Machine Learning for Diabetes Forecasting: A Way Forward for Early Intervention

BY

MD. Abir Hossain
ID: 193-15-3025

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

Supervised By

Raja Tariqul Hasan Tusher
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Md. Assaduzzaman
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JULY 2024

APPROVAL

This Project titled **"Machine Learning for Diabetes Forecasting: A Way Forward for Early Intervention"**, submitted by MD. Abir Hossain, ID No: 193-15-3025 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 13th July 2024.

BOARD OF EXAMINERS

Dr. S.M Aminul Haque (SMAH) — **Chairman**
Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

13.07.24

Ms. Nazmun Nessa Moon (NNM) — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

13.07.24

Mr. Deawan Rakin Ahamed Remal (DRAR) — **Internal Examiner**
Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

13.07.2024

Dr. Abu Sayed Md. Mostafizur Rahaman (ASMR) — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

I hereby declare that, this project has been done by us under the supervision of **Raja Tariqul Hasan Tusher (THT) (Assistant Professor), Department of CSE** 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:

Raja Tariqul Hasan Tusher
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Md Assaduzzaman
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Submitted by:

Abir Hossain
ID: 193-15-3025
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to almighty Allah for His divine blessing makes me possible to complete the final year project/internship successfully.

I am really grateful and wish our profound our indebtedness to **Raja Tariqul Hasan Tusher (THT), Assistant Professor,** Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

I would like to express our heartiest gratitude **Raja Tariqul Hasan Tusher** and **Md. Assaduzzaman** Department of CSE, for their kind help to finish my project and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

The medical industry and machine learning now have a stronger connection because to technological advancements. This work uses machine learning to forecast the prevalence of diabetes, a worldwide illness. Predicting the disease at its early stages is the goal in order to make treatment or management of the illness easier. I have worked with a dataset that has nine characteristics and 100,000 occurrences. This dataset includes data on blood glucose level, BMI, age, smoking history, heart disease, gender, and HbA1c level in addition to hypertension. These are the principal markers of diabetes. I have used the Random Forest Classifier to forecast the sickness. My study's findings have been contrasted with those of other machine learning techniques, such as Decision Tree Classifier and Logistic Regression. After comparison, it was discovered that, out of all of these methods, the Random Forest Classifier had the greatest accuracy and AUC score. Using this approach, I have determined that the accuracy is 95.67%, with an AUC score of 0.97. KNN has 88% accuracy, decision trees have 94%, and logistic regression has 88%. The findings demonstrate our study's remarkable success in accurately predicting diabetes. My research indicates that there is great potential for integrating computer science and medicine so that dangerous conditions like diabetes may be identified early on.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v
CHAPTER 1 INTRODUCTION	**1-3**
1.1 Introduction	1
1.2 Motivation	2
1.3 Objective	3
1.4 Research Questions	3
1.5 Expected Outcome	3
CHAPTER 2 BACKGROUND	**4-7**
2.1 Literature review	4
2.2 Overview of Existing Works	7
2.3 Challenges	7
CHAPTER 3 RESEARCH METHODOLOGY	**9-23**
3.1 Methodology	9
3.2 Data Collection	10
3.3 Data Preprocessing	12

3.4 Data Splitting 14

3.5 Model Selection 14

CHAPTER 4 EXPERIMENTATL RESULT AND DISCUSSION **24-30**

4.1 Performance Analysis 24

4.2 Model Evaluation 25

CHAPTER 5 IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY **31-33**

5.1 Impact on society 31

5.2 Impact on Environment 32

5.3 Ethical Aspects 32

5.4 Sustainability Plan 33

CHAPTER 6 SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH **34-35**

6.1 Summary of the Study 34

6.2 Conclusion 34

6.3 Future Research 35

REFERENCES **36-37**

LIST OF FIGURES

FIGURES	PAGE NO
Fig 3.1 Research Methodology	9
Fig 3.2.1: Heatmap of the dataset	11
Fig 3.2.2: Class Distribution of the dataset	12
Fig 3.3.3.1: Class distribution after applying SMOTE	14
Fig 3.5.1.1: Logistic Regression [22]	16
Fig 3.5.2.1: Decision Tree [23]	18
Fig 3.5.3.1: KNN [24]	19
Fig 4.2.3.1 Receiver Operating Characteristics of all models	27
Fig 4.2.4.1 Confusion Matrix of Logistic Regression	28
Fig 4.2.4.2 Confusion Matrix of Decision Tree Classifier	28
Fig 4.2.4.3 Confusion Matrix of KNN	29
Fig 4.2.4.4 Confusion Matrix of Random Forest Classifier	29

LIST OF TABLES

TABLES	PAGE NO
Table 2.1: Existing Works	8
Table 4.2.1.1: Comparative Analysis of Accuracy	26
Table 4.2.2.1: Classification Report	26
Table 4.2.2.2: Cross Validation Score	28
Table 4.2.5.1: Comparative Analysis with Existing Works	29
Table 4.2.6.1: Comparative Analysis of Accuracy	30

CHAPTER 1
INTRODUCTION

1.1 Introduction

Diabetes has become a worldwide health problem. Millions of lives have been affected, and it is now a major concern for the healthcare system. A person is generally diagnosed with diabetes if their blood sugar levels are higher than 4.4–6.1 mmol/L [1]. The pancreas produces the hormone insulin. It facilitates the bloodstream's absorption of glucose from diet. Diabetes may result from a lack of this hormone if there is any dysfunction. It can also result in malignancy, retinal failure, cerebral vascular dysfunction, cardiovascular malfunction, and other conditions. According to reports, there are at least 500 million diabetic patients worldwide; by 2030, that number might rise to 25 percent and 51 percent, respectively.422 million persons worldwide have diabetes and there were 1.6 million fatalities as per WHO on the world diabetes day on 2016 [3]. There are primarily three types of diabetes. Gestational, Type 1, and Type 2 [4]. When it comes to Type 1, the condition often begins as an inflammatory disease before the age of 20. This kind of diabetes is characterized by damage to the pancreatic cells that produce insulin. Insulin resistance develops in the body's organs when type 2 diabetes is present, increasing the demand for insulin. Right now, the pancreas isn't making enough insulin. Pregnant women are more prone to develop gestational diabetes due to inadequate pancreatic insulin production. Treatment is necessary for any of these different types of diabetes [5]. On the other hand, early precise detection of this condition may be vital to improving the patient's circumstances and enabling successful disease management. Machine learning is one of the many subfields of artificial intelligence. It has demonstrated significant influence and promise across a range of industries, including healthcare. I'm delving deeply into machine learning's predictive capabilities for diabetes in this project. Our objectives are to assist the patients and enhance the early diagnosis of this illness. The convergence of computer technology and medical information brings up new opportunities in the healthcare industry. We can delve deeper into large, complex data sets with the aid of machine learning. By combining these two in the ideal way, I can look forward to a brighter future when diabetes can be identified early. Prior research indicates that adipocyte size and anthropometric measures can be used to predict the risk of getting diabetes using standard statistical methods. [6] [7]

[8] [9]. Studies show that data mining can be a vital factor to know the unknowns and potentialities of the models that have been developed for medical fields [10] [11] [12]. The current state of diabetes prediction using techniques related to machine learning is examined in this research. In order to fully understand the most recent techniques, algorithms, and features used in comparable investigations, I have reviewed the relevant research. I have also described our approach of gathering data, preparing it, and creating models. The findings of this study could have consequences for policymakers, researchers, and medical professionals working to improve diabetes management and prevention efforts.

1.2 Motivation

Urgency of Innovation:

- The increasing prevalence of diabetes necessitates urgent development of innovative methods for early identification and treatment.
- Traditional diagnostic methods may not fully capture the complex interactions among diabetes risk factors, underscoring the need for advanced models.

Specific Focus on Bangladesh:

- In Bangladesh, where awareness about diabetes consequences is low, there is a heightened risk due to lack of understanding among the population.
- There is a critical need to introduce effective early-stage diagnostic methods tailored to local challenges and risk factors.

Role of Machine Learning:

- Machine learning offers revolutionary potential in improving diabetes prediction.
- The research aims to address the complexity of diabetes risk factors (biomarkers, medical history, demographics) and provide personalized treatment options.

Utilization of Machine Learning Techniques:

- Techniques such as Random Forests and Logistic Regression are employed to develop accurate predictive models.
- The objective is to offer trustworthy predictions that facilitate early treatments and significantly improve patient outcomes.

Impact on Healthcare Delivery:

- By leveraging the accuracy of machine learning algorithms, the study aims to advance diabetes prediction and revolutionize healthcare delivery.
- This approach promises to be proactive, individualized, and grounded in modern machine learning techniques, paving the way for a transformative change in diabetes management

1.3 Objective

- **Machine Learning Focus** : The study aims to leverage machine learning techniques for precise and early prediction of diabetes.

- **Development of Robust Models** : By exploring various algorithms and analyzing diverse patient datasets, the objective is to develop robust predictive models.

- **Timely Insights for Healthcare** : These models are intended to provide healthcare professionals with timely insights, facilitating preemptive interventions in diabetes management.

- **Enhanced Diagnosis and Treatment** : Ultimately, the goal is to improve the accuracy of diabetes diagnosis, leading to more individualized and efficient treatment plans in response to the widespread prevalence and significance of diabetes.

1.4 Research Questions

I had a lot of questions when I began this research project. There weren't many challenges, but there were some unanswered questions. Like as:

- Which machine learning method works best for predicting diabetes among Logistic Regression, Random Forest, and k-Nearest Neighbors?
- What specific information, like age, medical history, or other factors, is most influential in accurately predicting the likelihood of diabetes?
- How can the results from machine learning models be made easy to understand and useful for doctors to help patients prevent diabetes?
-

1.5 Expected Outcome

- To detect diabetes more efficiently.
- Select algorithm for this dataset.
- Preprocessing the dataset to find out the best possible outcome.
- Ablation study on the best performance model.
- Performance analysis and statistical analysis based on the confusion matrix.

CHAPTER 2

BACKGROUND

2.1 Literature Review

The title of my article is Diabetes Forecasting with Machine Learning: A Path to Early Intervention. The goal is to detect the disease at an early stage based on some circumstances and it may lead to prevent the disease by taking proper medication. Additionally, I have collected some papers on this topic which reflects the work on this same topic. The summary is added below:

Research done by Veena Vijayan V. and Anjali C [13] reveals that Diabetes results from elevated sugar levels in the blood. To predict and diagnose diabetes, various computerized information systems have been developed using different classifiers. The choice of the right classifier significantly enhances the accuracy and efficiency of the system. In this context, a decision support system is proposed, employing the AdaBoost algorithm with Decision Stump as the base classifier for classification. Additionally, Support Vector Machine, Naive Bayes, and Decision Tree are also utilized as base classifiers for the AdaBoost algorithm to verify accuracy. The AdaBoost algorithm with Decision Stump as the base classifier achieved an accuracy of 80.72%, surpassing 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. JayaMalini [14] where it is said that Diabetes, a potentially life-threatening condition, can give rise to various complications such as heart failure, vision impairment, and kidney diseases. Managing diabetes often involves frequent visits to diagnostic centers for consultations and reports, demanding both time and financial investment from patients. However, the advent of Machine Learning brings a promising solution. By harnessing advanced systems and data processing techniques, they have found the ability to predict whether an individual is at risk of developing diabetes. This predictive capability is crucial for early intervention, enabling timely and preventive measures. This research focuses on developing a sophisticated system that can accurately assess a patient's risk level for diabetes. Different classification methods, including Decision Tree, Artificial Neural

Network (ANN), Naive Bayes, and Support Vector Machine (SVM) algorithms, were employed for model development. The Decision Tree exhibited a precision of 85%, Naive Bayes showed 77%, and Support Vector Machine demonstrated 77.3% accuracy. These findings underscore the significant effectiveness of these methods in predicting the risk of diabetes with notable precision.

Another article of Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly [5] on diagnosis of diabetes using classification mining techniques reflects the lethal consequences of this disease. In their study, they have employed Decision Tree and Naïve Bayes algorithms to predict and diagnose diabetes. Through different techniques like validation and percentage split, they achieved varying levels of accuracy. Specifically, they obtained 74.86% accuracy using Decision Tree, 76.956% using the validation technique, and 79.56% using the Naïve Bayes algorithm. These findings highlight the effectiveness of these algorithms in accurately predicting diabetes, providing valuable insights into the potential application of these methods in healthcare settings.

By exploring data mining algorithms like GMM, SVM, Logistic Regression, ELM and ANN, M Komi et al. [15] have found that the application of data science methods extends beyond its traditional domains, offering valuable insights in various scientific fields. One such application involves making predictions based on medical data, particularly in the context of diabetes mellitus—a condition characterized by elevated blood glucose levels. While conventional methods rely on physical and chemical tests for diabetes diagnosis, data mining techniques have emerged as effective tools for predicting the risk of high blood pressure associated with diabetes. The experimental results demonstrate that ANN exhibits the highest accuracy among these techniques which is 89%. This underscores the potential of data mining, particularly through ANN, in providing accurate and early predictions for diabetes, thereby contributing to more effective healthcare practices.

Another paper published by Ioannis Kavakiotis et al. [16] has tried to make a sync between health and biotechnology. It is mentioned that the significant progress in biotechnology and health sciences has resulted in 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 methods in biosciences has become crucial in intelligently transforming this wealth of information into valuable

knowledge. Diabetes mellitus (DM), a group of metabolic disorders with a global impact on human health, has been a focal point of extensive research across various aspects. This study conducts a systematic review of the applications of machine learning and data mining techniques in diabetes research, covering Prediction and Diagnosis, Diabetic Complications, Genetic Background and Environment, and Health Care and Management. The most prominent category appears to be Prediction and Diagnosis. A diverse range of machine learning algorithms were employed, with 85% characterized by supervised learning approaches and 15% by unsupervised ones, particularly association rules. Among these algorithms, Support Vector Machines (SVM) emerged as the most successful and widely used. Clinical datasets were predominantly utilized for analysis. The applications discussed in the selected articles emphasize the value of extracting knowledge, leading to new hypotheses and deeper understanding, paving the way for further investigation in the realm of diabetes mellitus.

Mercaldo, F., Nardone, V., & Santone, A. they have proposed that feature selection is a great choice for increasing accuracy. They have used multi-layer perceptron, BayeNet, Random Forest, Hoeffding Tree, decision tree and Jrip for their research. They have indicated that Hoeffding Tree is the best fit for their research [17].

Another research led by Osman A.H. et al. [18] has combined K-means clustering and SVM to predict the disease. 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 research focuses on enhancing diagnostic accuracy and reducing misclassifications by identifying significant diabetes-related features. The selection of these features is crucial for the effectiveness of classifiers developed through knowledge discovery methods, addressing classification challenges in diabetes patient data. The study proposes an integrated approach, combining the SVM (Support Vector Machine) technique with K-means clustering algorithms for diagnosing diabetes. Experimental results indicate a high accuracy in distinguishing patterns between Diabetic and Non-diabetic patients, outperforming modern diagnosis methods in terms of performance measures. The Ttest statistical method demonstrated significant improvement results, particularly when applied to the UCI Pima Indian standard dataset using the KSVM (Kernelized SVM) technique. This suggests a promising avenue for more accurate and reliable diabetes diagnosis, contributing to advancements in healthcare practices.

Nilashi M, Ahmadi H, bin Ibrahim, & Shahmoradi, L. [19] have used Regression Tree and Classification to generate fuzzy rule. They have found the CART with noise removal can offer a better and effective result in order to diagnose diabetes at an early stage.

Huang Y X, Zhang Q, Meng X H, Rao D P, & Liu Q [20] have used various data mining techniques to predict diabetes. They have used real world data sets. They have used weka tools and SPSS to analyze data and predict respectively. They have used three methods and they are ANN, Logistic Regression and j48. And they have concluded that j48 gives the best result for their research.

2.2 Overview of Existing Works

TABLE 2.1: EXISTING WORKS

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

Numerous works have been done throughout all over the world on this particular topic to learn and explore more ways to diagnose the disease. Various criteria have been kept under considerations to perform the researches. 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 have not 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.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Methodology

Now, let's talk about how I went about figuring out the early signs of diabetes. My plan, or methodology, is like a map guiding us through the steps of looking at data and creating models. I wanted to be really accurate and use smart ideas. My main goal is to find patterns in the data that can help us spot diabetes early. This way, I can do something about it before it becomes a big problem. I am using both tried-and-true methods and the latest computer techniques to make my approach solid. As I go through collecting data, picking important features, using machine learning methods, and checking my work, my method is building a strong base for finding diabetes early. Let’s dive into the step-by-step details of my work.

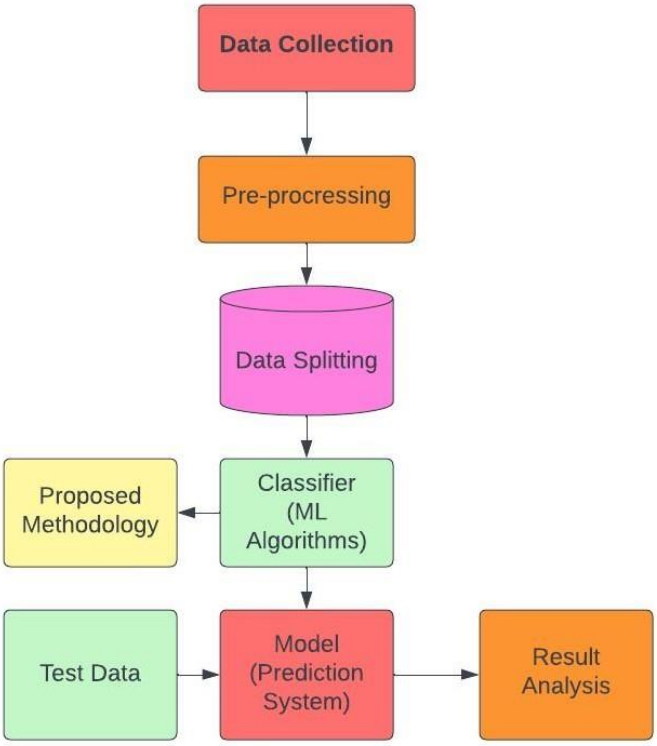

Fig 3.1 Research Methodology

3.2 Data Collection

The information I used in this study comes from a big collection of details about people's lives and health. It's like a snapshot that tells us things like age, gender, and if someone has diabetes. I got this data from a place called Kaggle [21], and it's free for anyone to use. The dataset has information about 100,000 people and looks at nine different things about them. These include how old they are, if they're male or female, whether they have high blood pressure or heart disease, if they smoke, and details about their weight (BMI), blood sugar levels, and something called HbA1c level. By looking at these details, I can find connections and patterns that help me understand more about diabetes and the factors linked to it. Having access to this data from Kaggle is like having a helpful tool for researchers and others interested in learning more about diabetes.

a. Gender: Gender of the patient has a huge impact on predicting diabetes.
b. Age: Age is also an important element because diabetes causes more to the older persons. In the dataset, I have the data of people from age 0 to 80.
c. Hypertension: Hypertension means high blood pressure. In our dataset 0 represents no hypertension and 1 represents that he or she has hypertension.
d. Heart Disease: This is another element that can increase the chance of being affected with diabetes.
e. Smoking History: It is also a vital factor in predicting diabetes. There are 5 types here, they are not current, current, former, no info, never and ever.
f. BMI: BMI is measured using the weight and height of a person. Higher BMI means a high chance of being diagnosed with diabetes.
g. HbA1c Level: It is the average blood sugar level of the patient. The higher the level, the higher the risk of diabetes. Normally, more than 6.5% of HbA1c level means diabetes.
h. Blood Glucose Level: It means the amount of glucose in the bloodstream at a given time. If the level is high, it increases the chance of diabetes.
i. Diabetes: It is the target variable in the dataset. 1 means the patient has diabetes and 0 means the patient doesn’t have.

Fig 3.2.1 shows the overall heatmap of the dataset:

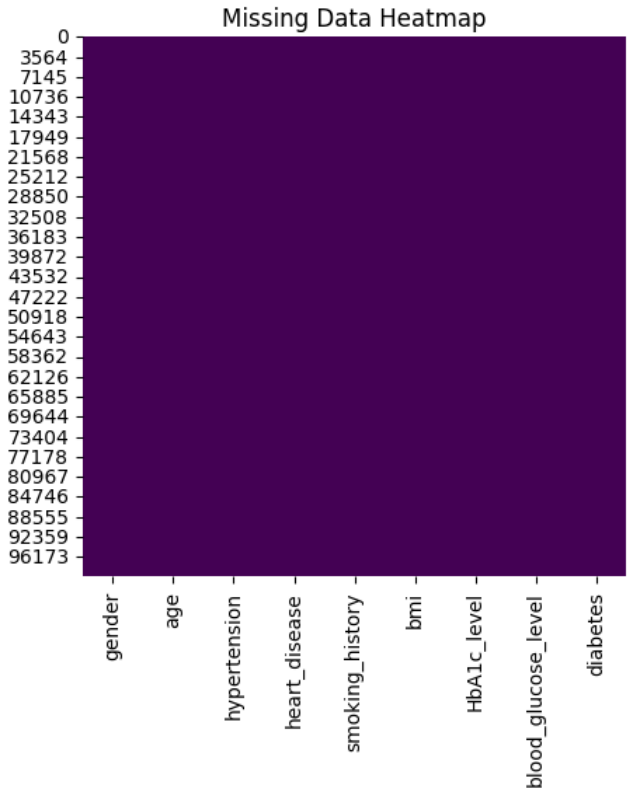

Fig 3.2.1: Heatmap of the dataset

This colorful map acts like a visual guide to make sure we've got everything we need. If it's all one color, especially bright red, it's like having all the pieces of a puzzle perfectly in place. So, when our map shows that vibrant red, it means I've gathered all the data necessary for our study, and there are no missing values in the dataset. This is fantastic because it not only makes my work easier but also boosts the chances of our study performing really well. Having a dataset without any missing values is like discovering a complete set of information. It not only simplifies the work process but also ensures that the data is dependable for our analysis and exploration. It sets a strong foundation for the next steps in our research, making the journey smoother and more reliable.

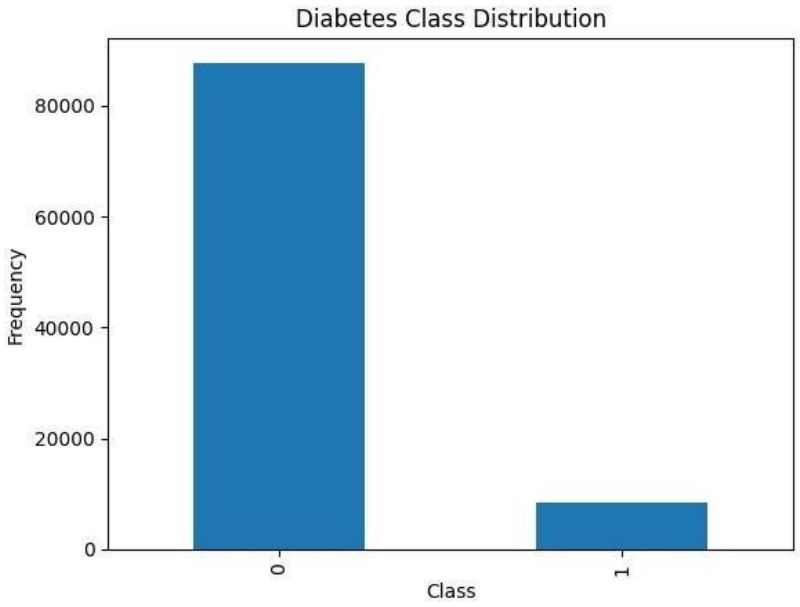

Fig 3.2.2: Class Distribution of the dataset

The colorful bar chart is like a picture that shows how many times there are information about people having diabetes and people not having diabetes in the dataset. The title 'Diabetes Class Distribution' tells us what the chart is about. On the chart, we see two bars—one for people with diabetes (marked as '1') and another for people without diabetes (marked as '0'). The height of each bar shows how many times we have that kind of information. If one bar is much taller than the other, it means we have more information about that group of people. This helps us understand how common or rare diabetes is in our dataset.

3.3 Data Preprocessing

In examining the dataset illustrated in section 3.2.1, it was evident that no information was missing. Despite this, I encountered duplicate values within the dataset, which I addressed by identifying and removing these repetitions. Duplicate values can lead to confusion, so ensuring each piece of data is unique is crucial. Additionally, I tackled the challenge of handling string values by encoding them into a numerical format. Computers prefer numbers over text, and this encoding process facilitates effective data analysis. In summary, the dataset was thoroughly prepared by checking for missing values, eliminating duplicates, and transforming string values into a numerical format, resulting in a clean and well-organized dataset ready for further analysis or machine learning applications.

3.3.1 Handling Duplicate Values

In our dataset study, we noticed something interesting – some info was repeated. To keep things clear, we decided to look closely and get rid of those extra bits. After this clean-up, our dataset went from 100,000 to 96,146 cases. We kept all the important details – still working with the same 9 features. This careful sorting is like making sure each person's info is unique. It makes our dataset cleaner and ready for us to learn cool things from it. It's a little tweak, but it makes our data more reliable and sets us up for some smart analysis.

3.3.2 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,' we used numbers 0 to 4 to represent different values – 'not current,' 'former,' 'no info,' 'never,' and 'ever.' This encoding ensures that these features are in a format the computer can easily work with, enhancing the efficiency of our analysis.

3.3.3 Handling Class Imbalance with SMOTE

In the dataset, I noticed that the number of cases for different classes might not be equal – we call this an imbalance. To fix this, I used a smart trick called SMOTE, which stands for Synthetic Minority Over-sampling Technique. SMOTE helps us to create a fairer playground by making more copies of the minority class. It's like giving everyone an equal chance to be heard. This ensures our analysis doesn't get biased because of more instances of one class compared to another. So, with the magic of SMOTE, I made sure the dataset represents everyone well, making it a better place for our computer to learn about the different outcomes. It removes the class imbalance which helps to train the model well and helps to get a better performance and prediction. Fig 3.3.3.1 describes how the dataset has performed after applying SMOTE:

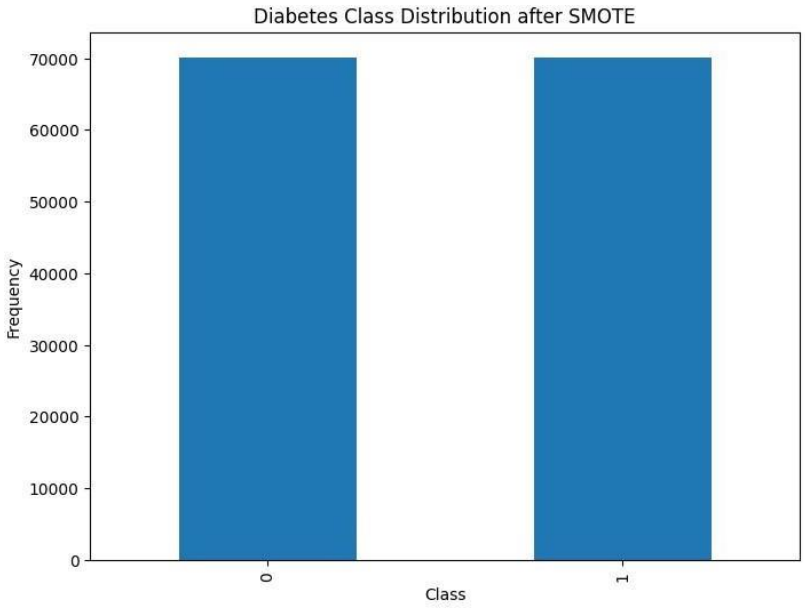

Fig 3.3.3.1: Class distribution after applying SMOTE

3.4 Data Splitting

To facilitate the training and evaluation of our machine learning models, a critical step involved the systematic splitting of our dataset. The dataset was partitioned into two subsets: a training set (denoted as `x_train` and `y_train`) and a testing set (denoted as `x_test` and `y_test`). This division was executed through the `train_test_split` function from the scikit-learn library. Approximately 80% of the data was allocated to the training set, allowing the models to learn patterns and relationships. The remaining 20% constituted the testing set, serving as an independent set to assess the models' generalization performance. A fixed random state (set to 100) was employed to ensure reproducibility in subsequent analyses. This meticulous separation of data sets enables a robust evaluation of the models, gauging their ability to apply learned insights to new, unseen data—a crucial aspect in ensuring the reliability of our predictive models.

3.5 Model Selection

As we delved into the realm of machine learning algorithms for our research, we encountered a diverse array of computational tools, each with its own unique set of capabilities. Among this selection process, a standout emerged, earning the coveted position of our preferred model—the Random Forest Classifier. This choice wasn't arbitrary; rather, it was grounded in the

model's exceptional performance, adaptability, and robustness in handling the intricacies of our dataset.

In the journey of model selection, this section serves as an introduction to the reasoning behind favoring the Random Forest Classifier over other contenders. Beyond a mere selection, this model stood out for its prowess in addressing common challenges such as overfitting, accommodating non-linear patterns, and providing valuable insights into feature importance. As we embark on the exploration of model intricacies, the Random Forest Classifier takes center stage, promising a comprehensive understanding of our data and effective prediction capabilities. We have

3.5.1 Logistic Regression

In the world of predicting outcomes, Logistic Regression plays a key role, especially in understanding and foreseeing discrete results. Let's dive deep into Logistic Regression, uncovering its complexities and acknowledging its significance in making predictions. At the core of Logistic Regression is the logistic function—a mathematical tool that gracefully turns combinations of features and coefficients into probabilities. Imagine it as an S-shaped curve, elegantly capturing how likely an event is to happen. The Logistic Regression equation tells a story of how coefficients (denoted as β) interact with features (represented as X), giving us a detailed understanding of how each feature influences the probability of a particular outcome.

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

$$\ln\left(\frac{P}{1 - P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (ii)$$

Here equation (i) represents the Logistic Function where P(Y=1) is the probability of event happening, e is the base of the natural logarithm, $\beta 0$ is the intercept and $\beta 1, \beta 2, \ldots, \beta n$ are the coefficients for the features $X_1, X_2, \ldots, X_n$.

And equation (ii) is the Logistic Regression Function.

Odds ratios, part of the Logistic Regression equation, provide insights into the likelihood of an event occurring compared to not occurring. These ratios, when interpreted using logarithms,

deepen our understanding of relationships within the data. To maintain the integrity of predictions, Logistic Regression employs regularization techniques like L1 and L2 regularization. These act as guardians, preventing overfitting and striking a balance between a model's complexity and its ability to generalize. In the realm of probabilistic classification, adjusting the threshold decision boundary allows for a tailored balance between precision and recall. This dynamic tuning ensures the model behaves according to the specific needs of the predictive task. As we navigate Logistic Regression, it transforms from a mere mathematical concept into an intuitive companion, helping us decipher patterns within datasets. Its interpretability, computational efficiency, and adaptability make Logistic Regression a timeless ally in the evolving landscape of predictive modeling methodologies.

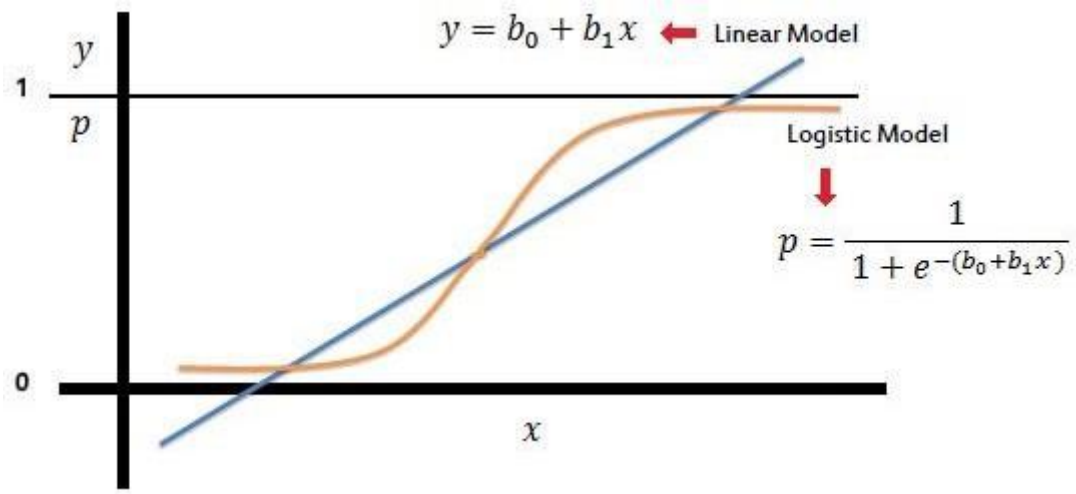

Fig 3.5.1.1: Logistic Regression [22]

3.5.2 Decision Tree Classifier

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. In the world of predictions, the Decision Tree Classifier is like a guide, helping us understand patterns and make choices. Let's take a closer look at this tool, figuring out how it works and why it's useful. Think of a Decision Tree as a map or flowchart. Each stop on the map is a decision,

like choosing between different paths. The path we follow depends on specific features or factors, and each path leads to a prediction. The Decision Tree makes decisions by splitting the data into groups based on features. It wants these groups to be as similar as possible when it comes to the outcome we want to predict. It's like sorting things into different piles to make the best prediction.

While the Decision Tree algorithm involves a series of decisions and branches based on features in the data, it doesn't have a single equation like some other models such as linear regression or logistic regression. The process is more like a series of conditional statements. However, there are key concepts and calculations used within the Decision Tree algorithm:

Gini Impurity:

- At each decision node, the algorithm calculates the Gini impurity, a measure of how often a randomly chosen element would be incorrectly classified.
- The formula for Gini impurity is usually represented as:

$$I_G(p) = 1 - \sum_{i}^{J} p_i^2 \dots \dots \dots \dots \dots (iii)$$

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

Information Gain:

- Information gain is used to decide which feature to split on. It measures the reduction in entropy or uncertainty after a dataset is split.
- The formula for information gain is:

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

  Where D is the dataset, A is the attribute to split on, D_v is the subset of D for a given value of A, and H is the entropy.

While these formulas are part of the Decision Tree algorithm, the overall process involves recursively selecting features and splitting the dataset based on certain criteria until a stopping

condition is met. The simplicity of Decision Trees lies in their hierarchical structure and the series of decisions rather than a single mathematical equation. Decision Trees is that they're easy to understand. One can follow the map of decisions step by step, seeing how the model reaches a particular prediction. However, sometimes Decision Trees can get too specific and try to memorize the data instead of understanding the real patterns. To avoid this, we can use tricks like pruning, which trims the tree to focus on the essential decisions. We can also use a bunch of Decision Trees together, called a Random Forest, to get more accurate predictions. So, in a nutshell, the Decision Tree Classifier is like a helpful guide that uses simple decisions to understand and predict outcomes. It's easy to follow but needs a bit of tweaking to make sure it's not too focused on tiny details.

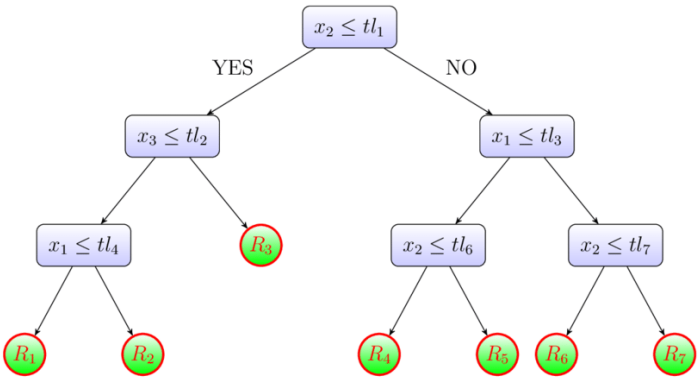

Fig 3.5.2.1: Decision Tree [23]

3.5.3 KNN

In the interesting world of making predictions, let's take a closer look at the k-Nearest Neighbors (KNN) method—it's like having friendly neighbors to help you out. Imagine the data as houses in a digital neighborhood. KNN, acting like a trustworthy neighbor, figures out how close or far these houses are from each other, creating a kind of digital community. Now, the "K" part is about deciding how many nearby opinions to consider, similar to choosing the right amount of friendly advice.

$$\text{Distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (v)$$

For multiple dimensions, the formula extends accordingly. Once distances are calculated, KNN selects the "k" nearest neighbors and makes predictions based on their collective input. The choice of distance metric and the way votes are combined depend on the specific implementation and requirements of the algorithm. Let's think of the decision-making process as a teamwork in this digital neighborhood. For choices like picking categories, KNN goes with what most neighbors say. When predicting numbers, it combines everyone's input to get an average. KNN likes things to be simple and fair, preferring not to get too confused in its digital community. Exploring the good and not-so-good sides of using KNN shows it's simple and easy to understand, a bit like getting advice from friendly neighbors. It works well when there are clear patterns in your data, although sometimes it might take a bit longer when dealing with a lot of digital neighbors.

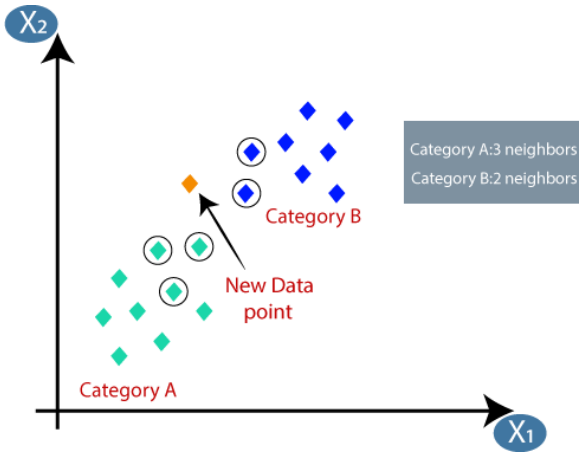

Fig 3.5.3.1: KNN [24]

3.5.4 Proposed Model (Random Forest Classifier)

The machine learning algorithm that has been used in this research is Random Forest Classifier or RF Classifier. Random forest classifier is a supervised machine learning technique. This algorithm is applicable for both classification and regression problems. One of the amazing features of this algorithm is it has the ability to handle continuous variables in case of regression problems and also it can handle the categorical variables in case of classification problems. It stands on the idea of ensemble learning. Ensemble learning is a technique which combines

multiple classifiers. This approach can solve complex problems and thus improve the performance of any model. Just like the name of the algorithm, it contains numerous decision trees on various subsets of the given dataset. It takes the average of the decision trees which can improve the accuracy of the dataset. The algorithm does not rely on a single decision. It takes the prediction of each tree and performs a voting operation. According to the majority of votes, the algorithm generates the final output and thus delivers more accurate results. So, it is obvious that the higher number of trees means higher accuracy of the algorithm.

Before we dive into the working procedure, we must know what ensemble technique is. Ensemble means combining multiple models. It means that a number of models is used in this algorithm to predict an output. There are two types of methods generally used by ensemble technique. They are:

a) Bagging: It is also known as Bootstrap Aggregation. It selects any random sample or subset from the given dataset. The final output in this technique is generated through majority voting. Random Forest Algorithm is one of the finest examples of this class.

b) Boosting: In this technique several simple models work in a combined way to generate a stronger and more accurate model. The final model generates a better output or accuracy.

As discussed above, RF classifier can be applicable in both classification and regression problems. We should know how this algorithm works differently for these two types of problems.

3.5.4.1 Regression Problems:

To solve the regression problems, mean squared error of MSE approach is followed. The formula is:

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (fi - yi)^2 \quad \dots\dots\dots\dots\dots\dots (vi)$$

In this equation, N means the total number of data points, fi represents the value that the model has returned and yi represents the actual value of data point i. The distance between each node

and predicted actual value is calculated through this formula. It gives the decision to select which branch is better choice for the forest.

3.5.4.2 Classification Problems:

To solve the classification problems, Gini index approach is mostly followed. The formula is:

$$Gini = 1 - \sum_{i=1}^{C}(p_i)^2 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (vii)$$

The class and probability are used by this formula to calculate the Gini of each branch. It determines which branch occurs frequently. pi means the relative frequency of the class which is being observed and c is the total number of classes.

Sometimes entropy formula is also used to see how the nodes branch in the decision tree. The entropy formula is:

$$Entropy = \sum_{i=1}^{C} -p_i * log_2(p_i) \dots\dots\dots\dots\dots\dots\dots\dots (viii)$$

Entropy examines the probability of a particular result to have the decision of which branch the node needs to take. It is more complex operation than the Gini approach because of the presence of logarithmic function in the formula.

The mathematics of RF classifier is very important to have a better view on the difference among node, leaf and branch. It gives a knowledge about using different formulas to come to a final decision.

3.5.4.3 Working Procedure of RF Classifier

The working procedure of this algorithm consists of multiple steps. The steps are described below:

a. Distinct subset of both data points and features is chosen to build each decision tree. In a mathematical way it can be said that n random records and m random features are selected from the dataset which has k numbers of records.

b. For every unique sample, unique decision trees are formed.

c. Every decision tree produces an output.

d. A majority voting operation is performed where the output having the most votes is considered as the final output or prediction.

3.5.4.4 Performance Enhancement Techniques

To achieve the desired output from any individual approach, training that model with necessary techniques is a must. In our study, we have applied several techniques to enhance the performance and accuracy of our proposed model. Those techniques are described below:

a) Increasing the number of trees:

It is one of the most efficient techniques to have a better output from our model. Increasing the number of trees mean the algorithm can capture more complex patterns of the dataset and thus decrease the impact of various noisy features of the dataset. The number of trees is defined as 'n-estimators'. Increasing the number increases the accuracy of the model potentially. Each tree learns from different subsets and thus improves the performance of the model. We have set the 'n-estimator' value as 500. It means that the algorithm will create 500 different decision trees.

b) Tuning Hyperparameters:

This is a very important step to optimize the performance of an algorithm. In case of RF Classifier, selecting the perfect hyperparameters can affect the output significantly. There are different hyperparameters like 'n_estimators', 'min_samples_leaf' and 'max_depth'. These hyperparameters can have a huge impact on number of trees, minimum sample required for leaf nodes and depth of each tree. We have used GridSearchCV. It is a wonderful technique to fetch the optimal value of the hyperparameter. This method runs a search operation across the predefined parameter grid which helps to get best possible combination of hyperparameters.

c) Cross Validation:

It is a common technique to measure the performance and capability of any machine learning model. In case of RF Classifier, it provides a more robust model accuracy. 'cross_val_score' function is applied in this technique. It is a function that perform k-fold cross-validation. The

average found from this k number of iterations provides the model's actual performance. The advantage we get by applying this method is that the uncertainty caused by train, test, splits is reduced by averaging the outcomes of multiple folds and thus it increases the accuracy of any model.

d) Ensemble Random Forest Models:

Ensemble of models is an important feature of machine learning which utilizes the collective performance of multiple models to increase the capability of a model. By unifying multiple models, the final output is received. 'VotingClassifier' plays an important role in case of combining multiple models. In our study a RF Classifier of 100 'n-estimators' and another RF Classifier of 200 'n-estimators' have been used. The strength of these method is that it increases the capability of the algorithm by making it able to handle complex relationships in the data.

The Random Forest Classifier comes out as a perfect example of ensemble learning which shows the tremendous power of it. By combining results from various decision trees, it creates a very precise and accurate model which gives the best possible result. It has the super ability to solve the issues related with overfitting. If the algorithm can be used accurately, it can work with any sort of data. This algorithm is easy to use, training data is very fast and finds the accurate results.

CHAPTER 4

EXPERIMENTATL RESULT AND DISCUSSION

4.1 Performance Analysis

To evaluate the performance and capability of the model we have proposed, we have relied on multiple parameters. These parameters show us how good and reliable a machine learning method is. The results of these parameters simply reflect the improvement we have made to the model for better performance and accuracy. Key metrics including precision, recall, f1score along with macro average, weighted average and ROC curve with the AUC score have been the selected parameters for the analysis. The parameters are described below:

Precision: This is a machine learning matrix which defines the ratio of 'True Positive (TP)' and combined positive predictions which means the total number of 'True Positive or TP' and 'False Positive or FP [28]. The equation is:

$$Precision = \frac{TP}{TP + FP} \dots\dots\dots\dots\dots\dots (ix)$$

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 machine learning matrix that indicates the ratio between True Positive and the combination of False Negative and True Positive [28]. The equation is as follows:

$$Recall = \frac{TP}{TP + FN} \dots\dots\dots\dots\dots\dots\dots (x)$$

F1-Score: This is a unified matrix operation of precision and recall. It defines the strength of a model of predicting the correct positive values and not making too many errors by predicting something as positive but it is not [29]. The equation of f1-score:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (xi)$$

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

Support: It indicates the number of instances in the test dataset of each class that has actually occurred. It helps the model to learn which occurrences are more common and which are rarer. It impacts the model on how well it can predict [30].

Accuracy: Accuracy is a term which defines the percentage of accurate prediction. It means that it is a ratio of total number of correct predictions and total number of predictions [31].

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \ldots\ldots\ldots\ldots\ldots (xii)$$

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

Macro Average: It is mainly the average of recall, precision and f1-score across all classes. It means, macro average of these machine learning matrices is acquired by averaging the values of individual classes [32].

Weighted Average: It is used to combine the predictions of various models. Here every model's contribution is measured according to the capacity or expertise of that model [33].

ROC Curve with AUC Score: This curve is a graphical visualization to examine how well the model is performing. It is a visual representation to understand the connection between false positive rate and true positive rate.

AUC of Area Under the Curve is a single value that represents the overall performance of a classifier based on its ROC. It has a range from 0 to 1. If the value of AUC is higher, the model is considered to be better.

4.2 Model Evaluation

Model evaluation is a crucial step for judging a model's accuracy and efficacy. The model we have applied in our research can be evaluated or examined by using different criteria. The examination results show the actual performance of the suggested model which has been used in our study. Machine learning matrix like precision, recall, f1-score, support and most

importantly accuracy. Moreover, we can look at the ROC curve, AUC score and we can judge the overall performance by explaining the confusion matrix. By closely looking at the results we can come to a decision.

As I have applied multiple performance enhancement techniques, it has made some changes in the accuracy of the model. Table 4.2.1.1 shows the difference that I have found.

4.2.1 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

Examining table 4.2.1.1 reveals that diverse techniques were employed to enhance the accuracy of the Random Forest Classifier. Notably, 'Tuning Hyperparameters' emerged as the most effective method, showcasing the success of our proposed approach. This underscores the significance of thorough training in optimizing the model's performance and accuracy enhancement. The best parameters I found for my research are n_estimators=300, max_depth=None, min_samples_leaf=1

4.2.2 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 overall performance of our proposed model can be found. Result of every class those are being considered in the report is great. It reflects how strong the model is while predicting diabetes. Precision, Recall and F1-Score all are same and it is 0.86. It shows the wonderful performance of the proposed way.

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 get an insight that my proposed methodology has achieved an accuracy of 95.67%. Cross-validation results suggest that the model is performing well, with average accuracies of about 97.21% (cv=5) and 97.40% (cv=10) across different subsets of the training data. These scores indicate that the model generalizes effectively to new data.

4.2.3 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:

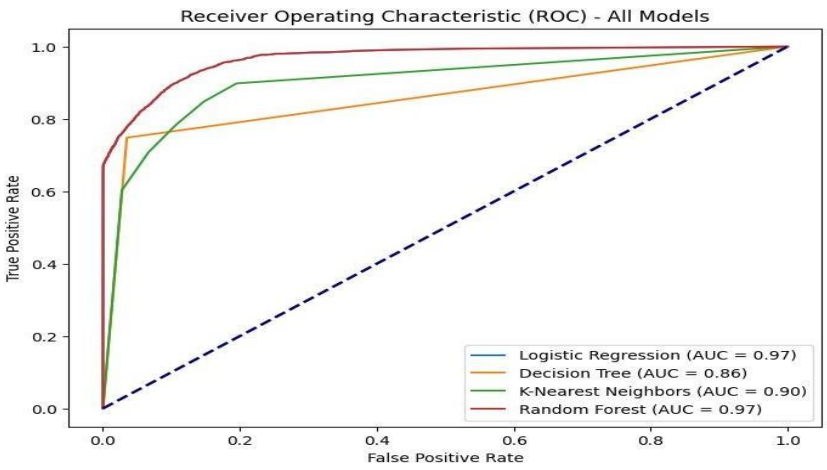

Fig 4.2.3.1 Receiver Operating Characteristics of all models

In evaluating the performance of diverse predictive models using the Area Under the Curve (AUC) metric, the results reveal distinctive strengths and weaknesses. Logistic Regression demonstrates an AUC score of 0.97, suggesting quite a good performance. Also, the Random Forest model excels with a high AUC of 0.97, indicating robust predictive accuracy. The Decision Tree performs well with an AUC of 0.86, showcasing good discriminatory ability. KNearest Neighbors (KNN) achieves a commendable AUC of 0.90, demonstrating strong capabilities in distinguishing between positive and negative instances. These findings offer valuable insights for model selection based on specific dataset characteristics and predictive requirements. We can clearly see the dominance of Random Forest classifier as per the AUC score though Logistic Regression also have the same score.

4.2.4 Confusion Matrix

As I have mentioned before, I have used four algorithms in completing my work. I have generated the confusion matrix for each algorithm. Here the figures show us the difference of confusion matrix among these algorithms:

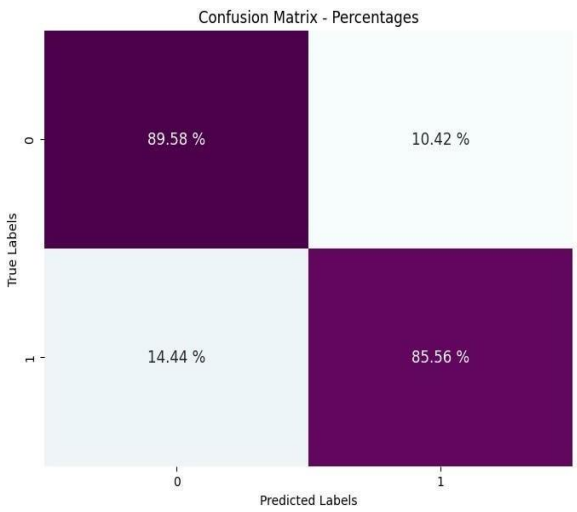

Fig 4.2.4.1 Confusion Matrix of Logistic

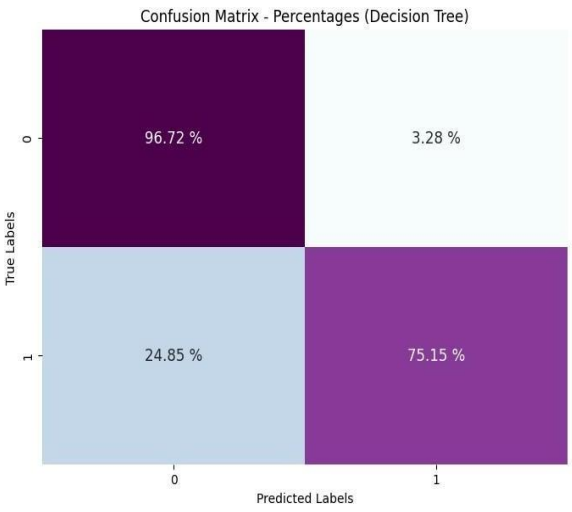

Fig 4.2.4.2 Confusion Matrix of Decision

Regression

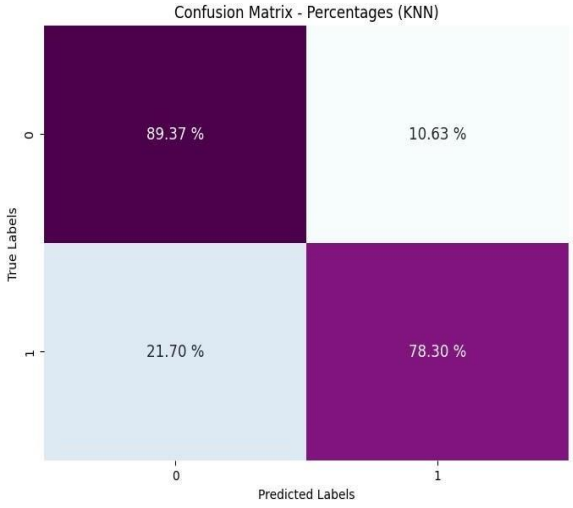

Fig 4.2.4.3 Confusion Matrix of KNN

Tree Classifier

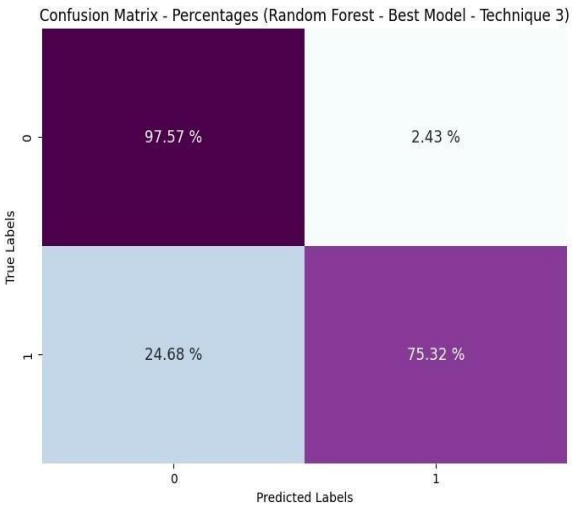

Fig 4.2.4.4 Confusion Matrix of Logistic Regression

4.2.5 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

4.2.6 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 analyzing Table 4.2.6.1, it is found that my proposed model is dominating in every aspect over other models. 95.67% is the accuracy of my proposed model which is pretty higher than other models. Decision Tree Classifier has the second highest accuracy percentage which is 94.81%. KNN classifier has also a good accuracy which is 88.38%. The algorithms I have used here, Logistic Regression offers the lowest accuracy which is 88.23%. It reflects how accurate my proposed model can be in case of predicting diabetes. The AUC score is also the highest among all. This score defines the ability of my proposed model in keeping balance between true positive and false positive rate. As the score is higher than any other models, it can be easily said that my proposed model has better ability to handle this.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on society

Predictive models for diabetes present a transformative potential with profound societal implications. These models serve as invaluable allies in the realm of health, akin to supportive friends guiding individuals on a path to well-being. Their primary strength lies in identifying diabetes risks early, offering a chance for preventive actions and healthier lifestyles. At a broader societal level, the implementation of these predictive models contributes significantly to the efficiency of healthcare systems. By intercepting potential diabetes cases in their early stages, the burden on healthcare resources is alleviated, and the occurrence of severe complications is curtailed. This not only benefits individuals directly but has a ripple effect on the overall well-being of our communities.

Beyond the healthcare landscape, the influence of predictive models extends to fostering a culture of health consciousness. These models act as gentle nudges, encouraging individuals to adopt healthier lifestyles and make informed choices. The communal shift towards proactive health practices is akin to a collective journey towards well-being, with predictive models playing a role as supportive companions.

However, the implementation of such technology requires a conscientious approach. Ensuring universal access to these predictive tools, regardless of individual backgrounds, is paramount. Equity in healthcare is a foundational principle that must be upheld. Additionally, safeguarding the privacy of health information is crucial, akin to maintaining a confidential pact between individuals and their healthcare providers. Fairness and non-discrimination are equally vital considerations in the deployment of predictive models, preventing any unintended disparities.

In essence, these predictive models symbolize not just a technological advancement but a societal shift towards a healthier and more informed collective. They stand as allies in our quest for well-being, urging us to embrace healthier lifestyles and navigate the healthcare landscape with confidence. As we integrate these models into our society, it is imperative to tread carefully, ensuring inclusivity, privacy, and fairness for all. In doing so, we can harness the

positive potential of predictive models to elevate the health and happiness of our society as a whole.

5.2 Impact on Environment

Predicting diabetes may not have a direct impact on the environment, but its positive effects contribute to overall sustainability. By accurately identifying individuals at risk, healthcare resources can be used more efficiently, reducing unnecessary medical tests and treatments. Moreover, the emphasis on lifestyle factors in diabetes prediction models promotes healthy habits, aligning with eco-friendly practices such as plant-based diets and sustainable transportation choices. The data-driven insights from large-scale analyses inform public health policies, allowing for targeted interventions and awareness campaigns that, in turn, reduce the overall burden on healthcare systems. Additionally, the development of machine learning models for diabetes prediction fosters technological innovations, potentially leading to more energy-efficient algorithms and computing solutions, contributing to a greener technological landscape. While the environmental impact may be indirect, the integration of predictive models in healthcare supports broader sustainability goals.

5.3 Ethical Aspects

Using computer programs to predict diabetes involves some important fairness and privacy issues. I needed to make sure people's private health information is kept safe, and they understand how their data will be used. It's crucial to be fair and not let the computer predictions discriminate against certain groups of people. I also had to be clear about how the computer makes its predictions, so everyone can understand. Making sure everyone, no matter where they are or how much money they have, can benefit from these predictions is really important. Doctors and patients need to work together with these computer predictions, not replace human judgment entirely. It's also important to think about how these predictions might affect people in the long run and make sure they get the right support and help. Everyone, including doctors, computer experts, and the people the predictions are about, should work together to make sure everything is done fairly and openly.

5.4 Sustainability Plan

In establishing a sustainability plan for the implementation of predictive models for diabetes, a multifaceted approach is essential. Continuous monitoring and evaluation will be integral, ensuring ongoing accuracy, effectiveness, and ethical adherence. User education and engagement initiatives will play a pivotal role, fostering community understanding and trust through workshops, webinars, and informational campaigns. Prioritizing data security measures, including encryption and access controls, is paramount to protect the privacy of health information. An adaptable design, capable of incorporating technological advances, ensures the longevity and efficiency of the predictive models. Inclusivity and accessibility considerations, addressing language preferences and cultural sensitivity, will guarantee accessibility to a diverse population. Compliance with policies and regulations, staying current with healthcare standards, and responsible use policies are crucial for ethical and legal adherence. Collaboration with stakeholders, financial sustainability planning, community feedback mechanisms, and continuous capacity building for healthcare professionals complete the comprehensive sustainability framework. This approach aims to not only ensure the enduring success of predictive models but also foster positive health outcomes and societal well-being in the long term.

CHAPTER 6

SUMMARY,CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

In summary, the study extensively explored machine learning algorithms for water diabetes prediction, with the Random Forest algorithm standing out as the top performer, achieving an impressive accuracy of 95.67%. Key performance metrics such as precision (86%), recall (86%), and F1 score (86%) underscore the effectiveness of the model, while the AUC score of 0.97 highlights its ability to distinguish between positive and negative instances. The comparative analysis with other algorithms, including Decision Tree, KNN, and Logistic Regression, provided valuable insights into their strengths and weaknesses. This in-depth examination of performance metrics contributes to a nuanced understanding of each model's suitability for diabetes prediction.

Beyond technical aspects, the study delved into the societal, environmental, and ethical implications of deploying these models. It emphasized the necessity for a sustainable plan in their implementation, considering the broader impact on various aspects. the research offers a comprehensive summary of findings, laying the groundwork for future advancements in diabetes prediction models. The study not only contributes to the field of machine learning but also underscores the importance of responsible and thoughtful deployment of predictive models in real-world scenarios.

6.2 Conclusion

In concluding my investigation into predicting diabetes risk, I delved into an array of computer models, namely Logistic Regression, Random Forest, Decision Tree, and k-Nearest Neighbors (KNN). Through a comprehensive examination of their predictive capabilities, I found that my proposed model outshone the others. The assessment of various criteria led to this judgment. Although the alternative algorithms demonstrated commendable performance, my proposed methodology consistently showcased superior results, taking into account factors such as accuracy, precision, recall, and F1 score.

In simpler terms, my model stood out in providing accurate predictions for diabetes risk. This not only establishes it as a robust choice for such tasks but also sets the stage for further advancements in the future. The study not only contributes to our understanding of different

machine learning models but also positions the proposed approach as a frontrunner in the critical domain of predicting diabetes risk.

6.3 Implication for Future Study

Even though I learned a lot about predicting diabetes, there are some things I need to keep in mind. The data I used is crucial, and if it doesn't represent different people well or has biases, my predictions might not work for everyone. Also, if the dataset is missing important information about diabetes, my predictions might not be as accurate. Another thing to consider is that the data is from a specific time, and things might change over time. Looking forward, there are exciting possibilities for improving diabetes prediction. I can focus on getting more varied and inclusive data to make our predictions more useful for different groups of people. Adding more relevant information, maybe from new technologies or complete health records, could also make the predictions better. Exploring data that changes over time can help to understand how diabetes risk evolves. Trying out new ways of combining different prediction methods might give even better results. As technology grows, keeping an eye on new techniques and using them in diabetes prediction can lead to more accurate and helpful predictions. There is also a great scope to deploy the model in a web or mobile app format which can help to easily detect the disease putting some certain value.

References

[1] National Diabetes Information Clearinghouse (NDIC), http://diabetes.niddk.nih.gov/dm/pubs/type1and2/#signs

[2] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516-76531.

[3] Alehegn, M., Joshi, R., & Alehegn, M. (2017). Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. International Research Journal of Engineering and Technology, 4(10), 426-436.

[4] Global Diabetes Community, http://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html

[5] Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. arXiv preprint arXiv:1502.03774.

[6] Schulze, M. B., Heidemann, C., Schienkiewitz, A., Bergmann, M. M., Hoffmann, K., & Boeing, H. (2006). Comparison of anthropometric characteristics in predicting the incidence of type 2 diabetes in the EPIC-Potsdam study. Diabetes care, 29(8), 1921-1923.

[7] Nyamdorj, R., Qiao, Q., Söderberg, S., Pitkäniemi, J. M., Zimmet, P. Z., Shaw, J. E., ... & Tuomilehto, J. (2009). BMI compared with central obesity indicators as a predictor of diabetes incidence in Mauritius. Obesity, 17(2), 342-348.

[8] Jia, Z., Zhou, Y., Liu, X., Wang, Y., Zhao, X., Wang, Y., ... & Wu, S. (2011). Comparison of different anthropometric measures as predictors of diabetes incidence in a Chinese population. Diabetes research and clinical practice, 92(2), 265-271.

[9] Lönn, M., Mehlig, K., Bengtsson, C., & Lissner, L. (2010). Adipocyte size predicts incidence of type 2 diabetes in women. The FASEB journal, 24(1), 326-331.

[10] Chen, H. Y., Chuang, C. H., Yang, Y. J., & Wu, T. P. (2011). Exploring the risk factors of preterm birth using data mining. Expert systems with applications, 38(5), 5384-5387.

[11] Chang, C. D., Wang, C. C., & Jiang, B. C. (2011). Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors. Expert systems with applications, 38(5), 5507-5513.

[12] Aslan, K., Bozdemir, H., Sahin, C., & Noyan Ogulata, S. (2010). Can neural network able to estimate the prognosis of epilepsy patients accorrding to risk factors?. Journal of medical systems, 34, 541-550.

[13] Vijayan, V. V., & Anjali, C. (2015, December). Prediction and diagnosis of diabetes mellitus—A machine learning approach. In 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) (pp. 122-127). IEEE.

[14] Sonar, P., & JayaMalini, K. (2019, March). Diabetes prediction using different machine learning approaches. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 367-371). IEEE.

[15] Komi, M., Li, J., Zhai, Y., & Zhang, X. (2017, June). Application of data mining methods in diabetes prediction. In 2017 2nd international conference on image, vision and computing (ICIVC) (pp. 1006-1010). IEEE. [16] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, 104-116.

[17] Mercaldo, F., Nardone, V., & Santone, A. (2017). Diabetes mellitus affected patients classification and diagnosis through machine learning techniques. Procedia computer science, 112, 2519-2528.

[18] Osman, A. H., & Aljahdali, H. M. (2017). Diabetes disease diagnosis method based on feature extraction using K-SVM. International Journal of Advanced Computer Science and Applications, 8(1).

[19] Nilashi, M., bin Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). An analytical method for diseases prediction using machine learning techniques. Computers & Chemical Engineering, 106, 212-223.

[20] Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences, 29(2), 93-99.

[21] "Diabetes prediction dataset," www.kaggle.com. https://www.kaggle.com/datasets/iammustafatz/diabetesprediction-dataset?select=diabetes_prediction_dataset.csv (accessed Jan. 02, 2024).

[22] "Logistic Regression," saedsayad.com. https://saedsayad.com/logistic_regression.htm

[23] javaTpoint, "Machine Learning Decision Tree Classification Algorithm - Javatpoint," www.javatpoint.com, 2021. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

[24] JavaTpoint, "K-Nearest Neighbor (KNN) Algorithm for Machine Learning - Javatpoint," www.javatpoint.com, 2021. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

[25] "Simple Gaussian Naive Bayes Classification — astroML 0.4 documentation," www.astroml.org. https://www.astroml.org/book_figures/chapter9/fig_simple_naivebayes.html (accessed Jan. 02, 2024).

[26] V. Alto, "Understanding AdaBoost for Decision Tree," Medium, Jan. 31, 2020. https://towardsdatascience.com/understanding-adaboost-for-decision-tree-ff8f07d2851

[27] "ML - Gradient Boosting," GeeksforGeeks, Aug. 25, 2020. https://www.geeksforgeeks.org/ml-gradientboosting/

[28] "Precision and Recall in Machine Learning - Javatpoint," www.javatpoint.com. https://www.javatpoint.com/precision-and-recall-in-machine-learning

[29] R. S. Punia , "How to Measure the Performance of Your Machine Learning Models: Precision, Recall, Accuracy, and F1…," Medium, May 06, 2023. https://medium.com/@mycodingmantras/how-to-measure-theperformance-of-your-machine-learning-models-precision-recall-accuracy-and-f1-855702df048b

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

[31] "Accuracy vs. precision vs. recall in machine learning: what's the difference?," www.evidentlyai.com. https://www.evidentlyai.com/classification-metrics/accuracy-precisionrecall#:~:text=Accuracy%20is%20a%20metric%20that

[32] K. Leung, "Micro, Macro & Weighted Averages of F1 Score, Clearly Explained," Medium, Jan. 09, 2022. https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f

[33] "WeightedAverage Formula," GeeksforGeeks, Apr. 28, 2022. https://www.geeksforgeeks.org/weightedaverage-formula/ (accessed Jan. 20, 2024).

ABIR

ORIGINALITY REPORT

25%	23%	12%	17%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	7%
2	Submitted to Daffodil International University Student Paper	3%
3	Xue-Hui Meng, Yi-Xiang Huang, Dong-Ping Rao, Qiu Zhang, Qing Liu. "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors", The Kaohsiung Journal of Medical Sciences, 2013 Publication	1%
4	Submitted to University of Hertfordshire Student Paper	1%
5	Submitted to Sheffield Hallam University Student Paper	1%
6	Sudipta Priyadarshinee, Madhumita Panda. "Chapter 22 Machine Learning Algorithms for Diabetes Prediction", Springer Science and Business Media LLC, 2022 Publication	<1%