
A Diabetes Prediction System Using Machine Learning in a Web-Based Application Software

By

| | |
|--------------------|----------------|
| Md. Rashadul Islam | ID:19202130169 |
| Md. Faysal Mahmud | ID:19202103174 |
| Sadia Sultana | ID:19202103184 |
| Dipa Rani | ID:19202103201 |
| Md. Sahal Kabir | ID:17183103051 |

Submitted in partial fulfillment of the requirements of CSE 498B

**Bachelor of Science in
Computer Science and Engineering**



Department of Computer Science and Engineering
Bangladesh University of Business and Technology

January, 2024

Declaration

We do hereby declare that the project works presented here entitled as, “A Diabetes Prediction System Using Machine Learning in a Web-Based Application Software” are the results of our own works. We further declare that the project has been compiled and written by us and no part of this project has been submitted elsewhere for the requirements of any degree, award or diploma or any other purposes except for this project. The materials that are obtained from other sources are duly acknowledged in this project.

Signature of Students

Md. Rashadul Islam

Id: 19202103169

Md. Faysal Mahmud

Id: 19202103174

Sadia Sultana

Id: 19202103184

Dipa Rani

Id: 19202103201

Md. Sahal Kabir

ID: 17183103051

Approval

I do hereby declare that the project works presented here entitled, A Diabetes Prediction System Using Machine Learning in a Web-Based Application Software is the outcome of the original works carried out by Md. Rashadul Islam, Md. Faysal Mahmud, Sadia sultana, Dipa Rani & Md. Sahal Kabir under my supervision. I further declare that no part of this project has been submitted elsewhere for the requirements of any degree, award diploma, or any other purposes except for this project. I further certify that the dissertation meets the requirements and standards for the degree of Doctor of Philosophy in Computer Science and Engineering.

Supervisor

Sudipto Chaki

Assistant Professor

Department of Computer Science Engineering

Bangladesh University of Business and Technology

Chairman

Md. Saifur Rahman

Assistant Professor & Chairman

Department of Computer Science Engineering

Bangladesh University of Business and

Technology

Dedication

We would like to dedicate this research to our loving parents, teachers, friends, and
who loved us for all their love and inspiration.

Acknowledgement

We are deeply thankful to Bangladesh University of Business and Technology (BUBT) for providing us such a wonderful environment to peruse our project. We would like to express our sincere gratitude to Sudipto Chaki, Assistant Professor, CSE, BUBT. We have completed our project with his help. We found the project area, topic, and problem with his suggestions. He guided us with our study, and supplied us many articles and academic resources in this area. He is patient and responsible. When we had questions and needed his help, he would always find time to meet and discuss with us no matter how busy he was. We also want to give thanks to our CSE department. Our department provide us logistic supports to complete our project with smoothly. We would also like to acknowledge our team members for supporting each other and be grateful to our university for providing this opportunity for us.

Abstract

The development of prediction algorithms to help with early diagnosis and prevention has been motivated by the prevalence of diabetes, a disease that is common and has large worldwide ramifications. The main goal is to make reliable diabetes risk forecasts available to people all around the world. This method provides a workable solution for early intervention and health-conscious decision-making by overcoming the shortcomings of existing systems and making machine learning approachable to all users. Choosing and preparing a substantial diabetic dataset from Kaggle is the first step in the study procedure. Preparing the dataset for analysis with meticulous data preprocessing ensures data quality. The performance and interpretability of the model are then improved by using feature selection techniques to determine the most informative variables. A variety of classification techniques, such as LR, RF, GBC, SVC, KNN, and DT, can be evaluated and chosen thanks to the dataset's division into training and testing sets. Based on their precision on the test data, the best-performing models among these classifiers are determined. The forecasts from these models are then integrated to produce a more powerful and dependable prediction engine using an ensemble technique. To select the most accurate model, the updated classifiers are thoroughly assessed. Users can enter their medical information and receive tailored diabetes risk projections using the final model that was chosen, which is integrated into a user-friendly web-based application. By democratizing the advantages of machine learning and advanced analytics, this program equips individuals to make wise decisions about their health. The findings of this study demonstrate how an ensemble technique can increase prediction reliability and accuracy. The created web-based application provides users with an easy-to-use interface that allows them to access and use the prediction engine from any location, aiding in the fight against diabetes' early diagnosis and intervention efforts. To enable people to prioritize their health and well-being, the project intends to make the benefits of machine learning and sophisticated analytics broadly available through a user-friendly online platform.

Contents

| | |
|--|------------|
| Declaration | iv |
| Approval | iv |
| Dedication | v |
| Acknowledgement | vi |
| Abstract | vii |
| List of Figures | xi |
| List of Tables | xii |
| List of Abbreviations and Acronyms | xii |
| 1 Introduction | 1 |
| 1.1 Introduction | 1 |
| 1.2 Problem Statement | 1 |
| 1.3 Research Background | 3 |
| 1.4 Research Objectives | 3 |
| 1.5 Motivations | 4 |
| 1.6 Significance of the Research | 4 |
| 1.7 Project Contribution | 5 |
| 1.8 Project Report Organization | 6 |
| 1.9 Summary | 7 |
| 2 Literature Review | 8 |
| 2.1 Introduction | 8 |
| 2.2 Related work | 8 |
| 2.3 Problem Analysis | 16 |
| 2.4 Summary | 19 |

| | | |
|----------|---|-----------|
| 3 | Methodology or Proposed Framework | 20 |
| 3.1 | Introduction | 20 |
| 3.2 | Proposed Framework | 20 |
| 3.2.1 | Searching Dataset | 20 |
| 3.2.2 | Database Selection | 21 |
| 3.2.3 | Data Pre-processing | 21 |
| 3.2.4 | Describe Each Part of the Framework | 22 |
| 3.2.5 | Data Splitting | 24 |
| 3.2.6 | Classification of Algorithm | 25 |
| 3.2.7 | Selecting Classifier | 30 |
| 3.2.8 | User-friendly UI | 31 |
| 3.3 | Summary | 31 |
| 4 | Implementation and Testing | 33 |
| 4.1 | Introduction | 33 |
| 4.2 | System Setup | 33 |
| 4.3 | Evaluation | 33 |
| 4.4 | Result and Discussion | 34 |
| 4.4.1 | Dataset: | 34 |
| 4.4.2 | Data Pre-processing: | 34 |
| 4.4.3 | Training Model & Accuracy Score: | 37 |
| 4.4.4 | Ensemble Method: | 37 |
| 4.4.5 | Web Interface: | 38 |
| 4.5 | Summary | 40 |
| 5 | Standards, Constraints, and Milestones | 41 |
| 5.1 | Introduction | 41 |
| 5.2 | Standards | 41 |
| 6 | Conclusion | 43 |
| 6.1 | Introduction | 43 |

| | | |
|-------|---------------------------------------|-----------|
| 6.2 | Conclusion | 43 |
| 6.3 | Limitation and Future Works | 44 |
| 6.3.1 | Limitation | 44 |
| 6.3.2 | Future Works | 44 |
| | References | 48 |

List of Figures

| | | |
|------|--|----|
| 3.1 | Process workflow of diabetes prediction system | 20 |
| 3.2 | Random forest classifier architecture | 25 |
| 3.3 | KNN architecture | 26 |
| 3.4 | SVM Classifier architecture | 27 |
| 3.5 | Decision Tree architecture | 28 |
| 3.6 | Logistic Regression architecture | 29 |
| 3.7 | Gradient Boosting Classifier architecture | 30 |
| 4.8 | Outlier Check with Boxplot | 35 |
| 4.9 | After removing outlier | 36 |
| 4.10 | Correlation Heatmap | 37 |
| 4.11 | ROC Curve | 38 |
| 4.12 | Web App Interface 01 | 39 |
| 4.13 | Web App Interface 02 | 39 |

List of Abbreviations and Acronyms

| | |
|------------|---------------------------------|
| AI | Artificial Intelligence |
| ML | Machine Learning |
| RF | Random Forest |
| SVC | Support Vector Classifier |
| GBC | Gradient Boosting Classifier |
| UI | User Interface |
| SVM | Support Vector Machine |
| LR | Logistic Regression |
| DT | Decision Tree |
| KNN | K Nearest Neighbor |
| CSS | Cascading Style Sheets |
| GUI | Graphical User Interface |
| RAM | Random Access Memory |
| UCI | University of California Irvine |

1 Introduction

1.1 Introduction

Diabetes is a common disease that has negative consequences on people worldwide. It results from several causes, including obesity and high blood sugar levels, which affect the way the hormone insulin works, causing improper carbohydrate metabolism and higher blood sugar levels. Diabetes is mainly brought on by the body not producing enough insulin. According to an IDF assessment, Bangladesh will rank eighth in 2045, among countries with the highest number of individuals (20-79 years) with diabetes (13.7 million cases)[1]. The best way to find out if someone has diabetes or not typically entails consulting a doctor. They will examine the patient physically and go over their medical background. This makes it easier to analyze any symptoms, habits, diabetes in the family, and other pertinent data. The doctor may prescribe specific tests to confirm the diagnosis after the initial evaluation. By entering exact medical data through a website or mobile app, we may cross-validate this forecast with the help of AI from any place in the world. It is important to remember yet that consulting medical advice is still necessary for a precise diagnosis and the most recommended option for care. The use of AI technology can be a helpful hand for prediction, but it shouldn't take the place of professional medical advice and knowledge.

1.2 Problem Statement

A diabetes prediction system based on machine learning has various benefits across multiple dimensions. First of all, logistic regression is a straightforward and understandable method for predicting diabetes. It is suitable for huge datasets due to its fast calculation and training times. Logistic regression also provides probabilities for predicting outcomes, which allows for a better understanding and assessment of prediction confidence. Furthermore, it can efficiently handle both binary and multiclass classification issues. The K-Nearest Neighbors (KNN) classifier has the advantage of not assuming any particular data distribution. It is a non-parametric

technique that can capture complex feature interactions. KNN performs well with small to medium-sized datasets and is appropriate for binary and multiclass classification applications. The Support Vector Classifier (SVC) works effectively in high-dimensional spaces and when there is a clear boundary between classes. It is capable of dealing with both linear and non-linear classification issues. SVC's regularization parameter enables for model complexity control, and it can handle binary and multiclass classification jobs successfully. Decision tree classifiers are simple to understand and apply. They are versatile because they can handle both numerical and category characteristics. Non-parametric decision trees can capture non-linear correlations between features. They can also handle missing values and outliers successfully, giving the model robustness. Random forest classifiers have a great degree of accuracy and robustness. Random forests prevent over-fitting and handle high-dimensional data well by averaging many decision trees. They also rank feature importance, which might help with feature choice and interpretation. Random forest classifiers are capable of handling binary and multiclass classification tasks, as well as missing values and outliers. Finally, a gradient-boosting classifier is an ensemble method for combining weak learners to produce a strong learner. It is extremely accurate and frequently outperforms other algorithms. Gradient boosting can handle numerical and categorical features and is resistant to missing data. It provides a feature relevance rating, similar to random forests, to aid in feature selection and interpretation. While these classifiers offer many advantages, it is critical to understand their limitations. Logistic regression is based on the assumption of a linear relationship between features and the target variable, therefore it may struggle with complicated interactions or non-linear correlations. During the prediction phase, KNN classifiers can be computationally expensive, necessitating careful consideration of dataset size and feature scaling. SVC is sensitive to hyperparameter selection and may necessitate the use of other algorithms to derive probability estimates. Decision trees can be over-fitting and have difficulty capturing complex relationships. When evaluating the cumulative influence of all the trees, random forests might be computationally expensive and difficult to analyze. Gra-

dient boosting classifiers are hyper parameter sensitive, and training time can be lengthy, especially when a large number of iterations are used.

1.3 Research Background

Jobeda Jamal Khanam[9] et al. Proposed A comparison of machine learning Algorithm for diabetes Prediction system The major goal of the study is to data mining and machine learning approaches to predict diabetes in Patients To accomplish the above goals the research makes use of Logistic Regression, support vector Machine, Decision Tree, k-nearest Neighbors, Random forest, Naive Bayes and Adaptive Boosting. The Lackings of the study is, that it neglects Particular information about the Preparation Procedures used, such as outlier removal, handling missing values and normalizing. Sanskruti Patel[10] et al. proposed predicting the risk of diabetes at an early stage. using a machine learning approach. The major goal of the study is to use machine learning technique to Create a Predictive model for the early identification of diabetes to accomplish the above goals the research make use of Gaussian, Naive Bayes, Random Forest, Support Vector classifier, and Multinomial Naive Bayes The lacking of the study is, the paper doesn't address how this imbalance was handled during model training and evaluation, which can affect them. Jingyu Xuc[11] et al. Proposed Diabetes Prediction Method Based on Machine Learning The major goal of the study is to Performance comparison of several machine learning algorithms for diabetes prediction. To accomplish the above goals the research makes use of SVM is used as a binary classification technique. The Naive Bayes classifier is used. Lackings of the study is that the findings would be seen from a more unbiased viewpoint if the limitations were discussed.

1.4 Research Objectives

We are pleased to present a revolutionary diabetes prediction system that employs advanced machine learning algorithms to give accurate and dependable predictions. We used an ensemble method in which we carefully chose and mixed predictions from six separate classifiers. We have determined the top three to four models

that consistently display higher performance through this procedure. We discovered the best classifier model for diabetes prediction from this ensemble of models. This model has demonstrated outstanding accuracy and robustness while dealing with varied datasets. We created a user-friendly web-based application using our high-performing classifier. Users can quickly access our cutting-edge prediction technology and make diabetes predictions using our web application. Users can acquire specific results and insights about their chance of having diabetes by entering their specific medical data into the program. This enables people to take preventative measures and make educated health decisions. Using the collective knowledge of several classifiers, our unique blending process ensures that our prediction engine produces dependable and exact results. We extensively tested and verified our method to assure its efficacy across a wide range of people and demographics. We hope to make the benefits of machine learning and advanced analytic available to everyone by providing this novel diabetes prediction system via a user-friendly online application. Users may now conduct their forecasts reliably and intuitively, allowing them to prioritize their health and well-being.

1.5 Motivations

The motivation behind working on this project because not every user understands artificial intelligence or knows of it. Therefore, developing an architecture for prediction is a bad idea. So we have created a user-friendly online web-based form that clearly explains the back-end code to make machine learning accessible to all users. The users would have the advantage of being able to cross-validate from anywhere with this.

1.6 Significance of the Research

Getting accurate and dependable forecasts, our diabetes prediction system employs powerful machine learning techniques. We painstakingly chose and mixed predictions from several different classifiers to construct an ensemble that regularly exceeds the competition. Our approach produces more trustworthy and precise results by

combining the strengths of logistic regression, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), decision trees, random forests, and gradient boosting classifiers. We created an easy-to-use web-based application to ensure accessibility. Users can enter their medical information to receive customized predictions and insights on their risk of developing diabetes. This enables people to make informed health decisions and take preventative interventions. Our prediction system has been rigorously tested and validated over a wide range of communities and demographics. We ensured the effectiveness and stability of our solution by adding cutting-edge analytic and a unique mixing process. Moving forward, we intend to improve our diabetes prediction system by investigating additional feature engineering techniques, refining the ensemble learning approach, and incorporating cutting-edge algorithms. We will continue to collect real-world data and work with healthcare professionals to validate and fine-tune our system for clinical application. We hope to make the benefits of machine learning and advanced analytic available to everyone by providing this novel diabetes prediction system via a user-friendly online application. Users can now undertake credible and intuitive forecasts with ease, allowing them to prioritize their health and well-being.

1.7 Project Contribution

The overall contribution of the research work includes,

- **Data Pre-processing:** Diabetes Dataset pre-processing is the first step where we need to resolve missing values, normalize features, and perform feature selection.
- **Selecting Six Classifiers:** After pre-processing the data, we choose six different classifiers to train and evaluate for diabetes prediction (those are - logistic regression, random forest, support vector machines, k-nearest neighbors, gradient boosting, and decision tree). Because each classifier has various strengths and weaknesses, using numerous algorithms allows us to experiment with different modeling approaches.

- **Training and evaluation:** After data pre-processing is complete, we train each of the six classifiers. As a part of this process, the dataset is divided into training and testing sets, the classifiers are fitted to the training data, and their performance is assessed using the proper metrics, including accuracy, precision, recall, and F1-score, on the testing data.
- **N-symbol-based machine learning classifier:** We select the top three or four classifiers out of the six classifiers depending on accuracy. The n-symbol technique is used to train and assess these chosen classifiers to increase prediction precision and accuracy.
- **Diverse Data Selection:** We include diverse data in the training and testing sets to improve the prediction models' robustness and generalizability. This entails taking into account elements including diabetes-related demographic data, clinical information, lifestyle factors, bio-markers, and environmental factors.
- **User Interaction:** We design a web-based application that enables user interaction after obtaining the best accuracy from the chosen classifiers. Based on input from the user or uploaded data, the program uses the trained models to generate predictions or risk assessments for diabetes. Users of this program can determine their risk of diabetes and get tailored information.

1.8 Project Report Organization

The rest of the book is organized in the following way. In Chapter 1, we will show the background and related research studies. After that,

- **In Chapter 2,** we will discuss about Literature review and its background. We also review some existing papers to identify strengths and weaknesses. Additionally, we explore authors good part and bad part to get a clear overview.
- **In Chapter 3,** consists of our Methodology. We will discuss about the workflow and its brief. The classification algorithms its uses and so on.

1.9 Summary

This chapter comprises a broad overview of the problem such as what are we specifically targeting, what are the purposes of our thesis work along with the motivation of the output of the thesis work. This section also represents the overall steps on which we carried out our thesis work.

2 Literature Review

2.1 Introduction

The diabetes prediction using machine learning literature review investigates the existing body of research that focuses on utilizing various machine learning approaches to predict and diagnose diabetes. The goal of this review is to identify trends, approaches, and issues in the field, while also highlighting notable studies that have contributed to breakthroughs in accurate diabetes prediction. This review will provide insights into the effectiveness, limitations, and potential future approaches for constructing strong machine-learning models for diabetes prediction by synthesizing the data of various investigations.

2.2 Related work

Priyanka Indoria et al. proposed A Survey: Detection and Prediction of Diabetes Using Machine Learning Techniques. The goal is to use machine learning methods like Artificial Neural Networks (ANNs) and Bayesian Networks (BNs) to improve the precision of identifying diseases like diabetes and cardiovascular disease. To categorize and diagnose diabetes and cardiovascular disorders, the study uses a variety of machine learning models, including Artificial Neural Networks (ANNs), Naive Bayesian Networks, and Probabilistic Neural Networks (PNN). To create precise prediction models, the research examines medical files and datasets pertaining to diabetes, cardiovascular diseases, and risk factors. With accuracy rates exceeding 80%, the suggested machine learning models, such as PNN and Naive Bayesian Networks, successfully diagnosed diabetes and determined the risk levels for cardiovascular illnesses. Since feature independence is assumed while using Naive Bayesian Networks, there may be reduced[2].

N.A. Farooqui et al. construct a Prediction Model for Diabetes Mellitus Using Machine Learning techniques. By examining several diabetes-related characteristics using the Pima Indians diabetes dataset, the aim is to predict diabetes at an

early stage. To categorize diabetic patients, the study uses a variety of machine learning methods, including Decision Trees, K-nearest neighbors, Support Vector Machines, and Random Forests. The Pima Indian diabetes dataset is used by the authors as the input data for training and assessing machine learning models. The study comes to the conclusion that the Random Forest classifier outperforms other machine learning methods and has the highest accuracy of 96.89% for predicting diabetes. To create their own prediction models, researchers interested in diabetes prediction using machine learning might refer to this study and the suggested technique[3].

S. Saru et al. proposed Analysis and prediction of Diabetes using machine learning. The aim is to develop an efficient predictive model to detect diabetes at an early stage by using data mining and classification algorithms to evaluate healthcare data, in particular the Pima Indian diabetes dataset. The analysis of the Pima Indian diabetes dataset uses data mining and classification techniques like Decision Trees, K-nearest neighbors, Naive Bayes, and Support Vector Machines. The main emphasis is on using the features of the given data to predict and diagnose diabetes. According to the research, the proposed prediction model achieved a high accuracy rate of roughly 94.44% for Decision Trees, 79.84% for Naive Bayes, and 93.79% for K-Nearest Neighbors when employing various machine learning approaches. The findings show that the suggested approach is capable of accurately predicting and diagnosing diabetes at an early stage[4].

Neha Sharma et al. proposed Diabetes Detection and Prediction Using Machine Learning. In order to combat the rising diabetes rates, notably in India, early diagnosis via IoT and machine learning is advocated as the major purpose of utilizing modern computing techniques. The strategy strives to enhance treatment effectiveness and quality while empowering patients with self-care and better interaction with medical staff. The method uses IoT for online data collection and remote monitoring, machine learning for diabetes analysis, fuzzy cognitive maps to help with

decision-making, mobile health apps for self-management, and medical imaging and sensors for non-invasive glucose sensing, all of which work together to improve diabetes care and prediction. The challenges include the security and privacy of patient data in the Internet of Things, the inconsistent dependability of machine learning predictions due to data quality, accessibility restrictions, the ethical use of data, and potential limitations of noninvasive sensing techniques like[5].

Ratna Aminah et al. proposed a Diabetes Prediction System Based on Iridology Using Machine Learning. Obtain a wide collection of iridology photos from people, including both normal volunteers and those who have diabetes. Standardize image sizes and clean the iridology photos to get rid of noise and artifacts. Apply image processing techniques to the iridology photos to uncover features that might be a sign of diabetes risk. Based on iridology images and medical data, the system should strive for high accuracy in forecasting diabetes risk. The percentage of accurate predictions over all of the model's predictions is known as accuracy. High recall shows the system is successfully recognizing a significant majority of diabetes-positive cases, whereas high accuracy means the system is making fewer false positive predictions. Users should be able to simply upload iridology photographs and receive their diabetes risk prediction thanks to a user-friendly user interface[6].

Mustafa S. Kadhmi et al. proposed An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approach. Aiming to Select the K-means clustering features that are most pertinent and have the biggest impact on diabetes prediction. To show the usefulness of the suggested classification approach, its performance is compared to that of well-known machine learning algorithms (such as Logistic Regression, Random Forest, and Support Vector Machine). Assess the trained model's prediction performance using suitable metrics including accuracy, precision, recall, F1-score, and ROCAUC. The accuracy and representativeness of the dataset are crucial to these predictions. Results that are less than ideal could be caused by noisy or insufficient data. Choosing pertinent features from

the dataset and evaluating their significance might be difficult and call for in-depth topic knowledge. The created Diabetes Prediction System provides an important tool for early diabetes detection and risk assessment, maybe assisting in early[7].

Safial Islam Ayon et al. proposed Diabetes Prediction: A Deep Learning Approach. The goal of this research is to create a Deep Learning-based diabetes prediction system. Improved prediction accuracy, handling complicated relationships within the data, and offering a scalable and reliable solution for diabetes risk assessment are the main objectives. assemble a diverse and thorough dataset of health-related variables from people, including both healthy and participants with diabetes. Use a deep neural network architecture to forecast the risk of diabetes, such as a convolutional neural network, recurrent neural network, or a combination of the two. The model's capacity for generalization may be hampered by restricted access to a broad and varied dataset. Deep learning models can be complicated and difficult to interpret, making it difficult to explain their judgments[8].

M. Rajeswari et al. proposed A Review of Diabetic Prediction Using Machine Learning. This research aims to use machine learning to predict and identify diabetes. Algorithms with a focus on early identification and increased accuracy. Based on a variety of datasets, including the Pima Indians Diabetes Dataset, different machine learning methods, such as Decision Trees, Naive Bayes, Support Vector Machines (SVM), etc., are used to categorize diabetes. The papers highlight the need for precise and effective prediction models by discussing challenges associated with diabetes prediction, categorization, and early detection. These research findings also show the degrees of accuracy attained by various algorithms. SVM and ensemble algorithms frequently produce higher rates of accuracy[12].

Muhammad Exell Febrian et al. proposed Diabetes prediction using supervised machine learning. The major objective is to use machine learning to forecast diabetes and avoid serious diseases in order to combat the rising prevalence of diabetes. Additionally, contrast the efficacy of the k-Nearest Neighbor (KNN) and Naive Bayes

algorithms for predicting diabetes. In order to predict diabetes using health variables, this work uses supervised machine learning techniques like K-Nearest Neighbor (KNN) and Naive Bayes algorithms. After data integration, cleaning, scaling, and normalization, 10-fold cross-validation is used to verify the model. Naive Bayes beat KNN in experiments on the Pima Indians Diabetes dataset, attaining greater average accuracy (76.07% vs. 73.33%) and precision (73.37% vs. 70.25%). A few of the study's drawbacks are its focus on a small dataset (Pima Indians with Diabetes), the exclusion of cutting-edge techniques like neural networks, and potential feature-dependency problems with Naive Bayes assumptions[13].

Prachet Bhuyan et al. proposed Primitive Diabetes Prediction using Machine Learning Models: An Empirical. The main objective of this project is to develop a prediction model for the early detection of diabetes using machine learning techniques. To accurately forecast diabetes in patients, it is important to identify pertinent features and their relationships. The approach uses a variety of machine learning methods, including Gaussian Naive Bayes, Random Forest, Support Vector Classifier, and Multinomial Naive Bayes, to forecast diabetes based on historical data of linked symptoms. The selection of features and data analysis are the main areas of attention in order to pinpoint the most important characteristics for accurate prediction. To ensure the data's accuracy and completeness, it has been challenging to preprocess and sanitize it. The performance and accuracy of the model can also be affected by how class imbalances are handled. Considering the[14].

Amani Yahyaoui et al. proposed A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques. This research compares deep learning (Convolutional Neural Network) with more conventional machine learning (Support Vector Machine and Random Forest) in order to develop a Decision Support System (DSS) for diabetes prediction. The methodology makes use of the 768 samples and 8 characteristics of the Pima Indians Diabetes dataset. The dataset consists of 60% training sets and 40% testing sets. Both the deep learn-

ing method (CNN) and the conventional machine learning classifiers (SVM and RF) are used to predict diabetes using the dataset. The effectiveness of each algorithm is evaluated in terms of accuracy, precision, recall, and f-measure. According to the trial findings, the Random Forest algorithm predicted diabetes with an accuracy of 83.67%. Using the SVM method[15].

S. M. Mahedy Hasan et al. proposed An Effective Diabetes Prediction System Using Machine Learning Techniques. This study aims to improve classification accuracy for diabetic patients' early diagnosis by using a tree-based machine learning model. In a Tree-Based Machine Learning model for categorization, the researchers apply the Decision Tree (DT), Random Forest (RF), and Extra Trees (ET) approaches. They also employ the Adaptive Boosting (AB) technique to increase the efficiency of these classifiers. To eliminate unnecessary features and increase prediction accuracy, feature selection is done using a Mutual Information (MI)-based method. The study notes difficulties that make enhancing prediction accuracy difficult, including missing values, pointless features, and unequal class distribution in the dataset. On the PIDD dataset, the suggested Tree-Based Machine Learning method for categorizing diabetes patients[16].

Muhammad Azeem Sarwar et al. proposed Prediction of Diabetes Using Machine Learning Algorithms in Healthcare. In order to determine which machine learning algorithm is most effective at predicting diabetes, it is necessary to compare the efficacy and precision of six different machine learning algorithms on a dataset of patient medical records. Naive Bayes, Support Vector Machine, Decision Tree, Logistic Regression, and Random Forest are the chosen algorithms. Random Forest (RF), Logistic Regression (LR), and K-Nearest Neighbors (KNN). The PIMA Indian dataset, which was retrieved from the UCI machine learning repository, is used to train and test the models. We carry out feature selection, model training, model evaluation, and data preprocessing. One limitation of the study is the size of the dataset and any attribute values that may be missing. The results show how

accurate each algorithm was at predicting diabetes. SVM and KNN accomplish[17].

P. Moksha Sri Sai et al. proposed a Survey on Type 2 Diabetes Prediction Using Machine Learning. The main objective of the study is to predict and prevent type 2 diabetes using machine learning algorithms on medical data. To implement preventative measures, accurate diabetes prediction, and early risk factor detection are desired. K-means, Logistic Regression, Support Vector Machine (SVM), K-nearest neighbor (KNN), Random Forest, Decision Tree, and Naive Bayes are only a few of the machine learning algorithms used in the research. In order to anticipate diabetes, these algorithms examine and analyze data from medical reports. The results show that SVM gives the highest accuracy (93%), followed by other techniques including KNN, logistic regression, and decision trees. Random Forest performed slightly worse with 77% accuracy[18].

Preetha S et al. proposed Diabetes Disease Prediction Using Machine Learning. The goal of this study is to create a machine learning-based system for estimating the likelihood that a patient would develop diabetes. The researchers investigate a dataset of patient data to find patterns associated with diabetes using classification methods, particularly Naive Bayes and K-Nearest Neighbor. The idea behind this method is that by using historical data to train the algorithms, new patient data may be reliably classified. Data quality, feature choice, model over-fitting, and finding the right amount of sensitivity and specificity for predicting diabetes are all important factors. The achieved accuracy rate and perhaps a performance comparison between the K Nearest Neighbor and Naive Bayes algorithms[19].

Meenakshi Rajput et al. proposed Diabetes prediction and analysis using medical attributes: A Machine learning approach. The objective is to create a prediction model that can correctly forecast a person's risk of acquiring diabetes based on their medical characteristics. Using machine learning to evaluate medical data, the project seeks to aid in the early detection and prevention of diabetes. assemble

an extensive dataset with each person's medical details, such as age, BMI, blood pressure, glucose and insulin levels, family history, and other pertinent information. Create new attributes that could improve the model's ability to anticipate outcomes as well. The study shows that based on medical characteristics, machine learning models, in particular Random Forest and Neural Network, may accurately predict the chance of diabetes. The accuracy and other evaluation measures showed that the Neural Network model performed the best overall[20].

Mitushi Soni et al proposed A Review of Diabetic Prediction Using Machine Learning Techniques. The major goal is to investigate how machine learning techniques might be used to predict and categorize diabetes. The methodology used in this study uses a variety of machine learning classification methods to examine and categorize data linked to diabetes. The study makes use of methods including ensemble approaches, Naive Bayes, Support Vector Machines (SVM), and Decision Trees. The accuracy of various algorithms for predicting diabetes is examined and compared by the writers. The intricacy of the disease and the requirement for early detection are just two of the difficulties in diabetes prediction identified by the research. The study comes to the conclusion that several machine learning algorithms, such as Decision Trees, Naive Bayes, and SVM, can accurately predict diabetes to differing degrees. The survey offers perceptions into[21].

2.3 Problem Analysis

In this part, we have compared the performances and limitations of some existing Approaches.

Table 1: The table compares the performance and limitations of existing approaches -Part 01

| Author Name | Methods | Focused Area | Lackings |
|----------------------|--|---|--|
| Priyanka Indoria [2] | Artificial Neural Networks (ANNs), Naive Bayesian Networks, and Probabilistic Neural Networks (PNN) | Use machine learning methods like Artificial Neural Networks (ANNs) and Bayesian Networks (BNs) to improve the precision of identifying diseases like diabetes. | The size of the sample and the variety of the dataset should both be covered in the study. A tiny or skewed dataset may cause overfitting and have a limited impact on the generalizability of the findings. |
| N.A. Farooqui[3] | Decision Trees, K-Nearest Neighbors, Support Vector Machines, and Random Forest. | The aim is to predict diabetes at an early stage | They mentioned accuracy as an evaluation metric but they haven't discussed other metrics like - Precision, recall, and F1-score especially when dealing with imbalanced datasets |
| S.Saru[4] | Decision Trees, K-Nearest Neighbors, Naive Bayes, and Support Vector Machine | To develop an efficient predictive model to detect diabetes at an early stage by using data mining and classification algorithms to evaluate healthcare data, in particular, the Pima Indian diabetes dataset | Used only 3 classifier Decision Trees, KNN and Naïve Bayes. Doesn't specify what no of data uses just said small amount of data |
| Neha Sharma[5] | uses IoT for online data collection and remote monitoring, machine learning for diabetes analysis, fuzzy cognitive maps to help with decision making, mobile health apps for self management, and medical imaging and sensors for non-invasive glucose sensing | Develop an efficient predictive model to detect diabetes at an early stage by using data mining and classification algorithms | The security and privacy of the patient data |

Table 2: The table compares the performance and limitations of existing approaches - Part 02

| Author Name | Methods | Focused Area | Lackings |
|----------------------------|---|--|--|
| Ratna Aminah[6] | Apply image processing techniques, medical data | Obtain a wide collection of iridology photos from people, including both normal volunteers and those who have diabetes. | The study only evaluated a relatively small sample size of 27 subjects, with 11 diabetic and 16 non diabetic individuals. The paper does not mention if the developed model was externally validated using an independent dataset. |
| Mustafa S. Kadhmi[7] | Logistic Regression, Random Forest, and Support Vector Machine, precision, recall, F1-score, and ROC-AUC. | Aiming at is Select the K- mean clustering features that are most pertinent and have the biggest impact on diabetes prediction | Choosing pertinent features from the dataset and evaluating their significance might be difficult |
| Safial Islam Ayon[8] | Use a deep neural network architecture to forecast the risk of diabetes | Create a Deep Learning-based diabetes prediction system | The model's capacity for generalization may be hampered by restricted access to a broad and varied dataset |
| Jobeda Jamal Khanam[9] | Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbors, Random Forest, Naive Bayes, and Adaptive Boosting | To use data mining and machine learning approaches to predict diabetes in patients. | It neglects particular information about the preparation procedures used, such as outlier removal, handling missing values, and normalizing. A brief explanation of these processes could improve the reader's comprehension. |
| Sanskruti Patel [10] | Gaussian Naive Bayes, Random Forest, Support Vector Classifier, and Multinomial Naive Bayes | Use machine learning technique to create a predictive model for early identification of diabetes. | The paper doesn't address how this imbalance was handled during model training and evaluation, which can affect them. |
| Jingyu Xue[11] | SVM is used as a binary classification technique. The Naive Bayes classifier is used. | Performance comparison of several machine learning algorithms for diabetes prediction | The findings would be seen from a more unbiased view point if the limitations were discussed. |
| M. Rajeswari[12] | variety of datasets, including the Pima Indian Diabetes Dataset, different machine learning methods, such as Decision Trees, Naive Bayes, Support Vector Machines (SVM) | To use machine learning to predict and identify diabetes | The performance of algorithms can be affected by factors like dataset size, feature selection, and preprocessing techniques. |
| Muhammad Exell Febrian[13] | K-Nearest Neighbor (KNN) and Naive Bayes algorithms. | To use machine learning to forecast diabetes and avoid serious diseases in order to combat the rising prevalence of diabetes. | The KNN and Naive Bayes algorithms' workings aren't explained in great length in the publication. To help readers comprehend the algorithms, step-by-step explanations, Equations and diagrams should be provided. |

Table 3: The table compares the performance and limitations of existing approaches - Part 03

| Author Name | Methods | Focused Area | Lackings |
|---------------------------|---|--|---|
| Prachet Bhuyan [14] | Gaussian Naive Bayes, Random Forest, Support Vector Classifier, and Multinomial Naive Bayes | To develop a prediction model for the early detection of diabetes using machine learning techniques | The paper doesn't address how this imbalance was handled during model training and evaluation, which can affect the model's performance. |
| Amani Yahyaoui[15] | Support Vector Machine and Random Forest,DSS | Compares deep learning (Convolutional Neural Network) with more conventional machine learning (Support Vector Machine and Random Forest) | The research analyzes the models' performance using conventional classification measures (accuracy, precision, recall, etc.), but it doesn't highlight potential problems like class imbalance or offer a thorough examination of the confusion matrix. |
| S. M. Mahedy Hasan[16] | Decision Tree (DT), Random Forest (RF), and Extra Trees (ET), the Adaptive Boosting (AB) | To improve classification accuracy for diabetic patient early diagnosis by using a tree-based machine learning model | missing values, irrelevant features, and imbalanced class distribution in the dataset, |
| Muhammad Azeem Sarwar[17] | Naive Bayes, Support Vector Machine, Decision Tree, Logistic Regression, and Random Forest are the chosen algorithms. Random Forest (RF), Logistic Regression (LR), and K-Nearest Neighbors (KNN) | To compare the efficacy and precision of six different machine learning algorithms on a dataset of patient medical records | The size of the dataset and any attribute values that may be missing |
| P. Moksha Sri Sai[18] | Researchers investigate a dataset of patient data to find patterns associated with diabetes using classification methods, particularly Naive Bayes and K-Nearest Neighbor | To predict and prevent type 2 diabetes using machine learning algorithms on medical data. | Outline the study's restrictions and potential areas for further investigation. This can direct researchers who want to build on your work |
| Preetha S[19] | Investigate a dataset of patient data to find patterns associated with diabetes using classification methods, particularly Naive Bayes and K-Nearest Neighbor. | Create a machine learning-based system for estimating the likelihood that a patient would develop diabetes | Data quality, feature selection, model overfitting, and finding an optimal balance between sensitivity and specificity in diabetes prediction |
| Meenakshi Rajput[20] | health-related information about people, such as their age, body mass index (BMI), blood pressure, glucose and insulin levels, family history, and other pertinent characteristics. | To create a prediction model that can correctly forecast a person's risk of acquiring diabetes based on their medical characteristics | The current study was limited to a small sample size and cannot be generalized |

2.4 Summary

This Chapter shows the considerable advances made in the use of machine learning for diabetes prediction. The papers looked at show that various algorithms and methodologies can effectively predict diabetes risk. However, issues like data availability, model interpretability, and generalization persist. More research is needed to address these difficulties and improve the accuracy and applicability of machine learning-based diabetes prediction algorithms. As technology advances, continuing collaboration between medical practitioners and data scientists will be critical in moving this discipline forward and, ultimately, enhancing diabetes care and prevention efforts.

3 Methodology or Proposed Framework

3.1 Introduction

This chapter represents the proposed model and illustrates the feasibility analysis, requirement analysis as well as methodology where we will discuss how we collected the data that we will be using in our research which is a machine learning-based that is used to predict diabetes. Additionally, this chapter represents the procedure of the source of data and the processes of collecting data.

3.2 Proposed Framework

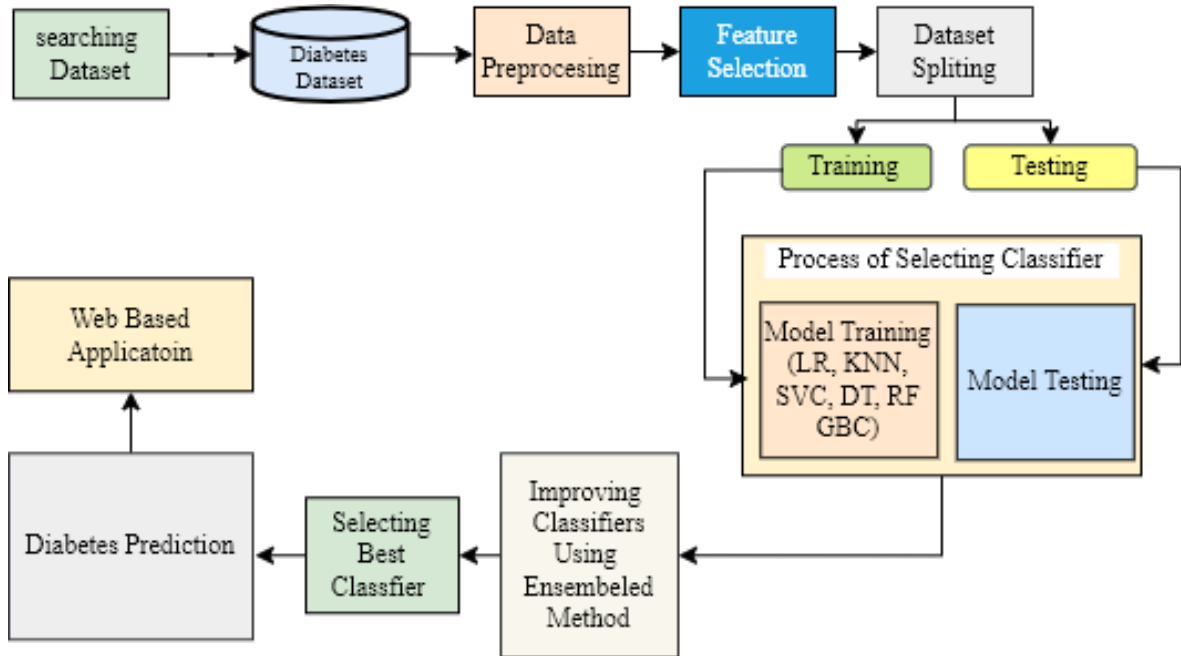


Figure 3.1: Process workflow of diabetes prediction system

3.2.1 Searching Dataset

This section describes how we identified relevant datasets for applying machine learning to predict diabetes. We chose datasets that were relevant to our research topic. To ensure that we acquired good datasets, we looked for key phrases and followed certain procedures. This procedure enabled us to obtain a large amount of usable data that we may utilize to train and evaluate our machine-learning models

for diabetes prediction. Here are several websites that search for datasets.

1. **Kaggle** - Kaggle[22] is a popular data science competition site that offers a variety of datasets, including those for diabetes prediction, as well as other resources for analysis and model creation.
2. **Google dataset search** - A Google service that allows users to search for datasets from a number of sources, including datasets potentially useful for diabetes prediction using machine learning[23].
3. **UCI Machine Learning Repository** - UCI[24] maintains a collection of datasets created exclusively for machine learning research, including diabetes prediction datasets.

3.2.2 Database Selection

The dataset we used was collected from Kaggle[22]. The datasets contain a number of medical predictor (independent) variables as well as one target (dependent) variable, Outcome. Independent variables include the patient's number of pregnancies, BMI, Diabetes pedigree function, insulin level, age, outcome, and so on. which requirements are more important than others.

3.2.3 Data Pre-processing

Data pre-processing is an important phase in the data analysis and machine learning processes because it ensures data quality, improves model performance, and prepares data for useful analysis and insights. The specific pre-processing processes to be taken are determined by the features of the data as well as the objectives of the analysis or modeling assignment

Data pre-processing is necessary for several reasons:

1. **Data Quality Improvement:** Raw data frequently contains errors, missing numbers, inconsistencies, and outliers. Pre-processing helps in identifying and correcting these errors, resulting in higher data quality and reliability.

2. **Enhanced Model Performance:** Many machine learning techniques are sensitive to data quality and distribution. Clean and well-structured data provides a solid platform for accurate predictions and insights, therefore proper pre-processing can lead to better model performance.
3. **Feature Engineering:** Data pre-processing allows for the production of new features or properties that may improve a model's prediction potential. This method entails selecting, altering, or combining existing features to create more informative data representations.
4. **Normalization and Scaling:** Different features in a dataset may have different scales, which could affect the performance of some algorithms. Normalization and scaling ensure that features have similar scales, allowing algorithms to converge faster and be more stable.
5. **Handling Missing Data:** Many real-world datasets have missing values. Data pre-processing is selecting how to handle missing values, which could include imputation (filling in missing values using statistical methods) or deleting missing value instances.

3.2.4 Describe Each Part of the Framework

Feature selection is a crucial phase in the machine learning pipeline that involves selecting the most relevant features to improve model performance, reduce over-fitting, and increase interpretability. Depending on the features of the data and the situation at hand, many methods, such as filter, wrapper, and embedding methods, can be used. Importance of feature selection:

1. **Improved Model Performance:** By focusing the model's attention on the most informative portions of the data, you can improve predicted performance by selecting only the most relevant characteristics.
2. **Reduced Over-fitting:** Reducing the number of features can assist in reducing over-fitting, which occurs when the model learns to fit noise in the data rather than capturing the underlying patterns.

3. **Faster Training:** When working with huge datasets or advanced models, having fewer features means faster training times.
4. **Enhanced Interpretability:** A model with fewer features is frequently easier to grasp and interpret, making it more useful for decision-making.

Feature Selection Methods:

Feature selection methods are widely classified into three types: filter methods, wrapper methods, and embedding methods.

1. **Filter Methods:** Filter methods are a type of feature selection strategy that evaluates the significance of features based on their own characteristics rather than the machine learning model. These methods are very handy when you need to quickly pre-process data before applying a model. Correlation analysis, mutual information, chi-squared testing, and variance thresholding are all common approaches used in filter methods. These methods assign feature scores or rankings based on established criteria, assisting in determining which aspects are most likely to contribute valuable information. Following the ranking of the characteristics, a limit is frequently used to choose the top features, rejecting those that fall below the threshold. Filter methods are efficient and can be especially useful when working with large datasets, as they allow for quick results.
2. **Wrapper Methods:** Wrapper approaches use the performance of a machine learning model as a criterion for evaluating feature subsets, which is a different approach. An algorithm repeatedly selects alternative feature combinations, trains the model on each subset, and evaluates the model's performance using techniques such as cross-validation in these methods. This iterative procedure enables wrapper methods to examine feature interactions and their impact on model performance. Wrapper approaches include forward selection, which adds features to the model one at a time, backward elimination, which removes features one at a time, and recursive feature elimination, which iteratively

removes the least significant features based on model performance. Wrapper approaches can provide more accurate feature selection, but they can also be computationally demanding.

3. **Embedded Methods:** Embedded methods embed feature selection into the model training process smoothly. Unlike filter and wrapper approaches, embedded methods use algorithms with built-in mechanisms for selecting or assigning weights to features during model training. This integration guarantees that feature selection is customized to the algorithm in use. LASSO (Least Absolute Shrinkage and Selection Operator) is a well-known embedding method that uses regularization to the model's coefficients, effectively downsizing and deleting less significant features. Decision tree-based approaches such as Random Forest and Gradient Boosting are also termed embedded methods because feature significance scores are automatically provided as part of the training process. These methods are helpful since they do feature selection while also training the model.

3.2.5 Data Splitting

Dataset splitting for training and testing involves splitting your dataset into two unique subsets: the one used for the training set and the other one used for the testing (or validation) set. The training set is used to train the machine learning model, allowing it to learn the underlying patterns and relationships in the data. The testing set, on the other hand, is kept separate and is used to evaluate the model's performance and generalization capabilities on unknown data. Using a different testing set, may check how well the model works on fresh, previously unknown cases and uncover any over-fitting or under-fitting concerns. This division ensures that the model's outcome measurements, such as accuracy or error, provide a reasonable estimate of its effectiveness when applied to real-world settings.

3.2.6 Classification of Algorithm

1. **Random Forest:** Random Forest is an ensemble approach for improving prediction by combining several decision trees. It reduces over-fitting by training trees on diverse data subsets and features. Trees "vote" for outcomes (classification) or average forecasts (regression), resulting in precise and consistent results. Randomness in feature selection improves robustness, and insights into feature relevance facilitate interpretation. While tuning prevents over-fitting, careful control of forest size and tree depth is essential. Random Forest, in essence, is a strong tool for precise and dependable machine learning across multiple domains[25].

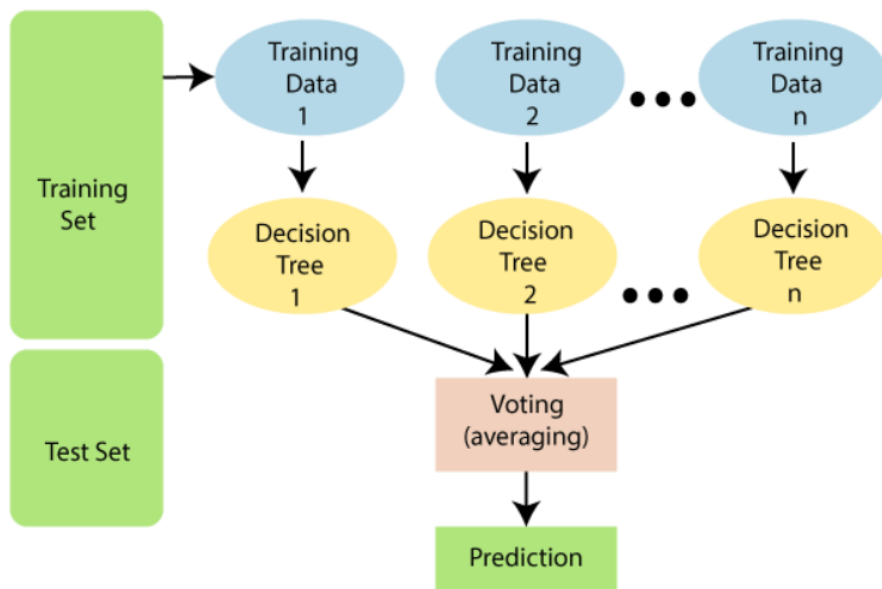


Figure 3.2: Random forest classifier architecture

2. **K-Nearest Neighbors:** K-Nearest Neighbors (KNN)[26] is a simple yet effective machine-learning technique that may be used for classification and regression problems. To create predictions, KNN considers a specified number of nearby data points, known as neighbors. When presented with a new data point, the algorithm computes the distances between it and all of the data points in the training dataset. The K data points having the shortest distances to the new point are then identified.

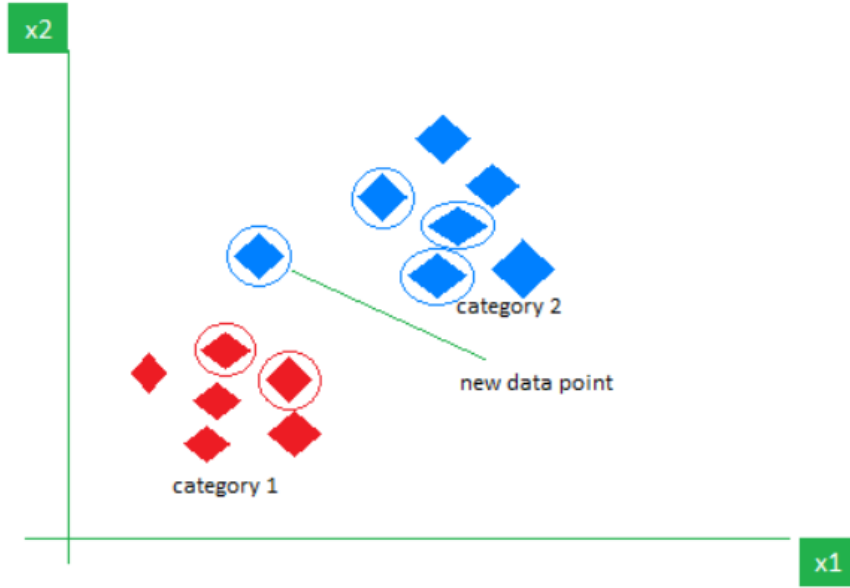


Figure 3.3: KNN architecture

The approach predicts the majority class among these neighbors for classification, while it estimates the average of the target values of the K neighbors for regression. KNN's simplicity and flexibility to diverse data distributions make it a versatile alternative, albeit careful selection of the data distribution is required.

3. **Support Vector Machine:** SVM is a strong machine-learning method that works by locating a hyperplane in a high-dimensional feature space that best separates data points of distinct classes. This hyperplane seeks to maximize the margin, which is the distance between the hyperplane and the nearest data points of each class, effectively separating them. If a linear separation is not achievable, SVM can utilize a kernel function to translate the data into a higher-dimensional space, allowing for non-linear separation. The flexibility of the decision boundary is influenced by the kernel (e.g., linear, polynomial, radial basis function). SVM looks for the best balance between maximizing margins and decreasing classification errors. It efficiently reduces complex classification issues to the task of determining the best hyperplane or decision boundary.

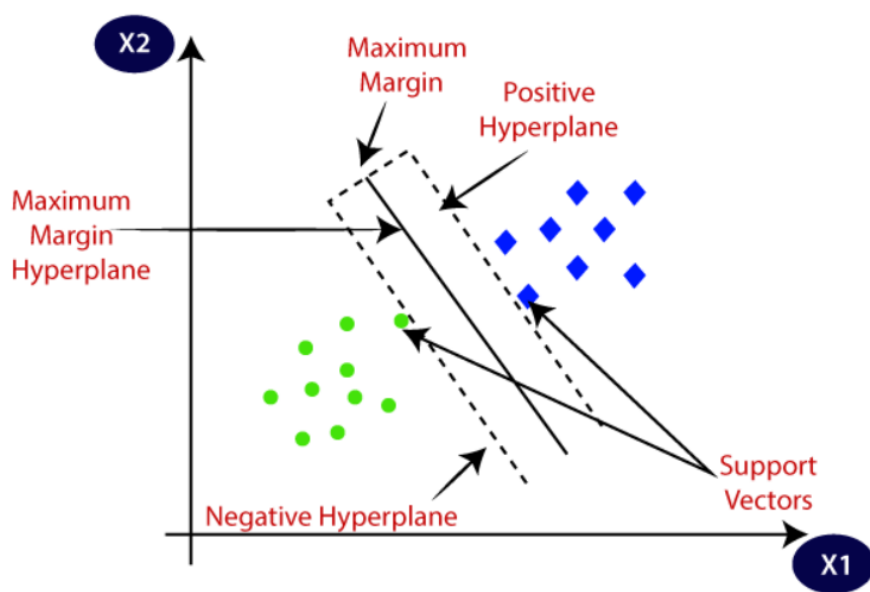


Figure 3.4: SVM Classifier architecture

The flexibility of SVM[27] to handle both linear and non-linear separations, as well as its strong theoretical background, makes it a versatile and commonly used classification technique in a variety of disciplines.

4. **Decision Tree:** A decision tree is a flexible machine-learning technique that may be used for classification as well as regression tasks. It uses a tree-like structure to model decisions and possible outcomes. The algorithm assesses features and makes binary judgments at each internal node, branching out based on feature circumstances, beginning with a root node. These judgments lead to leaf nodes, which assign final predictions or values. During training, the algorithm divides the data recursively into subgroups based on feature requirements that minimize impurity (for classification) or variance (for regression). This method generates a tree that learns patterns, relationships, and decision rules from data.

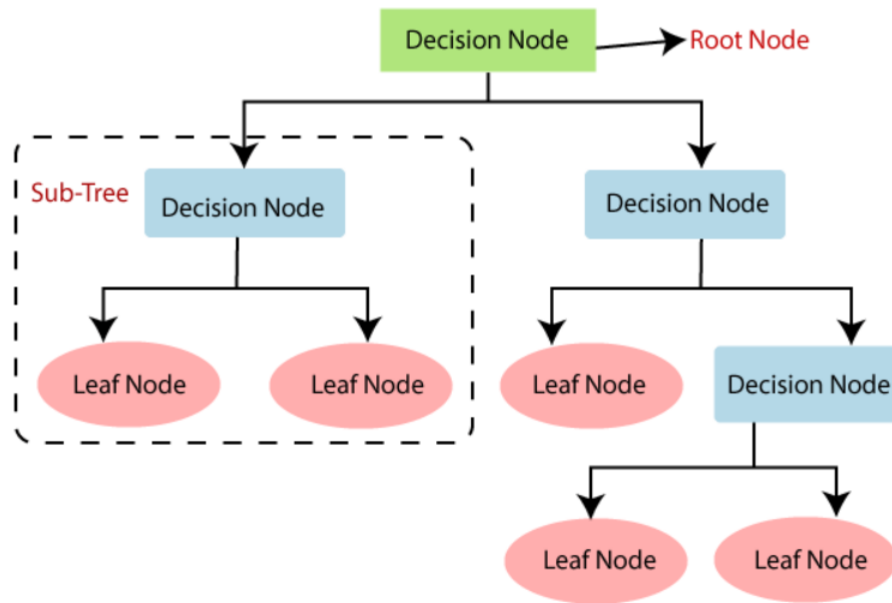


Figure 3.5: Decision Tree architecture

Decision trees[28] are simple to understand and can record complicated relationships. They are, however, prone to over-fitting and do not always generalize well. Pruning and ensemble methods (e.g., Random Forest) are frequently employed to solve these limitations and improve decision tree performance.

5. **Logistic Forest:** Logistic Regression and Linear Regression are quite similar. Linear regression is used to solve regression problems, whereas logistic regression is used to address classification problems. In logistic regression, the categorical dependent variable is predicted using a collection of independent variables. It predicts the output of a categorical dependent variable. As a result, the output must be either categorical or discrete. It can be Yes or No, 0 or 1, True or False, and so on, but instead of displaying exact values like 0 and 1, it displays probability values between 0 and 1.

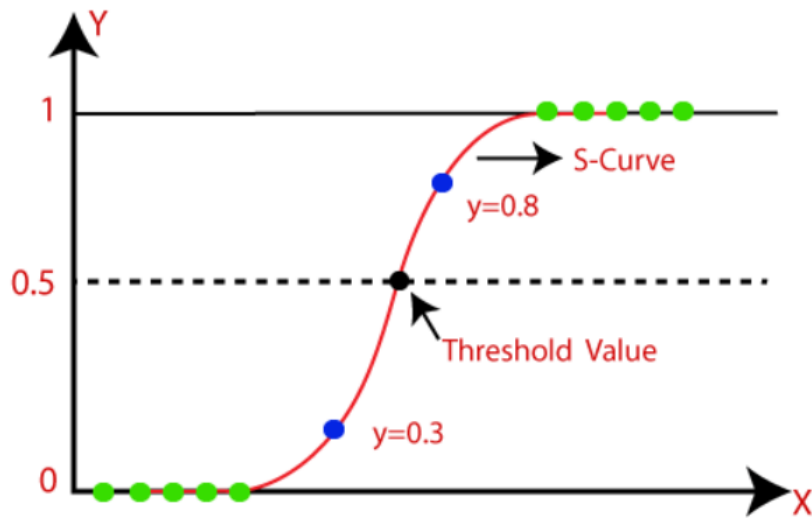


Figure 3.6: Logistic Regression architecture

Instead of generating a regression line, we fit a logistic function with an "S" shape[29] that predicts one of two maximum values (0 or 1). The logistic function curve indicates the possibilities of something like whether or not the cells are malignant, whether or not a mouse is fat based on its weight, and so on.

6. **Gradient Boosting Classifier:** Gradient Boosting Classifier is a powerful classification machine learning technique. It builds an ensemble model by sequentially joining numerous weak learners, often decision trees. The method focuses on minimizing prediction errors by learning from prior models' faults. It begins with an initial model and then adds new models successively, with each new model attempting to rectify the faults made by the ensemble up to that point. This is accomplished by modifying the weights of data points and maximizing the predictions of the new model. The final prediction is a weighted average of each model's projections.

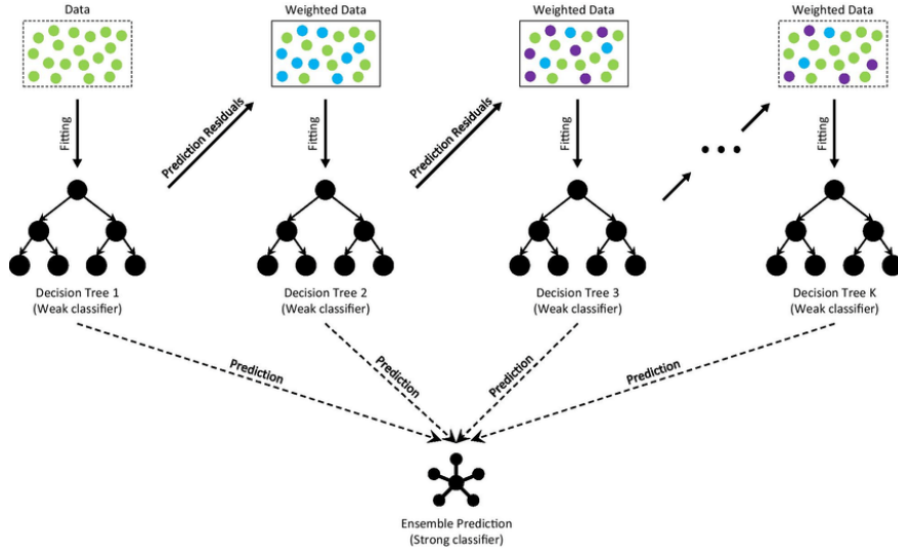


Figure 3.7: Gradient Boosting Classifier architecture

Gradient Boosting Classifier[30] is well-known for its ability to grasp complicated associations and effectively handle unstructured data. It is, however, susceptible to over-fitting and necessitates meticulous adjustment of hyper-parameters. Overall, the Gradient Boosting Classifier is a well-known and effective method for generating high predicted accuracy in a variety of machine learning tasks.

3.2.7 Selecting Classifier

The first step in predicting diabetes is to train multiple machine learning models on a labeled dataset, including Logistic Regression (LR), Random Forest (RF), Gradient Boosting Classifier (GBC), Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), and Decision Tree (DT). Each model discovers patterns and relationships in the data, allowing it to make precise predictions. The trained models are then evaluated against a different testing dataset to determine their individual predicting ability. To find the most promising models, we evaluate their accuracy on test data. For further examination, the top four models with the highest accuracy are chosen. These models have a high potential for accurate diabetes prediction and might be used in an ensemble approach.

The next stage focuses on improving the performance of the selected classifiers using

an ensemble technique. This involves combining the previous four classifier predictions to generate a more robust and accurate final model. The ensemble technique takes advantage of the strengths of various models while mitigating individual limitations. We evaluate the improved classifiers' accuracy on the testing data after implementing the ensemble approach and training them. This stage helps us to discover which of the selected models' upgraded classifiers has the highest predictive accuracy. We ensure the discovery and deployment of the best-performing classifier for diabetes prediction by applying this full process of training, testing, selection, and ensemble augmentation. This ultimately contributes to optimal healthcare decision-making. Also, We save the model for future use so that we don't have to repeat the process. Simply call the model and provide the input to receive the prediction.

3.2.8 User-friendly UI

We use Tkinter, a Python library that allows us to create graphical user interfaces (GUIs). While it is not generally used for web development, it may be used to create simple desktop apps with buttons, input fields, labels, and other features. It's most commonly used for developing standalone programs that operate on a user's PC rather than on the web. By providing specific Values in the input fields anyone can find whether a patient is diabetic or not. The fields are - Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and Outcome. If the Outcome is 0 it means the patient is non-diabetic and if 1 then the patient is diabetic.

3.3 Summary

In summary, there are several crucial milestones in our path to forecast diabetes using machine learning. It all starts with thorough dataset selection, followed by careful data treatment to assure correctness and readiness. Feature selection improves the model's capacity to capture relevant patterns by refining the dataset. The evaluation and selection of several categorization algorithms result in a strong prediction toolset, with each method responding to distinct data difficulties. The

critical phase of selecting the best classifiers is led by predicted accuracy, which leads to the selection of top-performing models. Further refinement using ensemble methods taps into these models' collective strength, resulting in a more dependable and potent forecasting tool. Using Tkinter a Python library helps to create a user-friendly GUI where users can cross-check their results.

4 Implementation and Testing

4.1 Introduction

In this section, we will provide a complete overview of the system. In this section we will highlight everything from preprocessing to the n-symbol method.

4.2 System Setup

We are using Google Colab for coding in this section. Google Colab is a free, cloud-based Python coding platform. Using Google Colab we have done cloud-based Python coding, machine learning, data analysis, and collaborative editing. In this section, we are using Django to build a website quickly and easily. It simplifies tasks handling databases, user authentication, and creating secure web applications. Django is a user-friendly framework for web development. In this project we also use PyCharm. PyCharm is an Integrated Development Environment for Python and tools for Django. we using PyCharm for efficient python development with features like intelligent code completion and integrated debugging. It's a tool for refactoring, testing, and Database management.

4.3 Evaluation

The Evaluation of the diabetes prediction system's performance uses historical data, as well as the effectiveness of the integrated machine learning algorithms, which include six classifiers those are - Random Forest (RF), Logistic Regression (LR), Gradient Boosting Classifier (GBC), Support Vector Machines (SVM), k-nearest Neighbors (KNN), and Decision Tree Classifier (DT). Our primary goal is to build a hybrid model that combines these classifiers and selects the best one through voting (hard and soft) and enhancing their performance.

4.4 Result and Discussion

4.4.1 Dataset:

The dataset contains a total number of 767 rows and 9 columns including - pregnancies, glucose, insulin, and so on. Have a look at the table-

Table 4: Diabetes Dataset

| Index | Preg. | Glucose | BP | SkinThick. | Insulin | BMI | DP Func. | Age | Outcome |
|-------|-------|---------|----|------------|---------|------|----------|-----|---------|
| 0 | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 2 | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 3 | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . |
| 764 | 2 | 122 | 70 | 27 | 0 | 36.8 | 0.34 | 27 | 0 |
| 765 | 5 | 121 | 72 | 23 | 112 | 26.2 | 0.245 | 30 | 0 |
| 766 | 1 | 126 | 60 | 0 | 0 | 30.1 | 0.349 | 47 | 1 |
| 767 | 1 | 93 | 70 | 31 | 0 | 30.4 | 0.315 | 23 | 0 |

4.4.2 Data Pre-processing:

Checking missing value and outlier if available then remove.

- Missing Value

Table 5: Missing Values Summary

| Feature | Missing Values |
|---------------|----------------|
| Pregnancies | 0 |
| Glucose | 0 |
| BloodPressure | 0 |
| SkinThickness | 0 |
| BMI | 0 |
| Age | 0 |
| Outcome | 0 |

- Outlier: Checking Outlier available or not -

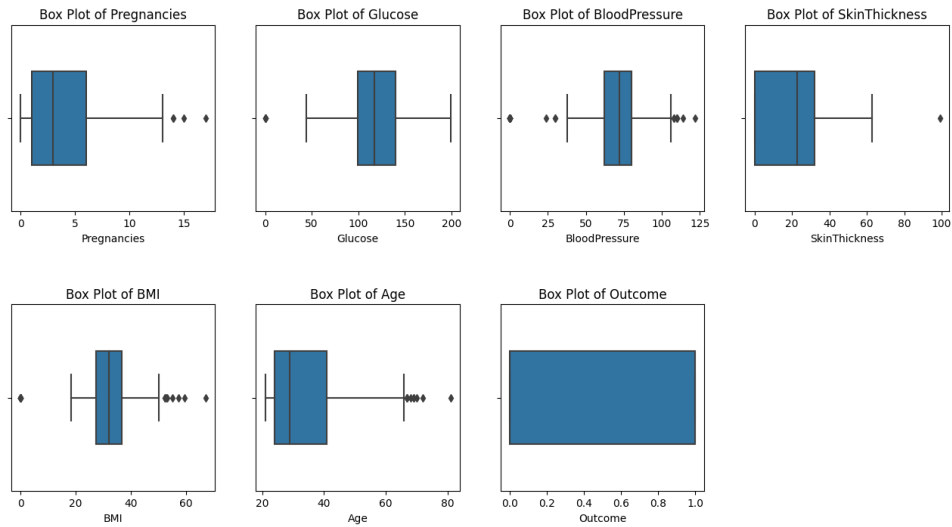


Figure 4.8: Outlier Check with Boxplot

We can see that the box plot outlier is available. This will impact statistical analyses and ML models. Some potential consequences are - the risk of overfitting, model robustness, data visualization, descriptive statistics, etc. To mitigate this impact various approaches can be employed. One of them is the Interquartile Range (IQR) method.

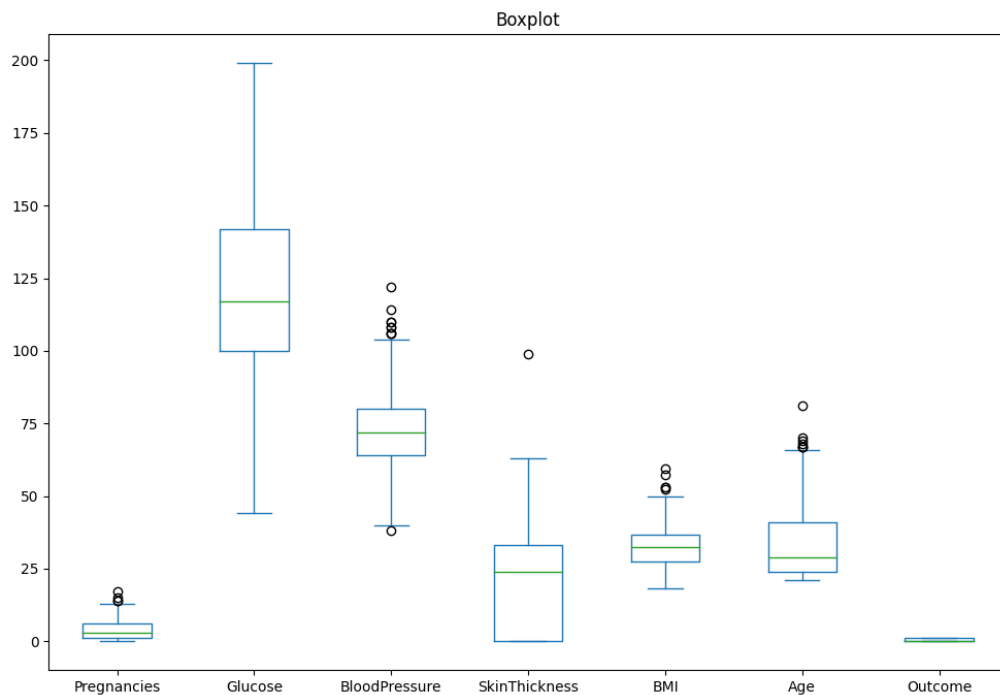


Figure 4.9: After removing outlier

- Correlation Heatmap

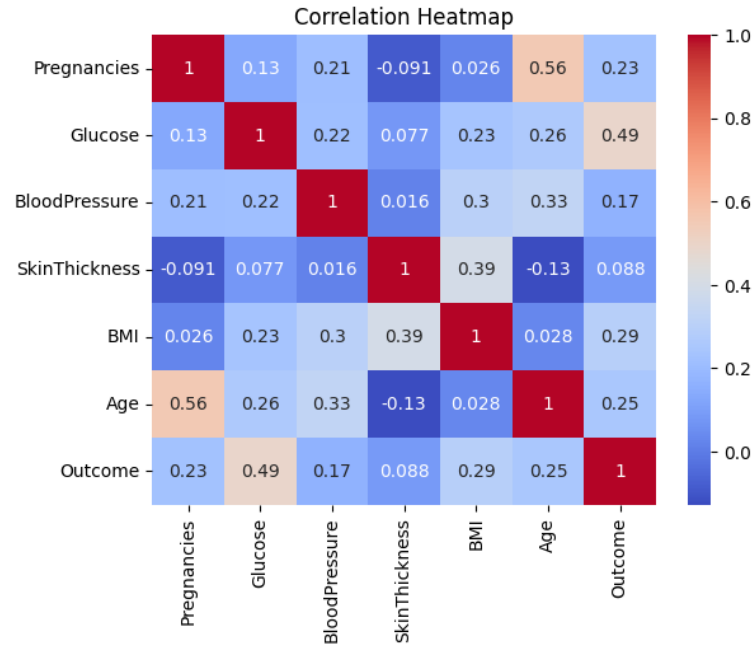


Figure 4.10: Correlation Heatmap

4.4.3 Training Model & Accuracy Score:

We split the data into two parts for training and testing where the ratio is 20:80. We have chosen six classifiers (RF, LR, GBC, SVC, KNN, DT) for model training and getting accuracy for each classifier.

Table 6: Classification Accuracy of Different Models

| Model | Accuracy (%) |
|---------------------|--------------|
| Logistic Regression | 74.31 |
| Random Forest | 77.78 |
| Decision Tree | 65.28 |
| KNeighbors | 75.00 |
| SVC | 75.00 |
| Gradient Boosting | 77.78 |

4.4.4 Ensemble Method:

we have done both hard and soft voting to enhance the overall performance in hard voting accuracy score of 74% and soft voting accuracy score of 80%. Here is the confusion matrix and classification report-

1. Confusion Matrix

$$\begin{bmatrix} 87 & 9 \\ 20 & 28 \end{bmatrix}$$

2. Classification Report

Table 7: Classification Report

| Class | Precision | Recall | F1-Score | Support |
|---------------------|-----------|--------|----------|---------|
| 0 | 0.81 | 0.91 | 0.86 | 96 |
| 1 | 0.76 | 0.58 | 0.66 | 48 |
| Accuracy | | | 0.80 | 144 |
| Macro Avg | 0.78 | 0.74 | 0.76 | 144 |
| Weighted Avg | 0.79 | 0.80 | 0.79 | 144 |

3. ROC Curve

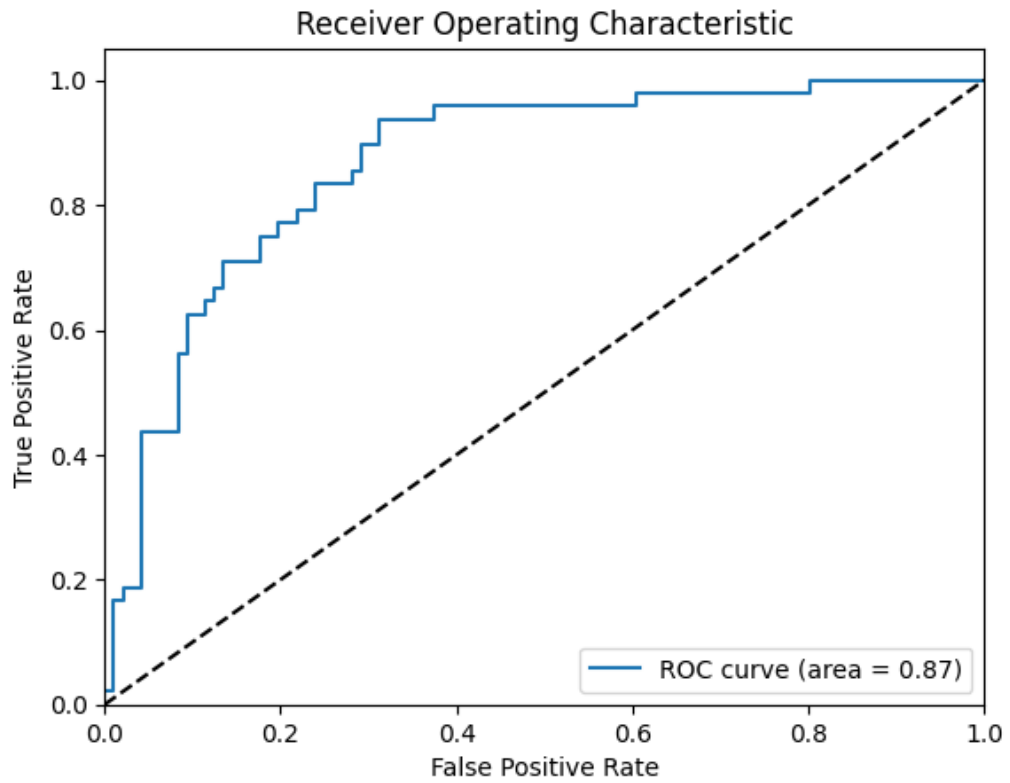
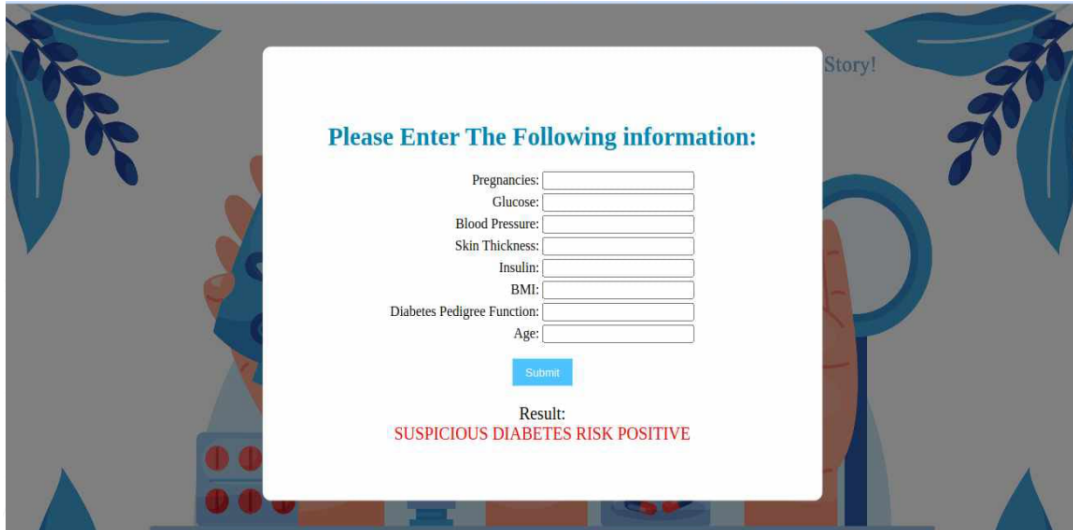


Figure 4.11: ROC Curve

4.4.5 Web Interface:



Please Enter The Following information:

Pregnancies:

Glucose:

Blood Pressure:

Skin Thickness:

Insulin:

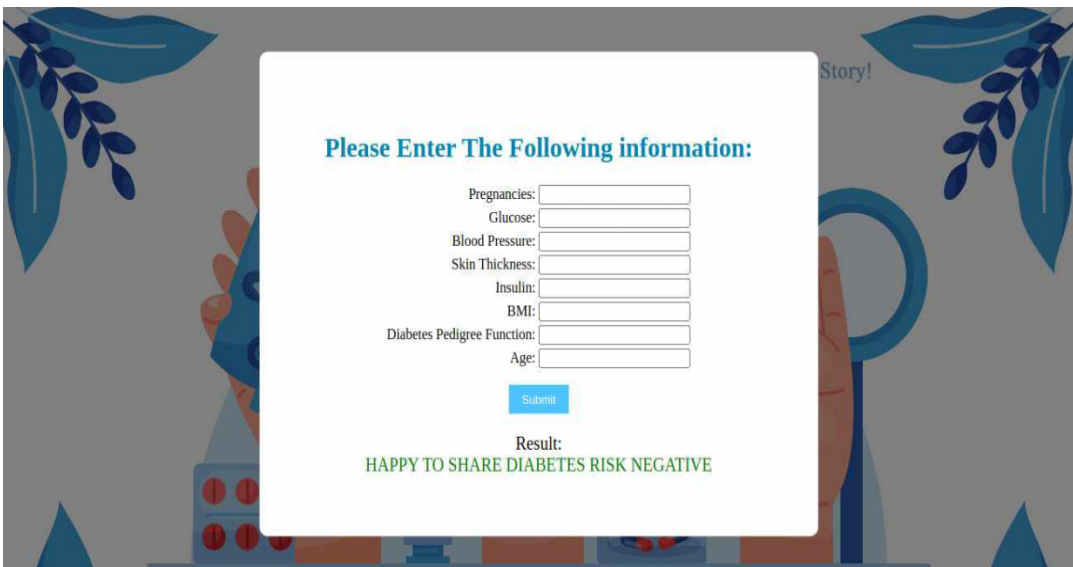
BMI:

Diabetes Pedigree Function:

Age:

Result:
SUSPICIOUS DIABETES RISK POSITIVE

Figure 4.12: Web App Interface 01



Please Enter The Following information:

Pregnancies:

Glucose:

Blood Pressure:

Skin Thickness:

Insulin:

BMI:

Diabetes Pedigree Function:

Age:

Result:
HAPPY TO SHARE DIABETES RISK NEGATIVE

Figure 4.13: Web App Interface 02

4.5 Summary

In this Implementation and Testing Chapter, we are working from pre-processing to the ensemble method, system setup, system evaluation, and system output/result and overall we have discussed everything.

5 Standards, Constraints, and Milestones

5.1 Introduction

The objective of this project is to give accessible and trustworthy diabetes risk forecasts worldwide by developing prediction algorithms for early diagnosis and prevention. Using machine learning methods, a large-scale diabetic dataset from Kaggle is carefully pre-processed to guarantee data quality for precise analysis. By using feature selection approaches, the project overcomes the shortcomings of current systems and assesses a variety of classification algorithms, including LR, RF, GBC, SVC, KNN, and DT, in order to make machine learning more approachable. By incorporating the most accurate model into an approachable online application, the goal is to democratize the advantages of machine learning and enable people to put their health first in the fight against diabetes. The next sections will go over the criteria, limitations, and project milestones that were reached.

5.2 Standards

Standards for the Project:

1. Ethical Guidelines:

- To protect user privacy and confidentiality, make sure that ethical standards are followed in the gathering, processing, and use of medical data.
- Respect accepted moral standards when creating and implementing machine learning algorithms to avoid prejudice and discrimination.

2. Data Quality Standards:

- Apply stringent data pre-processing methods to guarantee the dependability and correctness of the diabetic dataset that was obtained from Kaggle.
- Maintain and update the dataset regularly to reflect the most current and pertinent diabetes-related facts.

3. Model Performance Standards:

- Comply with industry standards for model performance when evaluating and

choosing classification approaches (LR, RF, GBC, SVC, KNN, DT) based on their accuracy on test data.

- To guarantee accurate diabetes risk estimates, set a threshold for the prediction engine's acceptable levels of sensitivity, specificity, and accuracy.

4. Ensemble Technique Standards:

- Put into practice and perfect ensemble approach to combine forecasts from several models, making sure the combination improves prediction accuracy and reliability.
- Regularly review and adapt group methods in light of new industry best practices.

5. User Interface Standards:

- Create and manage an accessible web application with an easy-to-use interface to make it accessible to users with different levels of technical expertise.
- Perform user testing to get input and keep the user experience getting better.

6. Security Standards:

- Adhere to industry standards for data encryption and secure transfer and put strong security measures in place to safeguard user-entered medical information.

7. Accessibility Standards:

- To encourage inclusivity, make sure the web application complies with accessibility standards (such as WCAG) and is usable by people of various abilities.

6 Conclusion

6.1 Introduction

The effectiveness of ensemble approaches in enhancing prediction reliability and accuracy is the highlighted part. The built web-based application provides users with an easy-to-use interface to cross-validate. This initiative adds to continuing global efforts in diabetes early detection and intervention by allowing individuals to prioritize their health and well-being through the widespread availability of machine learning and analytical analytics on a user-friendly web platform. The success of this project demonstrates the power of sophisticated technologies to improve public health outcomes.

6.2 Conclusion

The development and execution of prediction algorithms for early detection of diabetes and preventative measures represent an enormous breakthrough in healthcare. The major goal was to deliver reliable diabetes risk estimates around the world, allowing for early intervention and informed decision-making. The research process began with the selection and processing of a large diabetic dataset from Kaggle. To assure data quality, extreme data pre-processing processes were used, setting the groundwork for robust analysis. Feature selection approaches were used to improve the model's performance and interpretability by finding the most informative variables. An in-depth review of various classification algorithms was carried out, including Logistic Regression (LR), Random Forest (RF), Gradient Boosting Classifier (GBC), Support Vector Classifier (SVC), k-nearest Neighbors (KNN), and Decision Tree (DT). The dataset was divided into training and testing sets, allowing the models to be trained and evaluated on the test data. Among these classifiers, the best-performing models were determined. For better improvement, the Ensemble Method is used with different types of voting.

6.3 Limitation and Future Works

6.3.1 Limitation

The issue is that there's an accuracy score fluctuation. This makes it sometimes tricky to pick the best one. Also right now our web app is only console-based.

6.3.2 Future Works

- Design a Mobile Application so that people can access it from multiple devices.
- contact hospitals and doctors and add them to the board of this project.
- will try to shift to cloud services for any uninterrupted

References

- [1] Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., & Malanda, B. I. D. F. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes research and clinical practice*, 138, 271-281.
- [2] Indoria, P., & Rathore, Y. K. (2018). A survey: detection and prediction of diabetes using machine learning techniques. *International Journal of Engineering Research & Technology (IJERT)*, 7(3), 287-291.
- [3] Farooqui, N., Mehra, R., & Tyagi, A. (2018). Prediction model for diabetes mellitus using machine learning techniques. *Int. J. Comput. Sci. Eng*, 6(3).
- [4] Saru, S., & Subashree, S. (2019). Analysis and prediction of diabetes using machine learning. *International journal of emerging technology and innovative engineering*, 5(4).
- [5] Sharma, N., & Singh, A. (2019). Diabetes detection and prediction using machine learning/IoT: A survey. In *Advanced Informatics for Computing Research: Second International Conference, ICAICR 2018, Shimla, India, July 14–15, 2018, Revised Selected Papers, Part I 2* (pp. 471-479). Springer Singapore.
- [6] Aminah, R., & Saputro, A. H. (2019, September). Diabetes prediction system based on iridology using machine learning. In *2019 6th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)* (pp. 1-6). IEEE.
- [7] Kadhm, M. S., Ghindawi, I. W., & Mhawi, D. E. (2018). An accurate diabetes prediction system based on K-means clustering and proposed classification approach. *International Journal of Applied Engineering Research*, 13(6), 4038-4041.

- [8] Ayon, S. I., & Islam, M. M. (2019). Diabetes prediction: a deep learning approach. *International Journal of Information Engineering and Electronic Business*, 12(2), 21.
- [9] Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *Ict Express*, 7(4), 432-439.
- [10] Patel, S. (2021). Predicting a risk of diabetes at early stage using machine learning approach. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(10), 5277-5284.
- [11] Xue, J., Min, F., & Ma, F. (2020, November). Research on diabetes prediction method based on machine learning. In *Journal of Physics: Conference Series* (Vol. 1684, No. 1, p. 012062). IOP Publishing.
- [12] Rajeswari, M., & Prabhu, P. (2019). A review of diabetic prediction using machine learning techniques. *International Journal of Engineering and Techniques*, 5(4), 2395-1303.
- [13] Febrian, M. E., Ferdinan, F. X., Sendani, G. P., Suryanigrum, K. M., & Yunnanda, R. (2023). Diabetes prediction using supervised machine learning. *Procedia Computer Science*, 216, 21-30.
- [14] Sahoo, P., & Bhuyan, P. (2021). Primitive Diabetes Prediction using Machine Learning Models: An Empirical Investigation. *Turkish Journal of Computer and Mathematics Education*, 12(11), 229-236.
- [15] Yahyaoui, A., Jamil, A., Rasheed, J., & Yesiltepe, M. (2019, November). A decision support system for diabetes prediction using machine learning and deep learning techniques. In *2019 1st International informatics and software engineering conference (UBMYK)* (pp. 1-4). IEEE.
- [16] Hasan, S. M., Rabbi, M. F., Champa, A. I., & Zaman, M. A. (2020, November). An Effective Diabetes Prediction System Using Machine Learning Techniques. In *2020 2nd International Conference on Advanced Information and Communication*

Technology (ICAICT) (pp. 23-28). IEEE.

[17] Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018, September). Prediction of diabetes using machine learning algorithms in healthcare. In 2018 24th international conference on automation and computing (ICAC) (pp. 1-6). IEEE.

[18] Sai, P. M. S., & Anuradha, G. (2020, March). Survey on Type 2 diabetes prediction using machine learning. In 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC) (pp. 770-775). IEEE.

[19] Preetha, S., Chandan, N., Darshan, N. K., & Gowrav, P. B. (2020). Diabetes disease prediction using machine learning. *Int. J. Mod. Trends Eng. Res*, 6, 37-43.

[20] Rajput, M. R., & Khedgikar, S. S. (2022). Diabetes prediction and analysis using medical attributes: A Machine learning approach. *Journal of Xi'an University of Architecture & Technology*, 14(1), 98-103.

[21] Soni, M., & Varma, S. (2020). Diabetes prediction using machine learning techniques. *International Journal of Engineering Research & Technology (Ijert)* Volume,9.

[22] Kaggle. (n.d.). Retrieved July 21, 2023, from <https://www.kaggle.com>

[23] Google Dataset Search. (n.d.). Retrieved July 21, 2023, from <https://dataset-search.research.google.com>

[24] UCI Machine Learning Repository. (n.d.). Retrieved July 21, 2023, from <https://archive.ics.uci.edu>

[25] Random Forest Classifier. (n.d.). Retrieved August 1, 2023, from <https://www.javatpoint.com/machine-learning-random-forest-algorithm>

[26] KNN. (n.d.). Retrieved August 1, 2023, from <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning>

- [27] SVM Classifier. (n.d.). Retrieved August 1, 2023, from [https://www.javatpoint.com /machine-learning-support-vector-machine-algorithm](https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm)
- [28] Decision Tree Classifier. (n.d.). Retrieved August 1, 2023, from <https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm>
- [29] Logistic Regression Classifier. (n.d.). Retrieved August 1, 2023, from <https://www.javatpoint.com/logistic-regression-in-machine-learning>
- [30] GBC Classifier. (n.d.). Retrieved August 1, 2023, from <https://www.researchgate.net/figure/Architecture-of-a-Logistic-Regression-Model>