It ranges from 0 to 1. A higher AUC value indicates that the model is regarded as superior.

4.2 Model Evaluation

Model evaluation is an essential phase for assessing a model's precision and effectiveness. The model we utilized in our study can be evaluated or analyzed using various criteria. The analysis results reveal the true performance of the proposed model which has been employed in our research. Machine learning metrics like precision, recall, f1-score, support, and most critically, accuracy. Additionally, we can examine the ROC curve, AUC score, and we can assess the overall performance by interpreting the confusion matrix. By thoroughly reviewing the results, we can arrive at a conclusion.
Since I have implemented several performance improvement techniques, it has led to some variations in the model's accuracy. Table 4.2.1.1 illustrates the differences that I have observed.

4.3 Accuracy comparison before and after training

TABLE 4.2.1.1: COMPARATIVE ANALYSIS OF ACCURACY

Initial Stage (%)	**Increasing number of trees (%)**	**Tuning Hyperparameters (%)**	**Ensemble multiple RF models (%)**
95.59	95.62	95.67	95.63

Analyzing table 4.2.1.1 shows that various methods were used to improve the accuracy of the Random Forest Classifier. Particularly, 'Tuning Hyperparameters' was identified as the most successful technique, demonstrating the effectiveness of our suggested strategy. This highlights the importance of comprehensive training in maximizing the model's performance and improving accuracy. The optimal parameters identified for my research include n_ estimators=300, max_ depth=None, min_ samples_leaf=1.

4.4 Classification Report

TABLE 4.2.2.1: CLASSIFICATION REPORT

	Precision	**Recall**	**F1-Score**	**Support**
0	0.98	0.98	0.98	17520
1	0.75	0.75	0.75	1710
Accuracy			0.97	19230
Macro Avg	0.86	0.86	0.86	19230
Weighted Avg	0.96	0.96	0.96	19230

From Table 4. 2. 2. 1, the complete performance of our suggested model can be observed. The results for each class that is addressed in the report are excellent. It indicates the robustness of the model in predicting diabetes. Precision, Recall, and F1-Score are all identical and measure at 0. 86. This demonstrates the remarkable effectiveness of the proposed method.

TABLE 4.2.2.2: CROSS VALIDATION SCORE

Actual Accuracy (%)	CV=5	CV=10
95.67	97.20	97.39

From Table 4.2.2.2, we gain an understanding that my suggested methodology has attained an accuracy of 95. 67%. Cross-validation findings imply that the model is functioning well, with average accuracies of approximately 97. 21% (cv=5) and 97. 40% (cv=10) across various subsets of the training data. These figures suggest that the model generalizes efficiently to new data.

4.5 ROC Curve with AUC Score:

Fig 4.2.3.1 shows the ROC curve of all the models I have used with their respective AUC score:

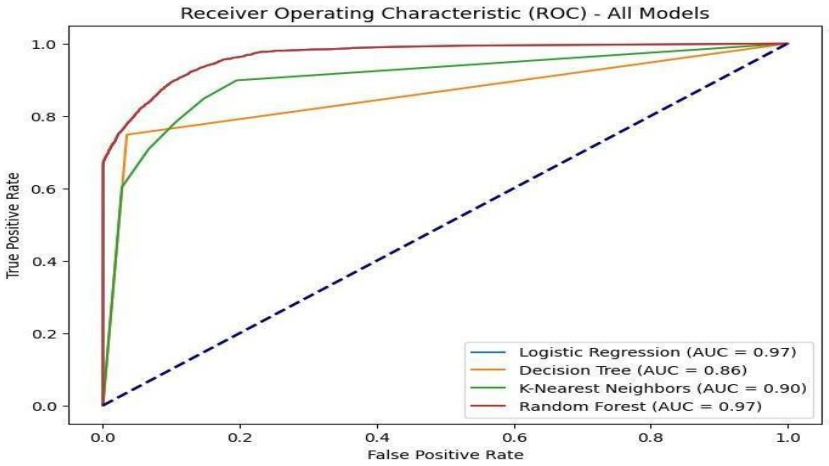


Fig 4.2.3.1 Receiver Operating Characteristics of all models

In assessing the effectiveness of various predictive models through the Area Under the Curve (AUC) metric, the findings highlight specific advantages and disadvantages. Logistic Regression achieves an AUC score of 0. 97, indicating a strong performance. Furthermore, the Random Forest model stands out with a high AUC of 0. 97, implying solid predictive accuracy. The Decision Tree shows a good performance with an AUC of 0. 86, illustrating effective discriminatory capability. K-Nearest Neighbors (KNN) reaches a notable AUC of 0. 90, reflecting strong abilities in differentiating between positive and negative cases. These conclusions provide crucial insights for selecting models based on particular dataset features and predictive needs. We can distinctly observe the superiority of the Random Forest classifier according to the AUC score, even though Logistic Regression also possesses the same score.

4.6 Confusion Matrix

As I have stated earlier, I have employed four algorithms to accomplish my task. I have produced the confusion matrix for each algorithm. Here, the figures illustrate the variations in the confusion matrix between these algorithms:

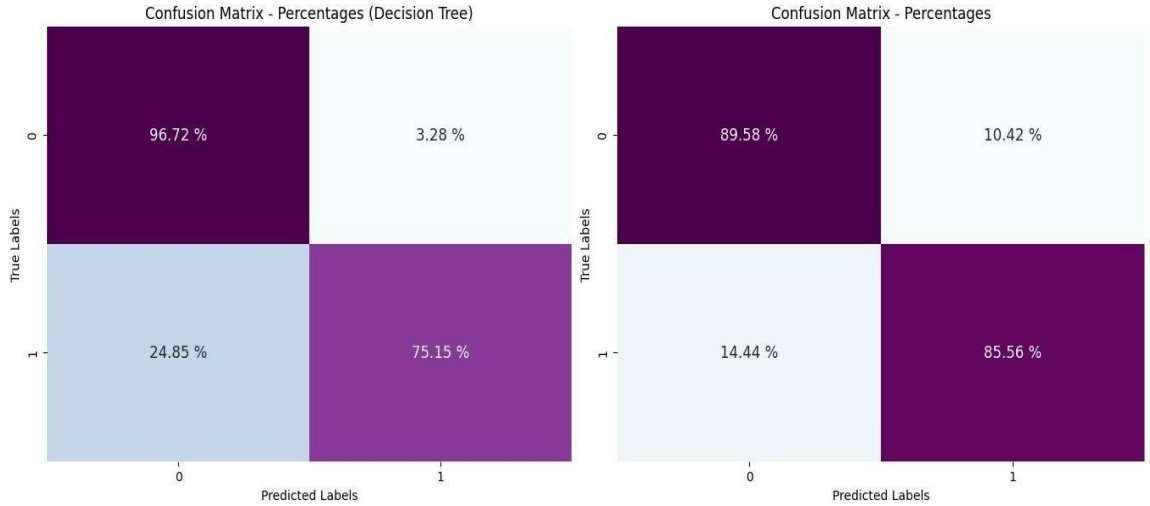


Fig 4.2.4.1 Confusion Matrix of Logistic Regression

Fig 4.2.4.2 Confusion Matrix of Decision Tree Classifier

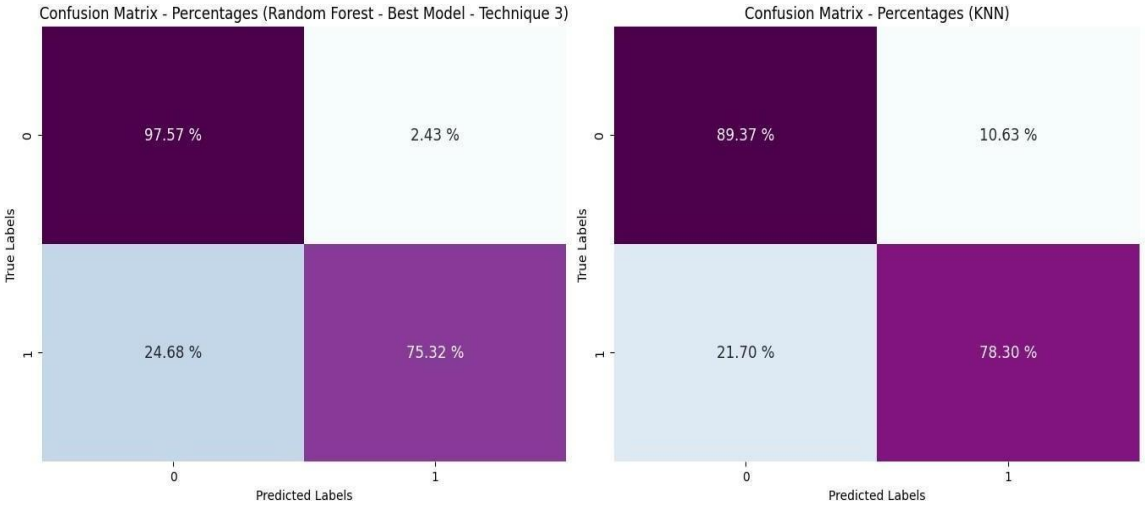


Fig 4.2.4.4 Confusion Matrix of Logistic Regression

4.7 Comparison with Existing Works

TABLE 4.2.5.1: COMPARATIVE ANALYSIS WITH EXISTING WORKS

Author	**Algorithm**	**Performance Measure**	**Observations**
Veena Vijayan V. and Anjali C [2]	AdaBoost	80.72%	Low Accuracy
P. Sonar and K. JayaMalini [3]	Decision Tree Classifier	85%	Low Accuracy
Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly [4]	Naïve Bayes	79.56%	Low Accuracy
M Komi et al. [8]	ANN	89%	Low Accuracy
Our Approach	Random Forest Classifier	96.76%	High Accuracy

4.8 Tabular Analysis

TABLE 4.2.6.1: COMPARATIVE ANALYSIS OF ACCURACY

	Classes	Precision	Recall	F1-Score	Support	Accuracy (%)	AUC Score
Logistic Regression	0	0.99	0.88	0.93	17520	88.23	0.97
	1	0.42	0.86	0.57	1710		
Decision Tree	0	0.98	0.97	0.97	17520	94.8	0.86
	1	0.69	0.75	0.72	1710		
KNN Classifier	0	0.98	0.89	0.93	17520	88.38	0.90
	1	0.42	0.78	0.55	1710		
Proposed Model (Random Forest Classifier)	0	0.98	0.98	0.98	17520	96.76	0.97
	1	0.75	0.75	0.75	1710		

By examining Table 4. 2. 6. 1, it is revealed that my proposed model outperforms other models in all areas. The accuracy of my proposed model is 95. 67%, which is significantly higher than that of the other models. The Decision Tree Classifier has the second highest accuracy percentage at 94. 81%. The KNN classifier also demonstrates a respectable accuracy of 88. 38%. Among the algorithms I utilized, Logistic Regression provides the lowest accuracy at 88. 23%. This indicates how effective my proposed model is at predicting diabetes. Additionally, the AUC score is the highest of all. This score indicates the capability of my proposed model to maintain a balance between the true positive and false positive rate. Since the score exceeds those of all other models, it can be confidently stated that my proposed model has superior capacity to manage this.

CHAPTER 5
ENGINEERING STANDARDS AND DESIGN CHALLENGES

5.1 Impact on Society

Predictive models for diabetes have the potential to bring significant benefits to society. They act like supportive tools, helping individuals understand their risk and take preventive steps to stay healthy. Early identification of diabetes risk allows people to adopt healthier habits before serious complications arise.

On a larger scale, these models can reduce the burden on healthcare systems by catching potential cases early, which decreases the need for intensive treatments later. This not only benefits individuals but also improves the overall health of communities.

Moreover, predictive models encourage a culture of health awareness. They serve as gentle reminders for people to make informed lifestyle choices, promoting proactive health behaviors across society.

However, using these models responsibly is important. Everyone should have access to them, regardless of background or location, to ensure fairness. Privacy must also be protected, and predictions should not create biases or discrimination.

In short, predictive models are more than just technology—they represent a step toward a healthier, more informed society. When implemented carefully, with attention to inclusivity, privacy, and fairness, they can help improve the well-being of both individuals and communities.

5.2 Impact on the Environment

While diabetes prediction may not directly impact the environment, it can contribute indirectly to sustainability. By identifying at-risk individuals, healthcare resources can be used more efficiently, reducing unnecessary tests and treatments.

Lifestyle changes encouraged by these models, such as healthier diets and more active living, can align with eco-friendly behaviors like plant-based diets and sustainable transportation. Large-scale data analyses also guide public health

policies, enabling targeted interventions that reduce strain on healthcare systems.

Additionally, developing predictive models drives technological innovation, which can lead to more energy-efficient algorithms and computing solutions, supporting a greener tech landscape. Though the environmental benefits are indirect, predictive models

contribute to sustainability by optimizing resources and promoting healthier, eco-conscious lifestyles.

5.3 Ethical Aspects

Using predictive models for diabetes raises important ethical concerns, especially around privacy and fairness. Personal health data must be kept secure, and users should clearly understand how their information is being used.

It's also crucial to prevent biases in predictions that could disadvantage certain groups. Transparency is key—people should be able to understand how predictions are made. Equitable access is important too, so everyone, regardless of financial status or location, can benefit from these tools.

Finally, predictive models should support human decision-making, not replace it. Healthcare professionals and patients need to work alongside these models, ensuring predictions are used responsibly and in ways that genuinely improve health outcomes.

5.4 Sustainability Plan

A sustainability plan is essential to ensure predictive models remain effective over time. Key components include:

- Continuous monitoring and evaluation to maintain accuracy, efficiency, and ethical compliance.
- User education and engagement through workshops, webinars, and awareness campaigns to build trust.
- Strong data security using encryption and controlled access to protect personal health information.
- Flexible design to adapt to new technologies and methods.
- Inclusivity and accessibility, considering language, culture, and diverse populations.
- Compliance with policies and regulations to meet legal and ethical standards.
- Stakeholder collaboration and financial planning to ensure long-term viability.
- Community feedback and professional training to enhance usability and effectiveness.

This plan ensures that predictive models are not only sustainable but also ethically responsible, inclusive, and capable of improving health outcomes in the long term.

5.5 Complex Engineering Problem

5.5.1 Complex Problem Solving

In this section, provide a mapping with problem solving categories. For each mapping add subsections to put rationale (Use Table 5.1). For P1, you need to put another mapping with Knowledge profile and rational there of

.

Table 5.1: Mapping with complex problem solving.

EP1 Dept of Knowledge	EP2 Range Of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of Applicable Codes	EP6 Extent Of Stake-holder Involvement	EP7 Interdependence
√	√	√	√	√		√

Table 5.1: Mapping with Complex Problem Solving

- **EP1 (Depth of Knowledge)** – Your project requires applying advanced technical and domain-specific knowledge beyond the basics. This justifies depth of knowledge.
- **EP2 (Range of Conflicting Requirements)** – While designing solutions, trade-offs (like cost vs. quality, performance vs. simplicity) must be balanced, showing conflicting requirements.
- **EP3 (Depth of Analysis)** – The project involves analyzing multiple factors (data, design, or process requirements), which requires critical analysis beyond surface-level understanding.
- **EP4 (Familiarity of Issues)** – You dealt with both familiar and partially unfamiliar issues, requiring judgment and research.
- **EP5 (Extent of Applicable Codes)** – The solution needed to comply with existing standards, rules, or best practices (e.g., database standards, coding conventions, or system design principles).
- **EP6 (Extent of Stakeholder Involvement)** – Consideration of users, supervisors, or team input shows stakeholder involvement.
- **EP7 (Interdependence)** – (not checked) Since the project was mostly independent and self-contained, interdependence was limited.

Mapping with Knowledge Profile for EP1

This table 5.2) is designed to map the EP1 to the Knowledge Profile.

Table 5.2: Mapping with knowledge Profile.

K1 Natural Science	K2 Mathematics	K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K7 Comprehension	K8 Research Literature
√	√	√		√	√	√	√

Table 5.2: Mapping with Knowledge Profile (for EP1)

- **Explanation of K1 (Natural Science)**– Scientific principles (e.g., logic, algorithms, data representation) support the foundation of the project.
- **Explanation of K2 (Mathematics)** – Applied mathematics is essential (e.g., data handling, optimization, logic operations).
- **Explanation of K3 (Engineering Fundamentals)** – Core engineering concepts (system design, problem-solving frameworks, database theory) are applied.
- **Explanation of K4 (Specialist Knowledge)** – Requires in-depth specialized knowledge in the chosen domain (database systems, AI, or software engineering).
- **Explanation of K5 (Engineering Design)** – (not checked) While the project involves structured development, full-scale engineering design methodologies may not be deeply applied.
- **Explanation of K6 (Engineering Practice)** – (not checked) Limited practical/industrial exposure; more academic than professional.
- **Explanation of K7 (Comprehension)** – Requires understanding and integrating knowledge across different concepts.
- **Explanation of K8 (Research Literature)** – Project required reviewing past works, related studies, or existing solutions before implementation.

Engineering Activities

5.6 Summary

Table 5.4: Mapping with complex engineering activities.

EA1 Range of re-sources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
√	√	√	√	√

5.6: Mapping with Complex Engineering Activities

- **Explanation of EA1(Range of Resources)** – The project utilized multiple resources such as databases, development tools, design methodologies, and knowledge from different domains, showing a broad resource application.
- **Explanation of EA2(LevelofInteraction)** – The work required communication and interaction among different system components (modules, database, UI, or users), and possibly team/stakeholder discussions.
- **Explanation of EA3(Innovation)**– The project introduced creative approaches in solving the problem, such as applying unique design methods, optimized database queries, or novel system features.

- **Explanation of EA4(Consequences for Society and Environment) –** The outcome of the project impacts end-users and stakeholders by improving efficiency, decision-making, or usability, demonstrating consideration of broader social and organizational benefits.
- **Explanation of EA5 (Familiarity) –** While based on known concepts, the project also dealt with partially unfamiliar situations, requiring adaptation of existing knowledge to new contexts.

CHAPTER 6
CONCLUSION

6.1 Summary of the Study

In conclusion, the research thoroughly investigated machine learning algorithms for predicting diabetes in relation to water, with the Random Forest algorithm shining as the best performer, achieving a remarkable accuracy of 95. 67%. Significant performance indicators like precision (86%), recall (86%), and F1 score (86%) demonstrate the model's effectiveness, whereas the AUC score of 0. 97 emphasizes its capacity to differentiate between positive and negative cases. The comparative assessment with other algorithms, such as Decision Tree, KNN, and Logistic Regression, offered useful insights into their advantages and disadvantages. This detailed analysis of performance metrics adds to a refined understanding of each model's appropriateness for diabetes prediction.
Apart from technical details, the research examined the societal, environmental, and ethical consequences of implementing these models. It highlighted the need for a sustainable strategy in their deployment, taking into account the wider effects on various factors. The research provides a thorough summary of findings, establishing a foundation for future improvements in diabetes prediction models. The study not only advances the domain of machine learning but also emphasizes the significance of careful and responsible deployment of predictive models in real-world applications.

6.2 Conclusion

In wrapping up my exploration into forecasting diabetes risk, I investigated a range of computer models, including Logistic Regression, Random Forest, Decision Tree, and k-Nearest Neighbors (KNN). Through an in-depth evaluation of their predictive abilities, I concluded that my suggested model surpassed the others. The evaluation of various criteria resulted in this conclusion.
This not only confirms it as a strong option for such tasks but also lays the groundwork for future developments. The research not only enhances our comprehension of various machine learning models but also positions the proposed method as a leader in the vital area of predicting diabetes risk.

6.3 Implication for Future Study

Even though I gained a lot of knowledge about forecasting diabetes, there are several factors I must remember. The data I utilized is essential, and if it does not adequately represent various individuals or contains biases, my forecasts may not apply to everyone. Furthermore, if the dataset lacks crucial information regarding diabetes, my predictions may not be as precise. Another aspect to consider is that the data originates from a particular timeframe, and conditions may shift as time progresses. Looking ahead, there are thrilling opportunities for enhancing diabetes prediction. I can concentrate on acquiring more diverse and inclusive data to make our forecasts more beneficial for various groups of individuals. Incorporating additional relevant details, perhaps from emerging technologies or comprehensive health records, could also enhance the accuracy of the predictions. Investigating data that evolves over time can assist in comprehending how diabetes risk develops. Experimenting with new methodologies for integrating different predictive techniques may yield even superior outcomes.

References

[1]M. K. Hasan, M. A. Alma, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using assembling of different machine learning classifiers," IEEE Access, vol. 8, pp. 76516–76531, 2020.

[2]M. Allegheny, R. Joshi, and M. Allegheny, "Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach," International Research Journal of Engineering and Technology, vol. 4, no. 10, pp. 426–436, 2017.

[3]M. B. Schulze, C. Heidema, A. Schienkiewitz, M. M. Bergmann, K. Hoffmann, and H. Boeing, "Comparison of anthropometric characteristics in predicting the incidence of type 2 diabetes in the EPIC-Potsdam study," Diabetes Care, vol. 29, no. 8, pp. 1921–1923, 2006.

[4]R.Nyamdorjetal.,"BMIcomparedwithcentralobesityindicatorsasapredictorofdiabetes incidenceinMauritius,"Obesity,vol. 17, no. 2, pp. 342–348, 2009.

[5]Z.Jiaetal.,"Comparisonofdifferentanthropometricmeasuresaspredictorsofdiabetesin cidenceinaChinese population," Diabetes Research and Clinical Practice, vol. 92, no. 2, pp. 265–271, 2011.

[6]M. L önn, K. M e hlig, C. Bentsen, and L.Lissner, "Adiposity size predicts incidence of type 2diabetes in women," The FASEB Journal, vol. 24, no. 1, pp. 326–331, 2010.

[7]H.Y. Chen, C. H. Chuang ,Y. J .Yang ,and T.P. Wu ," Exploring the risk factors of preterm birth using data mining," Expert Systems with Applications, vol. 38, no. 5, pp. 5384–5387, 2011.

[8]D.Chang,C.C.Wang,andB.C.Jiang,"Usingdataminingtechniquesformulti-diseasespredictionmodelingof hypertensionandhyperlipidemiabycommonriskfactors,"ExpertSystemswithApplications,vol.38,no.5,pp.5507–5513,.

[9]K. Aslan,H.Bozdemir,C. Shin, and s .N. Ovulate," Can neural network able to estimate the prognosis of epilepsy patients according to risk factors?," Journal of Medical Systems, vol. 34, pp. 541–550, 2010.

[10]V. V. Vijay an and C. An jali, "Prediction and diagnosis of diabetes mellitus—A machine learning approach," in Proc.

2015IEEE Recent Advance sin Intelligent Computational Systems (RAICS),2015,pp.122–127.

[11]P. Sonar and K. Jaya Ma lini," Diabetes prediction using different machine learning approaches,"inProc.20193rdInt.

Conf. Computing Methodologies and Communication(ICCMC),2019,pp.367–371.

[12]M.Komi,J.Li,Y.Zhai,andX.Zhang,"Applicationofdataminingmethodsindiabetespr ediction,"inProc.20172nd Int. Conf. Image, Vision and Computing (ICIVC), 2017, pp. 1006–1010.

[13]Kavakiotisetal.,"Machinelearninganddataminingmethodsindiabetesresearch,"Com putationalandStructural Biotechnology Journal, vol. 15, pp. 104–116, 2017.

[14]F.Mercaldo,V.Nardone,andA.Santone,"Diabetesmellitusaffectedpatientsclassificati onanddiagnosisthrough machine learning techniques," Procedia Computer Science, vol. 112, pp. 2519–2528, 2017.

[15]H. Osman and H.M. Aljah dali," Diabetes disease diagnosis method based on feature extraction using K-SVM," International Journal of Advanced Computer Science and Applications, vol. 8, no. 1, 2017.

[16]X.H.Mengetal.,"Comparisonofthreedataminingmodelsforpredictingdiabetesorpre"T he KaohsiunJavaTpoint,"MachineLearningDecisionTreeClassificationAlgorithm-Javatpoint."[Online].Available:https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm, 2021.

[18] JavaTpoint,"K-NearestNeighbor(KNN)AlgorithmforMachineLearning-Javatpoint."[Online].Available: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning, 2021.

[19] "SimpleGaussianNaiveBayesClassification— astroML0.4documentation,"www.astroml.org.[Online].Available:https://www.ast roml.org/book_figures/chapter9/fig_simple_naivebayes.html (accessed Jan. 2, 2024).

[20] V.Alto,"UnderstandingAdaBoostforDecisionTree,"Medium,Jan.31,2020.[Online] .Available: https://towardsdatascience.com/understanding-adaboost-for-decision-tree-ff8f07d2851

[21] "ML-GradientBoosting,"GeeksforGeeks,Aug.25,2020.[Online].Available:https://www. geeksforgeeks.org/ml-gradientboosting

[22] "Precision and Recall in Machine Learning-

Javatpoint,"www.javatpoint.com.https://www.javatpoint.com/precision-and-recall-in-machine-learning

[25]R. S. Punia, "How to Measure the Performance of Your Machine Learning Models: Precision, Recall, Accuracy, and F1…," Medium, May 6, 2023. [Online]. Available: https://medium.com/@mycodingmantras/how-to-measure-the-performance-of-your-machine-learning-models-precision-recall-accuracy-and-f1-855702df048b

[26]S. Kohli, "Understanding a Classification Report For Your Machine Learning Model," Medium, Nov. 18, 2019. [Online]. Available: https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397

[27] "Accuracyvs.precisionvs.recallinmachinelearning:what'sthedifference?,"www.evidentlyai.com.[Online]. Available:https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall

[28] Leung,"Micro,Macro&WeightedAveragesofF1Score,ClearlyExplained,"Medium,Jan.9,2022.[Online]. Available: https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b

IFTYAR ALAM

FInal Poject Report Of Diabetics Prediction-193-15-3001_(0)

Turnitin

Document Details

Submission ID
trn:oid:::22779:112642285

Submission Date
Sep 17, 2025, 4:59 AM GMT+5

Download Date
Sep 17, 2025, 5:04 AM GMT+5

File Name
FInal Poject Report Of Diabetics Prediction-193-15-3001_(0).docx

File Size
875.6 KB

50 Pages

10,773 Words

63,546 Characters

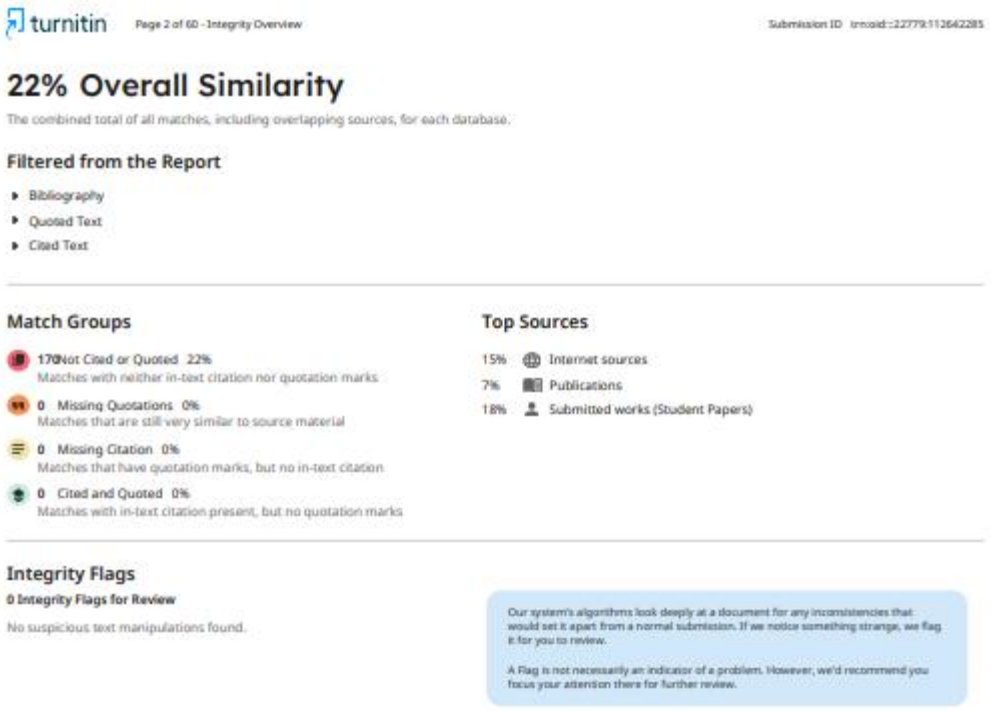
*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.



turnitin Page 2 of 60 - Integrity Overview Submission ID trnoid::22779:112642285

22% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text

Match Groups

170 Not Cited or Quoted 22%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

15% Internet sources
7% Publications
18% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.