LEVERAGING MACHINE LEARNING FOR DIABETES PREDICTION: ENSEMBLE MODEL

MCDONALD OTIENO OGUTU

A RESEARCH PROJECT SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY IN THE SCHOOL OF COMPUTING AND MATHEMATICS IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF THE DEGREE OF MASTER OF SCIENCE IN DATA SCIENCE OF THE COOPERATIVE UNIVERSITY OF KENYA

2025

DECLARATION

Declaration by the candidate

This proposal/thesis is my original work and has not been presented for the award of a degree in any other University or for any other award

24/11/2025

Signature Date

McDonald Otieno Ogutu – MDATC01/6052/2022

Declaration by the supervisors

I/We confirm that the work reported in this proposal/thesis was carried out by the candidate under our supervision and has been submitted with our approval as university supervisors

24/11/2025

Signature Date

Prof. Simon Karume

Department of Computer Science & Information Technology, The Cooperative University of Kenya

24/11/2025

Signature Date

Dr. Benson Kituku

Department of Computer Science, Dedan Kimathi University of Technology

DEDICATION

I dedicate this project to my family and mentors, whose unwavering support and encouragement have been my greatest source of strength throughout this journey.

ACKNOWLEDGEMENT

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

First and foremost, I extend my heartfelt appreciation to my supervisors, Dr. Kituku and Prof. Karume, for their invaluable guidance, continuous support, and constructive feedback throughout the course of this study.

I am also deeply thankful to the faculty and staff of Computer Science & Information Technology Department, whose expertise and resources provided a solid foundation for my research.

Special thanks to my colleagues and friends for their encouragement, insightful discussions, and moral support during challenging moments.

Lastly, I am profoundly grateful to my family for their patience, understanding, and unwavering belief in me throughout this academic journey.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

RF - Random Forest

GB - Gradient Boosting

SVM - Support Vector Machine

ML - Machine Learning

DT - Decision tree

RMSE - Root Mean squared error

MAE - Mean Absolute Error

OPERATIONAL DEFINITIONS

Machine Learning: Artificial intelligence branch that deals with algorithms and statistical methods to make forecasts.

Random Forest: This is an ensemble learning method; it constructs numerous decision trees and groups their forecasts to enhance forecasting accuracy.

Gradient Boosting Regression (GBR): This ML method sequentially builds models, in which case, every new model corrects errors from the previous ones.

Support Vector Machines (SVM): This is a classification and regression method that finds the optimal hyperplane to predict continuous values or segregate data into various classes.

Training Dataset: A collection of data used to teach a machine learning model to recognize patterns and make predictions.

Testing Dataset: A distinct set of data used to assess the performance and generalizability of a trained machine learning model.

ABSTRACT

Diabetes presents a great global health challenge, with delayed diagnosis significantly impeding effective management, particularly in resource-constrained regions. The critical shortage of medical professionals in regions like Kenya with a doctor-to-population ratio far below the WHO standard severely hampers timely screening and diagnosis diabetes. This deficit necessitates innovative, scalable tools, such as machine learning models, to assist in early prediction and intervention.This ~~project~~ research aimed to enhance timely and accurate diabetes prediction by developing an advanced ensemble machine learning model. Ahybrid dataset, compiled from the PIMA Indian (762 instances) and Hospital Frankfurt Germany (2000 instances) datasets, totaling 2762 datapoints, was utilized to improve generalizability beyond single-source limitations. The research employed a quantitative design which involved comprehensive data preprocessing, including the critical imputation of physiologically impossible zero values and feature standardization. After assessing multicollinearity, all independent variables were retained. Six machine learning algorithms; Logistic Regression, Decision Tree, K-Nearest Neighbors, Support Vector Machine, Random Forest, and XGBoostwere evaluated, undergoing hyperparameter tuning to optimize their performance. XGBoost and Random Forest consistently achieved the highest F1-scores (0.9974 and 0.9947 respectively) among individual classifiers. These two top-performing models were then selected as base learners for a StackingClassifier ensemble, which utilized a Logistic Regression meta-learner. The developed ensemble model demonstrated exceptional predictive capabilities, achieving an F1-score of 0.9974 and a near-perfect ROC-AUC of 0.9999. This performance matched XGBoost's F1-score and marginally surpassed its ROC-AUC. Implemented in Python, this

research underscores the significant potential of advanced ensemble machine learning to deliver highly accurate and robust diagnostic solutions, thereby contributing to earlier diabetes detection and improved health outcomes, particularly in underserved healthcare environments.

Key words: machine learning, support vector machine, gradient boosting, Random Forest, accuracy

1. CHAPTER ONE

1.0 Introduction

Diabetes is a disease affecting many people globally, causing serious health problems (World Health Organization, 2020) hence the need for early detection and treatmentTo respond to this, in the recent decade, data science has come up with powerful machine learning tools in the healthcare sector, providing innovative disease prediction and management solutions (Kandhare et al., 2025). This utilization of data science aligns with a broader public health requirement tied to the United Nations Sustainable Development Goals (UN SDGs), specifically Goal 3: Good Health and Well-being. Achieving this goal requires reducing premature mortality from non-communicable diseases (NCDs), including diabetes, by one-third by 2030 (United Nations, 2015). When diabetes is detected early and accurately, timely treatment and management can be implemented, significantly lowering the risk of severe, often fatal complications (such as cardiovascular disease, kidney failure, and stroke) and reducing associated healthcare costs. By improving diagnostic precision and accessibility, we directly contribute to this global mission, paving the way for a healthier, more productive population and lessening the immense social and economic burden caused by chronic disease.

1.1. Background of the study

Diabetes ranks among the top prevalent diseases globally. The World Health Organization (WHO, 2021) asserts that this condition's prevalence among adults of over 18 years is 8.5% and has caused 6.7 million deaths worldwide in 2021. The disease accounts for a substantial portion of premature deaths, alongside cardiovascular conditions, cancer, and respiratory diseases. Despite a decline in diabetes-related deaths from 2000 to 2010,

statistics show a resurgence between 2010 and 2016, with mortality rates expected to increase further (World Health Organization, 2020). The disease also comes along with huge health cost implications. In 2021, the disease caused global health spending of at least US$960 billion. In Sub-Saharan Africa, the burden of diabetes is expected to impact significantly, with its prevalence anticipated to increase 2.5 times between 2021 and 2045. The spending on health related to diabetes is expected to go up from a total of 12.6 billion USD to 46.7 billion USD (IDF, 2021). Since 2015, diabetes has been treated with a lot of concern in Kenya. The disease burden in Kenya has been exacerbated by late diagnosis (Manyara et al., 2024). A national survey conducted in 2015 revealed that 51% of diabetes cases in urban areas had not been diagnosed. Similarly, a cross-sectional study conducted by Manyara (et al., 2024) on some 50 patients in Nairobi revealed that 52% were diabetic but had not been diagnosed. This situation is being worsened by the fact that Kenya has a critical shortage of medical professionals, with only 1 doctor available for every 5,263 people according to the Kenya National Bureau of Statistics (KNBS, 2024)This is far below the WHO-recommended ratio of 1:1,000. This shortage severely hampers timely diagnosis and management of chronic diseases such as diabetes. As a result, many cases remain undetected until complications arise, contributing to increased morbidity and preventable mortality. The gap in early screening and diagnosis is especially alarming given the rising burden of diabetes in the country. There is need for innovative, scalable tools such as machine learning models that can help in early detection and intervention, especially in underserved communities where access to qualified healthcare providers is limited.

In healthcare, machine learning techniques can be employed to aid in detecting diseases early enough hence their treatment. The techniques analyze medical dataset to predict results hence lowering the costs of identifying complex diseases (Gadekallu et al., 2020). By incorporating machine learning approaches, researchers and studies have succeeded in building machine learning models that are able to detect diabetes early enough, hence avoiding severe effects of the disease and ensuring early medication. Despite this breakthrough, a good percentage of these models are single classifiers which are prone to overfitting in cases of small datasets and this results into poor generalizability (Alnagashi et al., 2024). Another limiting factor with single classifiers is that they are affected by noise and outliers, struggle when it comes to bias-variance tradeoff and they are not robust enough to capture complex patterns. In a bid to improve the predictive capabilities of single classifier machine learning models in diabetes, recent studies have gone the ensemble way (Mienye & Sun, 2022). Compared to single machine learning models, ensemble models do better on accuracy, flexibility and high predictive capabilities. However, majority of the studies and researchers who have developed ensemble models classifiers to predict the risk of the disease have mostly used PIMA Indian dataset (Mousa, Mustafa, & Marqas, 2023). This dataset comes with limitations such as the size; it only has 768 instances. Secondly, it represents a given ethnic group thus devoid of diversity hence not able to consider the different genetic predispositions to diabetes present in other population characteristics like the Europeans and Africans. To add on, using PIMA dataset excludes considerations such as environmental factors and lifestyle which largely influences the risk of diabetes. Lastly, because of the nature of this dataset, it limits generalizability and adoption in healthcare globally.

It is on this basis that this project sought to develop from a hybrid of datasets (PIMA (768 datapoints) and Hospital Frankfurt Germany (2000 datapoints)), an accurate ensemble machine learning algorithm from single machine learning models to enhance diabetes prediction. Two best performing models from Random Forest, Support Vector Machines, XGBoost, and k-nearest neighbor were used to develop the ensemble model.

1.2. Statement of the Problem

Despite advances in medical technology, early detection of diabetes remains a significant challenge, particularly in underserved communities (Ebekozien et al., 2024). A considerable number of people with the potential of developing diabetes mellitus are not diagnosed on time. According to (Musau et al., 2020), one of the major causes of high diabetes burden in Kenya is delayed diabetes diagnosis. This has necessitated employment of machine learning approaches for early detection of diabetes. However, most ML models that have been developed to predict diabetes have been single machine learning classifiers which come with numerous weaknesses. They often overfit, especially with complex models and small datasets, leading to poor generalization. They struggle with the bias-variance tradeoff, are sensitive to noise and outliers, and may have limited capacity to capture complex patterns among others. In bid to improve predictive capabilities of single classifier machine learning models in diabetes, recent studies have gone the ensemble way. Compared to single machine learning models, ensemble models do better on accuracy, flexibility and high predictive capabilities. However, majority of the studies and researchers who have developed ensemble models to predict diabetes have mostly used PIMA Indian dataset. This dataset comes with limitations such as the size; it only has 768 instances. Secondly, it represents a given ethnic group thus devoid of diversity hence not

able to take into account the different genetic predispositions to diabetes present in other population characteristics like the Europeans and Africans. To add on, using PIMA dataset excludes considerations such as environmental factors and lifestyle which largely has an effect on the risk of diabetes. Lastly, because of the nature of this dataset, it limits generalizability and adoption in healthcare globally.

It is on this basis that this project sought to develop from a hybrid of datasets (PIMA (768 datapoints[1] and Hospital Frankfurt Germany (2000 datapoints[2], an accurate ensemble machine learning algorithm from single machine learning models to enhance diabetes prediction.

1.3. Study Objectives

1.3.1. General objective

This project's primary goal is to develop an ensemble machine-learning model for predicting diabetes.

1.3.2. Specific objectives

a. To review existing machine learning models used in diabetes prediction.
b. To model six single classifiers; Logistic regression, XGBoost, Decision trees, SVM, K-NN and Random Forest, in predicting diabetes cases
c. To determine the best two machine learning models that are highly accurate in predicting diabetes
d. To develop an ensemble ML model from the two best-performing ML algorithms

[1] https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
[2] https://www.kaggle.com/datasets/johndasilva/diabetes

e. To evaluate the performance of the resultant ensemble ML model compared to single classifiers

1.4 Research questions

a. What are the current machine learning models and used in existing literature for the prediction of diabetes?
b. How do Logistic regression, XGBoost, Decision trees, SVM, Random Forest, and K-NN compare in terms of performance in predicting diabetes cases?
c. Which two ML models exhibit the highest accuracy in predicting diabetes cases?
d. How can the two best-performing machine learning algorithms be effectively combined into an ensemble model?
e. What is the performance of the resultant ensemble machine learning model?

1.5 Significance of the study

This study holds an important place in public health, specifically in diabetes prediction. This research study plays part in developing a diagnostic model for predicting diabetes thus helping in early detection of diabetes which in turn allows for proper medication and management of the disease. This goes a long way in reducing the burden and cost associated with the disease.

A meta-analysis of machine learning models presents the medical field with valuable insights as to which machine learning approaches to adopt after considering the strengths and weaknesses of each. This will give researchers and medical professionals guidance when choosing appropriate machine learning models to use in predicting diabetes. To machine learning models developers and policy makers, this project offers a robust,

validated ensemble architecture and serves as evidence to advocate for and invest in scalable, data-driven solutions for early disease detection.

The predictive accuracy of the developed ensemble machine learning model from best performing single classier models is expected to improve the accuracy significantly. This innovation will lead to more robust and flexible prediction systems that can adapt to diverse datasets and populations, enhancing the generalizability of the models.

As a country, in the spirit of primary healthcare where the government of Kenya is focusing more on preventive approaches to health as opposed to curative approaches, this innovation will come in handy in helping through screening of members of households at the community level. This will lead to early detection and medication thus preventing the disease from progressing to advanced stages leading to out-of-pocket catastrophic costs for the families.

1.6 Expected outcomes of the study

The initial expectation of this project was to model single machine learning classifiers with LR, DT, XGBoost, RF, SVM, and K-NN and determine among them the best performing model in predicting diabetes. The second expectation was to determine the two best-performing single ML classifiers that demonstrate the best accuracy in predicting diabetes. This comparison was critical in understanding the benefits of adopting advanced predictive methods. Finally, the study focused on modeling an ensemble ML algorithm by combining the two best-performing algorithms to enhance diabetes prediction accuracy. These findings would provide practical recommendations for integrating ensemble machine-

learning models into diabetes prediction systems, thereby improving early detection and intervention strategies.

1.7 Justification of the study

Early detection of diabetes would enable intervention thus reducing long term health consequences and healthcare costs associated with the disease. Dialysis, which manages renal failure occasioned by diabetes is prohibitively expensive at even one session per week through NHIF (now SHIF), costing the Kenyan exchequer a lot of money to cover these patients for an entire year. Catastrophic medical costs often bankrupt families, selling their belongings to defray these expenses and dramatically reduce the quality of life for those involved. By providing a robust and specific means of early detection, this research contributes significantly to public health in helping address disproportionate rates of diabetes plaguing underserved communities. A web-based app could be developed from this model and be used at the community level or level 2 facilities to aid in diabetes screening and patient referral, enhancing early detection efforts. Additionally, this research would enrich existing literature, providing valuable insights and data for further studies.

1.8 Scope of the study

This study deals with the building and evaluation of machine learning models for diabetes prediction among females. This research compared six single machine learning classifiers, namely LR, DT, XGBoost, RF, SVM, and K-NN, on their predictive performance for diabetes cases. It identifies the two best models from the performance evaluation and then goes further to create an ensemble model aimed at improving predictive accuracy.

Two datasets were be used in the research to train and test models, including the PIMA and the Hospital Frankfurt Germany dataset. The study overcomes the limitations of bias towards single data source and increases heterogeneity and completeness in the combined dataset. The evaluation metrics for the models were accuracy along with precision, recall and F1-score to check how well the model is at predicting these variables.

As part of the study, machine learning models were developed and their performance evaluated using comparative analysis then an ensemble model was constructed using two top performing algorithms. The study then evaluated the efficiency of this ensemble model and compared it with single models for superiority in predictive performance.

1.9 Limitations of study

The limitation of this study stems from the fact that it focused on particular ML classifiers such as XGBoost, K-NN, LR, Random Forest, Decision trees and Support vector machine. In as much as these techniques are highly popular in the field, we have numerous other techniques that are equally good and can give good performance. Therefore, this exclusion might limit the comprehensiveness of this project's findings.

2. CHAPTER TWO

2.0 LITERATURE REVIEW

2.1. Introduction

This section explores research on how ensemble machine learning classifiers are used to predict diabetes early on in patients' health journeys. It discusses various ML techniques that have been used in this area. Both individual methods and combined models and points out what they excel at and where they fall short. The chapter also looks at how researchers choose which features to focus on and how they fine-tune the parameters, for their models. Lastly it examines the data sets that have been used in studies focused on predicting diabetes. By combining the results of studies, this literature review hopes to offer a grasp of the latest advancements and pinpoint areas where more information is needed while also supporting the necessity for the suggested ensemble learning model.

2.2 Systematic review of machine learning classifiers

The systematic review of machine learning classifiers involved six . These classifiers were selected to cover the main types of modeling approaches used in healthcare: linear models (Logistic Regression), instance-based models (K-NN), kernel-based models (SVM), and powerful ensemble models (Random Forest and XGBoost, representing bagging and boosting) (Hajaj et al., 2025).

This diversity was key to fairly comparing how different mathematical approaches handle the complex patterns in your data. While other models were available, they were left out to keep the experiment focused and to manage the huge amount of computing time required for hyperparameter tuning (Bischl et al., 2023). By concentrating on this specific, high-

potential group, the study maximized the certainty of finding the most reliable and effective models for real-world use.

2.2.1 Logistic regression

Logistic regression method to predict the chances of occurrence of a binary result. This works by calculating a linear grouping of the input characteristics, each multiplied by their weight. This linear combination is then changed by the logistic function whose result is probability.

The probability function is:

$$P(Y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}$$ ……. equation 1

In this formula, β_0 represents the intercept, and β_1, β_2, …, β_n are the weights associated with the input features X_1, X_2, …, X_n. The exponent in the denominator is the negative of the linear collection of the features and corresponding weights. By applying the logistic function, this linear combination is mapped to a probability between 0 and 1, which can then be used to categorize the result as either positive or negative class.

Rajendra and Latifi (2021) investigated the prediction of diabetes using Logistic Regression as the core classification algorithm, aiming to boost performance through various techniques. Their analysis utilized two distinct datasets: the well-known PIMA Indians Diabetes dataset and a separate study based on rural African Americans in Virginia. The study incorporated methodological enhancements such as feature selection and the use of ensemble techniques (specifically Max Voting and Stacking) to overcome the limitations of the single Logistic Regression model. The researchers reported that while the base model provided a foundational prediction, performance was significantly improved by these

techniques, achieving a highest accuracy of 78% on Dataset 1 (PIMA) and 93% on Dataset 2 (Vanderbilt). The conclusion highlighted that high accuracy is dependent not just on algorithm choice, but on a robust pipeline incorporating data preprocessing, feature selection, and ensemble methods.

2.2.2 K-Nearest Neighbour

K-Nearest Neighbors works on the assumption that like data points tend to be close to one another in the feature space.

To classify or predict a new data point, the model estimates the 'k' nearest neighbors in reference the new point using a distance measure, known as Euclidean distance. The algorithm then bases its prediction on the majority class of these neighbors (for classification) or the average of their values (for regression).

In classification, if you have a new data point and want to determine its class, KNN looks at the 'k' closest training examples and assigns the class that is most common among them.

In regression, KNN computes the mean of the 'k' nearest neighbors to predict the output value.

Chandra & Nasien (2023) applied the K-Nearest Neighbour (KNN) algorithm to predict diabetes using a dataset of 390 observations. They trained the model with 80% of the data and tested it on the remaining 20%, using $k = 3$ as the optimal parameter. The model achieved a predictive accuracy of 93.58%, demonstrating the strong potential of KNN in medical diagnosis tasks such as diabetes prediction.

2.2.3 Support Vector Machine

This is a supervised learning technique employed in not only classification but also in regression. The technique identifies the best hyperplane to separate different categories in

the data, with the aim of optimizing the margin between the categories. Support Vector Machines seeks the hyperplane with the greatest margin from the nearest data points, known as support vectors.

In case the data cannot be linearly separated, SVM can use the kernel trick to map it to a higher-dimensional space which then permits linear separation. For regression, SVM finds a function that fits the data while keeping a specified margin of error. SVM perform optimally in high-dimensional spaces and are resilient to overfitting.

Ekong et al. (2024) developed a non-invasive diabetes classification model using six risk factors and SVM with 5-fold cross-validation. The model achieved 98% accuracy, with precision, recall, and F1-scores ranging from 97 to 99%, showing its strong reliability in distinguishing diabetic from healthy cases

2.2.4 Decision Trees

These are supervised machine learning methods that display versatility and is easy to use when doing classification or regression.

This algorithm partitions the data into sub-groups according to the values of the input features, thus creating a structure similar to a tree so that every internal node stands for a feature-based decision while also, every leaf node representing the final prediction. The splits are designed to maximize data separation based on information gain when it comes to for classification or variance reduction when it comes to regression.

The objective of the decision tree is to sub-divide the data so that each resulting subset is as pure as possible, this then implies that the datapoints in each subset basically belong to one class. The sub-groups in regression trees are supposed to minimize variance in every sub-group or subset.

Taser (2021) applied bagging and boosting with six decision tree-based classifiers for diabetes prediction. The ensemble methods outperformed individual classifiers, with AdaBoost and bagging with NBTree achieving the highest accuracy of 98.65%.

2.2.5 Random Forest

Random forest can be described as a resultant ensemble technique employing decision trees to enhance prediction and to be able to manage overfitting. It works by creating decision trees and then go ahead and combining their predictions. Below is a summary of how it works:

Multiple Decision Trees: RF creates numerous DT as it trains them. Every tree is created from a random partial set of the training set and a random partial set of features for each partition.

Bagging: This is a methodology that uses an approach called bootstrap aggregating where every tree get to be trained on various random samples of the same dataset. This reduces variance and overfitting that a single decision tree may exhibit.

Feature Randomness: This is a case whereby each split uses a subset of features picked randomly instead of using all features. The main aim of picking random features in to boost the model robustness and reduce cases of correlation among various individual trees.

Aggregation: Random forest then predicts by considering the output of each tree after the trees have been built. In classification, prediction will be based on e class that majority of the trees will predict. That class will be selected. In regression, final prediction will be determined by the average prediction of all the trees.

Raghavendra and Kumar (2020) evaluated the PIMA Indian Diabetes dataset from the UCI repository using the Random Forest algorithm combined with feature selection methods

(forward selection and backward elimination based on entropy). The experiment, conducted in R Studio, achieved a classification accuracy of 84.1%. The findings suggest that Random Forest, when paired with feature selection, predicts diabetes more effectively and with fewer attributes, reducing the need for less important diagnostic tests.

2.2.6 XGBoost

XGBoost is a machine learning technique that does well when it comes to not only classification but also regression. It constructs a series of decision trees in sequences, and every tree corrects the errors of the preceding tree using gradient boosting techniques. XGBoost uses regularization to avoid overfitting and advanced tree pruning methods to improve model efficiency. It also manages missing values internally and supports parallel processing, making it ideal for large datasets and complex problems. Its high accuracy, speed, and scalability have made it a well-known approach in machine learning competitions.

Wardhani and Akbar (2022) employed Extreme Gradient Boosting (XGBoost) for diabetes risk prediction. XGBoost, which improves upon gradient boosting through regularization and iterative error correction, was implemented using a tree-based approach. The model achieved an accuracy of 98.71%, highlighting its effectiveness in accurately predicting diabetes.

2.3 Review of ensemble machine learning models for diabetes prediction

Lately, following the enormous breakthrough in the machine learning domain, learning techniques have been devised which are quite good at enhancing the detection of cases of diabetes mellitus. In ensemble learning, one has a learning technique, known as the super learner, increasing accuracy by combining outputs from different machine learning

algorithms. In their paper, Doğru et al. (2023) developed the super learner framework based on four classifier models: LR, DT, RF, and gradient boosting; and a meta-learning component using SVM. The study evaluated this model on three datasets to confirm its efficacy. The datasets were the PIMA Indian dataset, 130 US hospitals dataset and early stage diabetes risk dataset. The variables involved in PIMA Indian dataset were diabetes status (target variable), age, BMI, glucose, pregnancy, blood pressure, insulin, skin thickness and diabetes pedigree function. The variables in 130 US hospitals dataset included 47 features such as patient demographics (age, race, gender), hospital encounter details, lab results (like HbA1c), medications administered, and prior health visit history. The findings proved that the super learner system performed very well in comparison with each base learner model as a fore-runner of the high-accuracy system for diagnosing diabetes: early-stage risks were forecast with 99.6% accuracy, PIMA data at 92%, and diabetes data from 130 US hospitals at 98%. One the gaps identified in this study is the use of PIMA Indian dataset which is relatively small (768 instances) and use of accuracy instead of F1-score as evaluation metric. Accuracy for classification problems in healthcare can be highly misleading when the dataset is imbalanced. A model could achieve high accuracy simply by correctly predicting the majority of non-diabetic cases while failing to identify most actual diabetic patients (producing many dangerous False Negatives).

In their paper, Dutta et al. (2022) put forward a new dataset (Bangladesh diabetes dataset – 520 instances) for diabetes from Bangladesh and an automated classification pipeline with a weighted ensemble of machine learning classifiers, namely NB, RF, DT, XGBoost, and LightGBM. It implements Grid Search hyperparameter optimization on K-fold cross-

validation, critical hyperparameters, feature selection, and missing value imputation. The results indicated a statistically significant improvement in diabetes prediction performance with the proposed weighted ensemble (Decision Tree + Random Forest + XGBoost + Light GBM) along with preprocessing techniques to attain 0.735 for accuracy and 0.832 for AUC. The study further showed that statistical imputation and RF-based feature selection combined with the identified ensemble technique had given the best results for predicting the risk of diabetes early enough. The gaps identified in this study were that the dataset was relatively small (520 instances) and also relied on accuracy as the main evaluation metric.

Mahesh et al. (2022) employing Bayesian networks and radial basis function, came up with a blended ensemble machine learning model to aid in diabetes prediction. They used the developed ensemble model to compare the performance of LR, DT, SVM, K-NN and RF. The study found that the resultant developed ensemble model performed better than the traditional single-learning classifiers by achieving an accuracy score of 97.11%. This study used the PIMA Indian dataset which is relatively small and also relied on accuracy as the main evaluation metric.

Atif, Anwer, and Talib (2022) opine that single classifier models come with several limitations, the main one being compromising on accuracy because of the models' lack of generalizability over different datasets. In remedying the situation, they developed a model employing hard voting classifier by combining SVM, LR and DT algorithms. In evaluating their model, they used the PIMA dataset and the Early-Stage Diabetes Risk Prediction Dataset. The results indicated that the ensemble model had superior outcomes with accuracies of 81.17% on PIMA dataset and 94.23% on Early-Stage Diabetes Risk

Prediction Dataset. The conclusion from this study was that the ensemble models based on hard voting classifiers especially enhanced prediction in terms of accuracy and reliability different sets of datasets notwithstanding (Atif et al., 2022). This study used the PIMA Indian dataset which is relatively small and also relied on accuracy as the main evaluation metric.

Abnoosian, Farnoosh, and Behzadi (2023) utilizing imbalanced Iraqi Patient Data, employed a pipeline-based multi-classification approach in prediction of diabetes in three unique groups: the non-diabetic, prediabetes and the diabetic. The study used base classifiers like Gaussian Naive Bayes (GNB), k-NN, RF, SVM, AdaBoost and DT. Since the dataset was imbalanced, they used a weighted ensemble method anchoring on the Area Under the Receiver Operating Characteristic Curve (AUC). To optimize performance, the study employed grid search and Bayesian optimization for hyper-parameter tuning. The resultant ensemble machine learning model compared to single machine learning models, showed superior performance. It gave an accuracy of 0.9887, F1-score of 0.9851, precision of 0.9861, a recall of 0.9792 and an AUC of 0.999. This study used the Iraqi patient dataset. This dataset was relatively large (100,000 instances) but was highly imbalanced with 91,500 not diabetic and only 8500 diabetic patients.

A study by Qi et al. (2023) developed an ensemble learning model which they named KFPredict. This model incorporated various input models with significant features and different single classifier models for diabetes prediction. The model was developed by creating neural network model (KF_NN), that was multi-input in nature. It employed recursive feature in decision trees elimination algorithm together with correlation method to determine significant and non-significant variables. The KF-NN model was then infused

with KNN, SVM and RF for the purpose of soft voting and hence creating a predictive model. The results showed that KFPredict attained accuracy score of 93.5%, sensitivity score of 85% and specificity score of 98%. Compared to single classifier models, this was an improvement of 18.18%. This research underscored the effectiveness of the KFPredict approach in giving robust prediction outcomes on PIMA diabetes data (Qi et al., 2023). The PIMA Indian dataset used in this study is relatively small and the study also relied on accuracy as the main evaluation metric.Singh et al. (2021) modelled an ensemble machine learning algorithm to predict diabetes employing XGBoost, RF, SVM, Neural Network, and DT. They named it eDiaPredict. Different evaluation metrics like specificity, Gini Index, minimum error rate, area under the curve (AUC), accuracy, area under the convex hull, sensitivity and minimum weighted coefficient were used to evaluate its performance. PIMA Indian dataset was used and eDiaPredict achieved an accuracy score of 95%. According to this study, eDiaPredict enhanced the prediction of diabetes through ensemble modeling. This study also used the PIMA Indian dataset which is relatively small and also relied on accuracy as the main evaluation metric.

Bhuvaneswari Amma (2024) developed a voting classifier known as En-RfRsK which integrated three single ML classifiers namely, K-Nearest Neighbor, Random Forest and Radial SVM to predict the risk of diabetes. To evaluate this ensemble model, the PIMA data was used. The results showed that En-RfRsK was superior to single classifiers by achieving an accuracy score of 88.89% thus showing its usefulness and effectiveness in enhancing the predictive performance. Bhuvaneswari Amma (2024) also used the PIMA Indian dataset which is relatively small and also relied on accuracy as the main evaluation metric.

In predicting blood sugar levels Yang et al. (2024) used an enhanced stacking ensemble approach that incorporated three enhanced Long Short-Term Memory (LSTM) network models like single classifiers. To ensure adaptive weighting, the study used improved Nearest Neighbor Propagation Clustering Algorithm. To evaluate the model's performance, the study used the OhioT1DM dataset. Results showed that this model achieved a Root Mean Square Error (RMSE) of 1.425 mg/dL, Mean Absolute Error (MAE) of 0.721 mg/dL, and Matthews Correlation Coefficient (MCC) of 0.982 for a 30-minute prediction horizon. The model achieved an RMSEs of 3.212 mg/dL and 6.346 mg/dL, MAEs of 1.605 mg/dL and 3.232 mg/dL, and MCCs of 0.950 and 0.930 for 45-minute and 60-minute horizons respectively. The study concluded that LSTM ensemble technique improved RMSE & MAE by 27.92% and 65.32%, in that order compared to non-ensemble model (Yang, Chen, Huang, & Li, 2024). In this study, PIMA Indian dataset was used. The dataset is relatively small hence prone to overfitting.

2.4 Gaps identified

Table 2.1: Table summary of literature review

Reference	Algorithm	Data source	Gap	Accuracy/F1-score obtained
A. Doğru et al, 2023	Logistic regression, DT, RF, and Gradient boosting, SVM	PIMA Indian dataset, 130-US hospitals dataset, early-stage diabetes risks dataset	Relatively small dataset (768 instances)	92%
A. Dutta et al, 2022	Naive Bayes, Random Forest (RF), Decision Tree (DT), XGBoost (XGB), and LightGBM (LGB)	DDC dataset Bangladesh	Relatively small l dataset (from the literature)	73.5%

T. Mahesh et al, 2022	Logistic regression, Decision Tree, Support Vector Machine, K-Nearest Neighbors and Random Forest.	PIMA Indian dataset	Relatively small dataset (768 instances)	97.11%
M. Atif et al, 2022	Logistic regression, Decision Tree, and Support Vector Machine	PIMA Indian dataset, Early-Stage Diabetes Risk Prediction Dataset	Relatively small dataset (768 instances)	94.23%
K. Abnoosian et al, 2023	Gaussian Naive Bayes (GNB), k-NN, random forest (RF), SVM, AdaBoost and decision tree (DT)	Iraqi Patient Dataset of Diabetes	Imbalanced dataset (8500 diabetic and 91,500 not diabetic)	Accuracy of 98.87%, F1-score of 98.51%
H. Qi et al, 2023	Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN)	PIMA Indian dataset	Relatively small dataset (768 instances)	93.5%
A. Singh et al, 2021	XGBoost, Random Forest, Support Vector Machine (SVM), Neural Network, and Decision Tree	PIMA Indian dataset	Relatively small dataset (768 instances)	95%
N. Bhuvaneswari, 2024	K-Nearest Neighbor, Random Forest and Radial Support Vector Machine	PIMA Indian dataset	Relatively small dataset (768 instances)	88.89%

A review of recent studies on machine learning models for diabetes prediction reveals that a wide range of algorithms such as Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, K-Nearest Neighbors, Gradient Boosting, and Naive Bayes have been widely applied. The PIMA Indian Diabetes dataset remains the most frequently used data source across the literature, appearing in studies by Doğru et al. (2023), Mahesh et al.

(2022), Atif et al. (2022), Qi et al. (2023), Singh et al. (2021), and Bhuvaneswari (2024). Despite the variety of algorithms explored, a dominant limitation is the over-reliance on relatively small datasets, particularly the PIMA dataset which only contains 768 instances. This limits the generalizability and robustness of model performance in real-world, diverse populations.

In addition to small sample sizes, other studies such as Dutta et al. (2022) and Abnoosian et al. (2023) have used region-specific datasets like the DDC Bangladesh dataset and the Iraqi Patient Dataset. However, these also come with limitations. For instance, the Iraqi dataset is significantly imbalanced, with a disproportionately higher number of non-diabetic cases (91,500) compared to diabetic ones (8,500), posing challenges for fair model training and evaluation.

2.5 Conclusion

From the literature review, studies on diabetes prediction predominantly rely on relatively small and often imbalanced datasets most notably the PIMA Indian dataset with only 768 instances and Bangladesh diabetes dataset which has only 520 instances. The other dataset used to develop the enseble model was the Iraqi diabetes dataset which was found to be highly imbalanced. This limitation constrains the development of robust and generalizable machine learning models. Additionally, while a variety of algorithms such as Random Forest, Support Vector Machines, XGBoost, and Decision Trees have been tested, few studies have explored ensemble approaches built from multiple strong-performing base learners, especially using diverse data sources.

The review also found that most ensemble machine learning models developed based their evaluation metric on accuracy. Accuracy in classification problems in healthcare can be mislead especially when the dataset is imbalanced (as is common in disease prediction). A model can achieve high accuracy simply by correctly predicting the majority of non-diabetic cases while failing to identify most actual diabetic patients (producing many dangerous False Negatives).

It is on this basis that this project seeks to bridge these gaps by developing an accurate ensemble machine learning model for diabetes prediction. The model will be built from a hybrid of datasets—the PIMA dataset and the Hospital Frankfurt Germany dataset (with over 2,000 datapoints)—to improve data diversity and model generalizability. From among the single machine learning models (Random Forest, Support Vector Machines, XGBoost, and K-Nearest Neighbor), the two best-performing algorithms will be selected and combined to form the ensemble model. This approach aims to enhance predictive accuracy and support early diagnosis, especially in resource-limited settings where delayed detection contributes to rising diabetes-related morbidity and mortality. The study will also use F1-score as the main evaluation metric in the evaluation of performance of the single classifier models and even the resultant ensemble model because it provides a critical balance between Precision and Recall, which is essential in a medical diagnostic context where minimizing both missed diagnoses (False Negatives) and false alarms (False Positives) is crucial for patient safety and resource allocation.

2.6 Conceptual framework

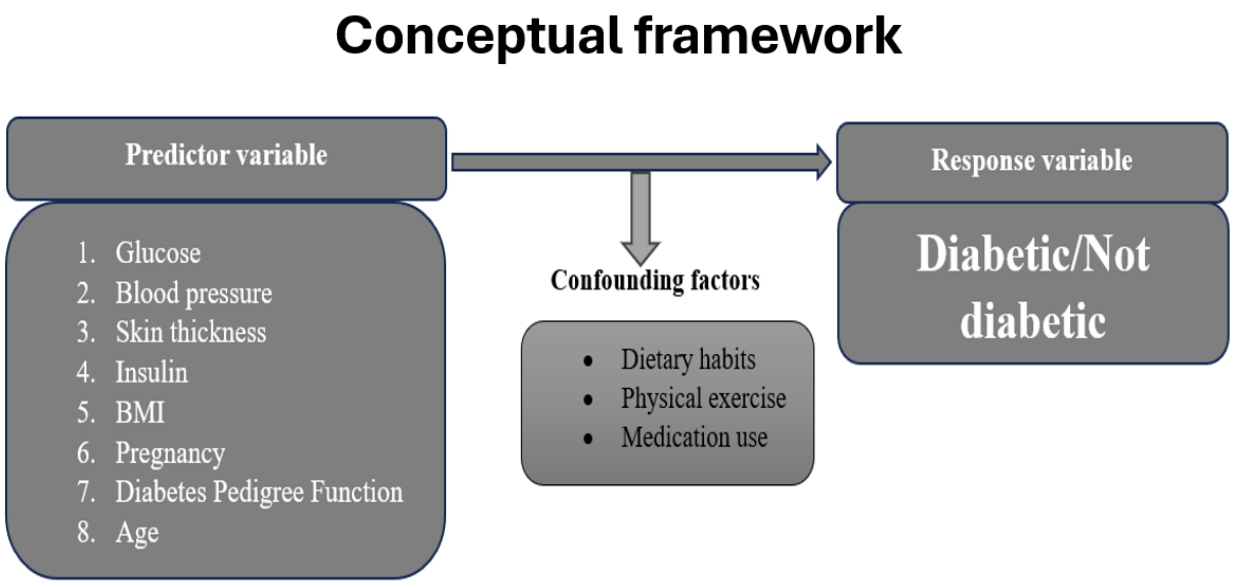


Figure 2.1: Conceptual framework

3. CHAPTER THREE

3.0 METHODOLOGY

3.1 Introduction

In machine learning, systems make sense of data fed to them, generating algorithms making them accurate in prediction. The algorithms are utilized to solve various issues, including predicting diseases in healthcare including the chances of being diabetic. This research project developed 6 machine learning models, namely; logistic regression, k-NN, SVM, DT, RF, and XGBoost, and introduced various evaluation metrics to identify the two best-performing models for developing an ensemble model. This chapter covers dataset description, the data pre-processing steps, model development, model evaluation metrics, and the selection process for the best models.

3.2 Data source and description

Aggregated data was sourced from two sources; PIMA Indian dataset and Dataset of diabetes, taken from the hospital Frankfurt, Germany all of which are available in Kaggle. The Indian PIMA dataset has 768 data points while the Hospital Frankfurt dataset has 2000 data points. The variables in the datasets are diabetes status, glucose level, blood pressure, skin thickness, insulin level, BMI (body mass index), pregnancy, diabetes pedigree function and age.

Table 3.1: Dataset attributes

No.	Attribute	Data type	Description
1	Pregnancy	Numeric	The number of pregnancies
2	Glucose	Numeric	Glucose plasma levels two hours after consuming glucose
3	Blood pressure	Numeric	Diastolic blood pressure (mm Hg)
4	Skin thickness	Numeric	Thickness of the skin fold on the triceps of the upper arm (mm)
5	Insulin	Numeric	Insulin serum levels in the blood two hours after the glucose test (lh/
6	BMI	Numeric	Body mass index [weight in kg/(Height in m)]
7	Diabetes Pedigree Function	Numeric	Numerical value that estimates a person's genetic risk of developing diabetes based on their family history.
8	Age	Numeric	Patient age
9	Diabetes status	Categorical	Diagnostic results (Diabetic or not diabetic)

3.3 Data pre-processing steps

3.3.1 Outlier Removal

To manage outliers, the study employed techniques such as the interquartile range method. This process maintains data integrity by excluding data points that are abnormally big or small according to the range that will be defined by the methods. The range allowable for non-outliers was defined as 1.5 times the interquartile range (IQR) applied to both the lower and upper boundaries of the data distribution. After applying these methods, no outliers were found, allowing all data points (2,768) to be retained for subsequent modeling phases

3.3.2 Dealing with Missing Values

Measures of central tendency (median) technique of imputation was used to manage missing values. This approach ensures that the dataset remains comprehensive and usable.

3.3.3 Data Standardization/Normalization

The project employed Min-Max scaling method to ensure that the range of features are normalized. This will ensure that data fall in a given range, always between 0 and 1. This

in turn strengthened the robustness of machine learning models that are sensitive to feature scaling.

3.3.4 Encoding

Label encoding was employed to assign numerical values to each category of a qualitative variable. This applied to the dependent variable only as it was the only categorical variable in the dataset. Label Encoding was sufficient because the dependent variable is dichotomous (two classes), meaning it only requires two unique numerical identifiers (0 and 1) which Label Encoding provides without creating redundant columns.

3.3.5 Feature Selection

Selecting the variables was very key to improving the performance of the model as this reduces complexities in computation. We calculated the correlation coefficients between attributes and the final output and visualize these correlations using a heat map. Variables having smallest correlation coefficients, which are not significant to the output, were to be eliminated. This ensured that only significant variables are maintained in the dataset, improving the models performance in terms of efficacy and accuracy.

3.4 Model development and evaluation

During modeling construction phase, the project did a series of steps to build, evaluate, and optimize the machine learning model. These steps helped to make the final model accurate and robust.

3.4.1 Splitting of the Data

The research dataset was divided into two splits: one was the training set, and the other was the testing set. The training set had 80% of the data while the testing set had 20% of the data. An 80/20 train-test split was used because it provides enough data for training

while still keeping sufficient unseen data to reliably evaluate model performance, and the choice of train/test proportions directly impacts the performance of resampling methods such as bootstrap, cross-validation, and repeated random splits (Vrigazova, 2021)..

3.4.2 Performance Analysis

To evaluate the performance of each model, the project used several evaluation metrics:

Accuracy: This was to quantify how much of the total number of samples were correctly predicted. It measures the model reliability on classification of a positive or negative cases. On the other hand, when training dataset that is imbalanced, accuracy can be misleading; it quickly goes upwards without paying attention to detection of positive class.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN}}$$

The subject should be classified as either diabetic or not diabetic. Out of the classification, we had subjects truly classified and falsely classified hence the below;

True positive (TP): The subject is predicted as positive (diabetic) and is diabetic.

True negative (TP): The subject is predicted as negative (not diabetic) and is actually not diabetic.

False positive (FP): The subject is predicted as positive (diabetic) and is not diabetic.

False negative (FN): The subject is predicted as negative (not diabetic) but is actually diabetic.

Precision: Indicates the proportion of true positive rightly predicted among all predicted positives.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}}$$

Recall: Reflects the ability of the model to identify all relevant instances (true positives).

$$Recall = \frac{TP}{TP + FN}$$

F1-Score: A harmonic mean of precision and recall, providing a single metric for model evaluation.

Area Under the Receiver Operating Characteristic (ROC) Curve (AUC): Evaluates the model's power to differentiate between positive and negative classes.

In this research, while metrics such Accuracy, Precision, Recall were used to evaluate the single classifiers, the ensemble model was developed based on F1-score metric. This is because Accuracy can be misleading in imbalanced datasets, as a model might appear effective by simply predicting the majority class. Precision, while indicating the reliability of positive predictions, can lead to numerous missed actual cases (low Recall) if prioritized exclusively. Conversely, Recall, crucial for minimizing dangerous false negatives (missed diagnoses in diabetes), can result in an unacceptably high rate of false positives (unnecessary alarms for healthy individuals) if optimized in isolation. Therefore, to strike a vital balance between accurately identifying true diabetic cases and ensuring the credibility of positive predictions, the F1-score was selected as the primary metric for determining the two best-performing single classifier models. The F1-score effectively captures the harmonic mean of Precision and Recall, making it particularly suitable for this medical diagnostic context where both types of errors carry significant consequences.

3.4.3 Hyperparameter Tuning

The project conducted hyperparameter tuning in order get the best model performance. It was done by adjusting the parameters of the model to find the best settings to improve

accuracy and other performance metrics. We did not only use grid search but also random search to systematically try a range of hyperparameter values.

Approach architecture

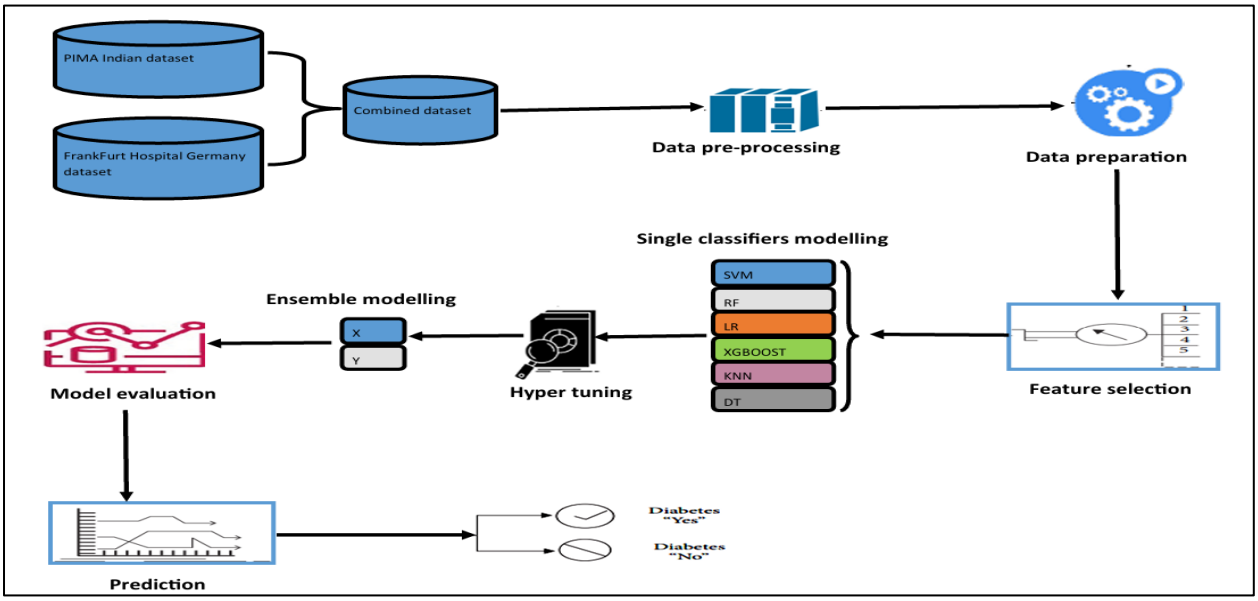


Figure 3.1: Approach architecture of the ensemble model

3.5 Ethical considerations

This study used secondary data sourced from publicly available datasets on Kaggle, namely the PIMA Indian Diabetes dataset and the Frankfurt Hospital Germany dataset. These datasets are anonymized and do not contain personally identifiable information. Therefore, the study does not raise significant privacy or confidentiality concerns.

Ethical use of data has been ensured by adhering to the terms and conditions set by the dataset providers. Proper citations are provided to acknowledge the original data sources. No attempts were made to re-identify individuals or link the data to other external sources.

Since the research does not involve direct interaction with human subjects or collection of primary data, formal ethical clearance was not required. However, the study follows general ethical research standards, including integrity in reporting results, avoidance of data manipulation, and transparency in methodology.

4. CHAPTER FOUR

4.0 MODEL DEVELOPMENT, ANALYSIS AND RESULTS

4.1 Introduction

This chapter provides the technical foundation of the project by outlining the systematic steps taken to develop, tune, and evaluate machine learning models for predicting diabetes. It details the exploratory data analysis, model training, ensemble modeling, and performance evaluation, all aimed at constructing an optimal ensemble model that delivers accurate diabetes predictions.

4.2.2 Missing Values

The project employed a crucial first step involving the use of a missing values heatmap. This visualization was instrumental in detecting the presence and spatial distribution of any missing data points across the dataset. By distinctly highlighting missing entries, the heatmap offered an immediate and clear visual assessment of data completeness. From figure 4.1 below, it can be observed that there were no missing data.

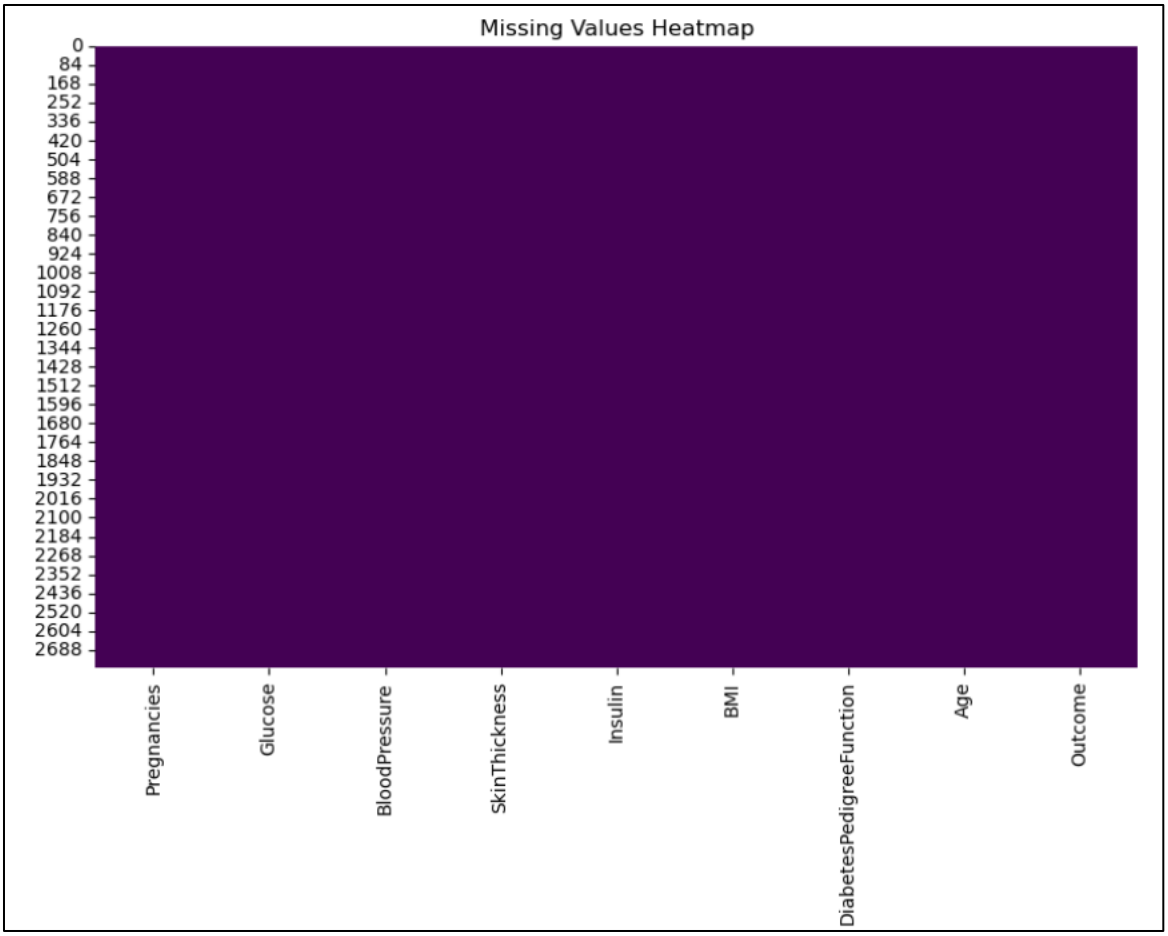


Figure 4.1: Missing values heatmap

4.2.3 Class balance analysis

To gain a comprehensive understanding of the dataset's core, the distribution of the target variable (diabetes status) was analyzed, providing a clear picture of the class balance. This step was vital for identifying potential class imbalance that might necessitate specific handling during model training**.** The analysis result is as shown below;

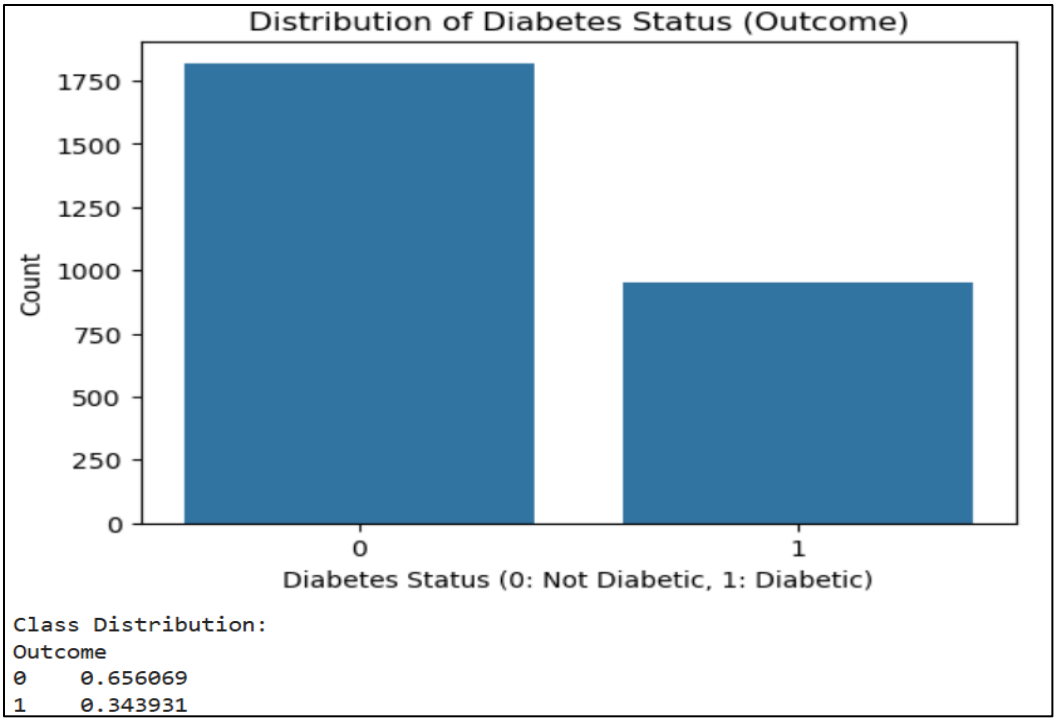


Figure 4.2: Distribution of diabetes status

4.2.4 Feature scaling

To prepare the dataset for effective model training and to mitigate the influence of varying scales among features, Standardization was employed. This technique, also known as Z-score normalization, transforms each numerical feature to have a mean of zero and a standard deviation of one. The implementation involved a two-step process: first, the central tendency (mean) and spread (standard deviation) were computed for each feature based on the imputed dataset. Subsequently, each individual data point was transformed by subtracting its feature's calculated mean and then dividing by its standard deviation. This systematic re-scaling ensures that all features are brought to a comparable scale, preventing those with inherently larger ranges or magnitudes from disproportionately impacting the learning process of scale-sensitive algorithms. This is particularly crucial for models such as Logistic Regression, Support Vector Machines and K-Nearest Neighbors, as it

contributes to more stable model convergence and improved predictive performance. The dataset, now uniformly scaled, was then prepared for subsequent model development.

Table 4.1: First 5 rows after features scaling

```
Features after scaling (first 5 rows):
   Pregnancies   Glucose  BloodPressure  SkinThickness   Insulin       BMI  \
0     0.679232  0.839738       0.149033       0.882845 -0.713633  0.181135
1    -0.825341 -1.127124      -0.163012       0.509169 -0.713633 -0.685773
2     1.281062  1.932439      -0.267027      -1.296931 -0.713633 -1.094459
3    -0.825341 -1.002244      -0.163012       0.135494  0.123547 -0.500007
4    -1.126256  0.496317      -1.515209       0.882845  0.782604  1.357654

   DiabetesPedigreeFunction       Age
0                  0.478509  1.432495
1                 -0.369130 -0.181079
2                  0.616712 -0.096154
3                 -0.934224 -1.030329
4                  5.579704 -0.011229
```

4.2.5 Encoding of Categorical Variable

The dependent variable "diabetes status" was encoded as a binary variable (0 = non-diabetic, 1 = diabetic) to enable model compatibility.

4.3 Descriptive statistics

4.3.1 Measures of central tendency and dispersion

Table 4.2: Summary statistics for independent variables

	mean	median	min	max	std
Pregnancies	3.742775	3.000	0.000	17.00	3.323801
Glucose	121.863439	117.000	44.000	199.00	30.503499
BloodPressure	72.385838	72.000	24.000	122.00	11.988549
SkinThickness	29.205925	29.000	7.000	110.00	9.032220
Insulin	140.669798	126.000	14.000	846.00	82.887979
BMI	32.593895	32.400	18.200	80.60	7.103462
DiabetesPedigreeFunction	0.471193	0.375	0.078	2.42	0.325669
Age	33.132225	29.000	21.000	81.00	11.777230
Outcome	0.343931	0.000	0.000	1.00	0.475104

4.3.1 Distribution analysis

The individual characteristics of all variables were explored through histograms, which revealed the distribution patterns of each numerical variable. Complementing this, box plots were generated for each variable to effectively identify outliers and understand the spread and central tendency of the data.

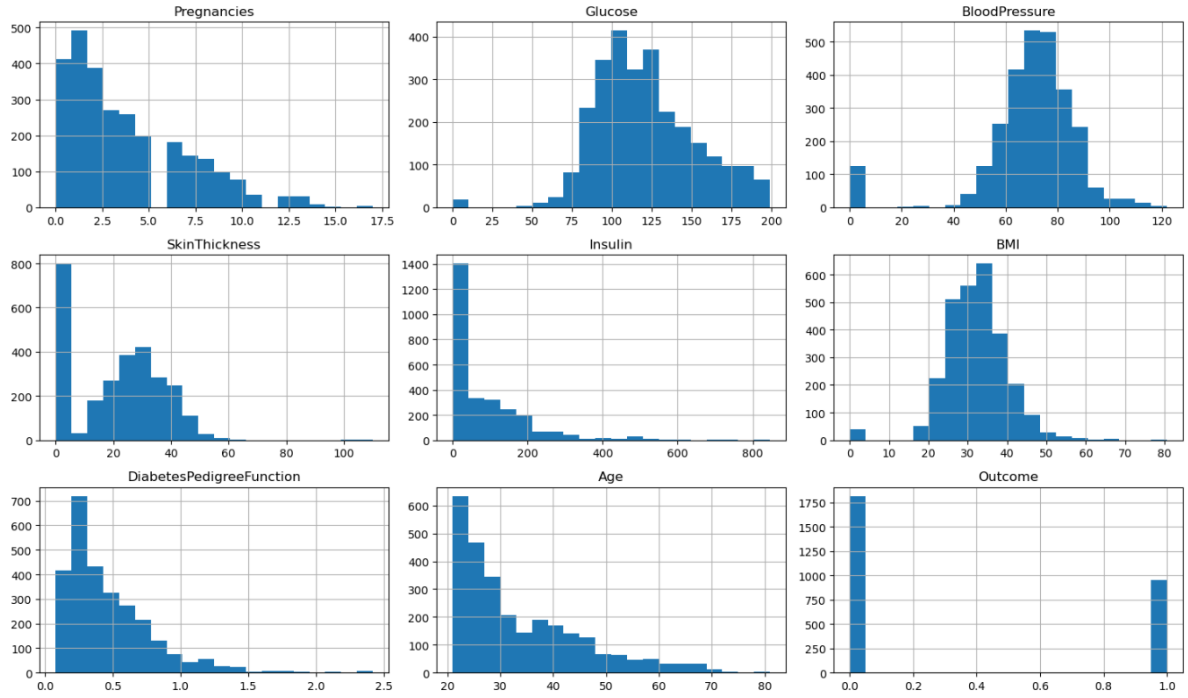


Figure 4.3: Histogram of features

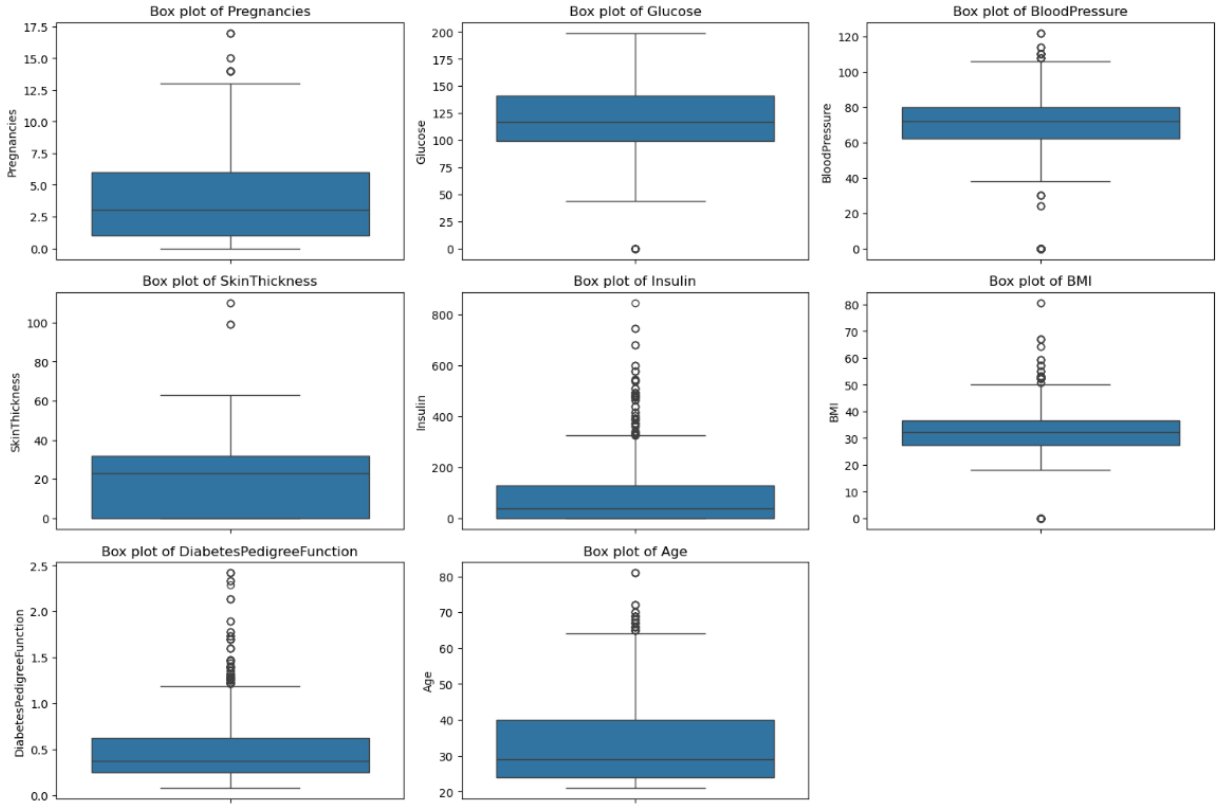


Figure 4.4: Boxplots of features

4.3.2 Feature Selection

Feature selection was conducted to identify the most relevant independent variables for model training and to address potential issues of multicollinearity. Correlation heatmaps were employed to assess the relationships between independent variables. During this assessment, no significant cases of multicollinearity were identified among the independent variables as can be observed in figure 5, therefore, all independent variables were retained for subsequent analysis, ensuring the preservation of their individual predictive utility.

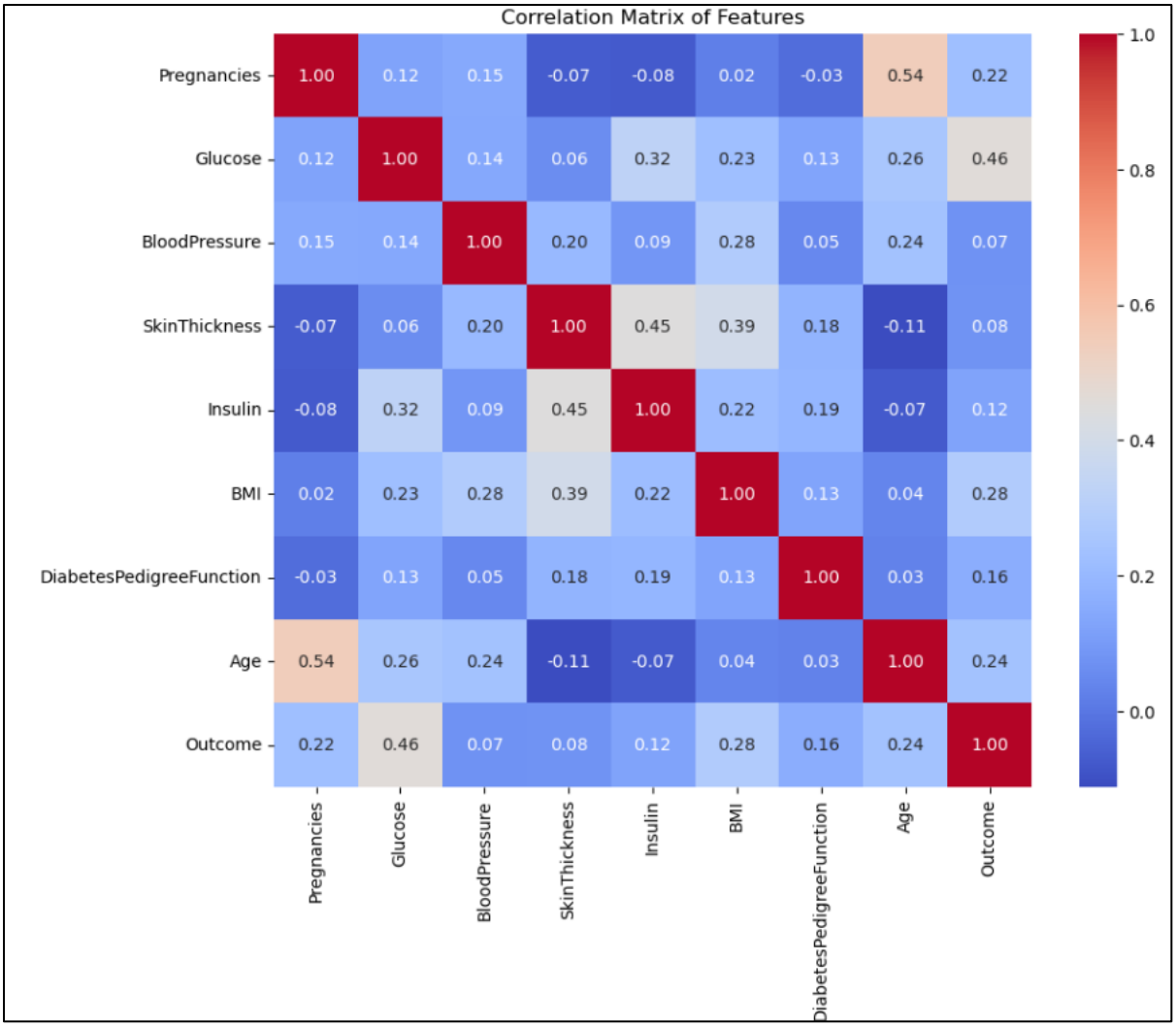


Figure 4.5: Heatmap showing correlation between independent variables

4.4 Model building and Evaluation

4.4.1 Logistic regression

The Logistic Regression model was initially evaluated using its default hyperparameters to establish a baseline performance. Subsequently, hyperparameter tuning was performed using GridSearchCV to optimize its predictive capabilities on the Diabetes dataset. This type of optimization systematically tested every combination of the regularization strength (C) and penalty type (penalty) hyperparameters, selecting the configuration that yielded the highest ROC-AUC score across 5-fold cross-validation.

Table 4.3: Default logistic regression performance

```
Default Logistic Regression Performance:
  Accuracy: 0.7726
  Precision: 0.7211
  Recall: 0.5550
  F1-Score: 0.6272
  ROC-AUC: 0.8355
```

With its default parameters, the Logistic Regression model demonstrated an Accuracy of 0.7726, indicating that approximately 77.26% of the predictions were correct. It achieved a Precision of 0.7211, meaning that 72.11% of the instances predicted as positive for diabetes were indeed positive. The Recall stood at 0.5550, suggesting that the model correctly identified 55.50% of all actual diabetes cases. The F1-Score, a harmonic mean of precision and recall, was 0.6272, providing a balanced measure of the model's performance. The ROC-AUC (Area Under the Receiver Operating Characteristic Curve) was 0.8355, indicating a good ability to distinguish between diabetic and non-diabetic individuals.

Table 4.4: Tuned Logistic Regression Performance

```
Tuned Logistic Regression Performance:
  Accuracy: 0.7780
  Precision: 0.7361
  Recall: 0.5550
  F1-Score: 0.6328
  ROC-AUC: 0.8355
```

Hyperparameter tuning achieved a best cross-validation ROC-AUC of 0.8348 during the tuning phase.

Upon evaluating the Logistic Regression model with these optimized hyperparameters on the held-out test set, a slight improvement in performance was observed. The Accuracy increased marginally to 0.7780. Similarly, Precision saw a slight rise to 0.7361. The Recall remained constant at 0.5550, indicating no change in the model's ability to identify actual positive cases at this threshold. Consequently, the F1-Score improved slightly to 0.6328. The ROC-AUC remained consistent at 0.8355, suggesting that while other metrics saw minor shifts, the overall discriminative power of the model was maintained.

Table 4.5: Confusion matrix for tuned LR model

```
 Confusion Matrix (Tuned):
[[325  38]
 [ 85 106]]
 Classification Report (Tuned):
              precision    recall  f1-score   support

           0       0.79      0.90      0.84       363
           1       0.74      0.55      0.63       191

    accuracy                           0.78       554
   macro avg       0.76      0.73      0.74       554
weighted avg       0.77      0.78      0.77       554
```

The confusion matrix for the tuned Logistic Regression model further details its performance:

True Negatives (TN): 325 (correctly predicted non-diabetic)

False Positives (FP): 38 (incorrectly predicted as diabetic - Type I error)

False Negatives (FN): 85 (incorrectly predicted as non-diabetic - Type II error)

True Positives (TP): 106 (correctly predicted as diabetic)

The classification report provides more granular insights per class:

Class 0 (Non-Diabetic): The model achieved 79% precision and 90% recall, with an 84% f1-score, indicating strong performance in identifying non-diabetic cases.

Class 1 (Diabetic): For the diabetic class, the model showed 74% precision and 55% recall, resulting in a 63% f1-score. This highlights that while the precision for diabetic predictions is reasonable, the recall for this class is lower, meaning a notable portion of actual diabetic cases were missed (false negatives).

In summary, hyperparameter tuning led to minor improvements in accuracy, precision, and F1-score for the Logistic Regression model, affirming its solid baseline performance while showing that even a robust model can be slightly refined through optimization. The consistency in ROC-AUC suggests the overall ranking ability of positive and negative instances remained high.

4.4.2 Decision tree model

Operating with its default parameters, the Decision Tree model demonstrated exceptionally high performance metrics. It achieved an Accuracy of 0.9892, correctly classifying nearly 99% of instances. Its Precision was 0.9843, meaning a very high proportion of positive predictions were correct, and its Recall was also 0.9843, indicating excellent identification of actual positive cases. Consequently, the F1-Score was 0.9843, reflecting a strong balance between precision and recall. The ROC-AUC of 0.9880 further affirmed the model's outstanding discriminative capability.

Table 4.6: Default Decision Tree performance

```
--- Evaluating Decision Tree with Default Parameters (on Test Set)
Default Decision Tree Performance:
  Accuracy: 0.9892
  Precision: 0.9843
  Recall: 0.9843
  F1-Score: 0.9843
  ROC-AUC: 0.9880
```

Upon evaluating the Decision Tree model with these 'tuned' hyperparameters on the held-out test set, the performance remained identical to the default model. The Accuracy, Precision, Recall, and F1-Score all remained at 0.9892, 0.9843, 0.9843, and 0.9843 respectively. The ROC-AUC also stayed consistent at 0.9880. This indicates that for this dataset, the default parameters of the Decision Tree, which allow it to fully learn the training data, were already optimal or very close to optimal in terms of the chosen evaluation metrics.

Table 4.7: Tuned Decision Tree Performance

```
Tuned Decision Tree Performance:
  Accuracy: 0.9892
  Precision: 0.9843
  Recall: 0.9843
  F1-Score: 0.9843
  ROC-AUC: 0.9880
```

The confusion matrix for the tuned Decision Tree model illustrates its high accuracy, detailing the classification outcomes: 360 True Negatives (correctly predicted non-diabetic) and 188 True Positives (correctly predicted as diabetic), alongside only 3 False Positives (incorrectly predicted as diabetic) and 3 False Negatives (incorrectly predicted as non-diabetic). The classification report further breaks down this high performance per class: for Class 0 (Non-Diabetic), the model achieved 99% precision, 99% recall, and 99% F1-score, demonstrating almost perfect identification. Similarly, for Class 1 (Diabetic), the

model showed very strong performance with 98% precision, 98% recall, and 98% F1-score, indicating highly accurate detection of actual diabetic cases with very few misses or false alarms.

In summary, the Decision Tree model demonstrated an exceptionally high level of performance on this dataset right out of the box. The hyperparameter tuning process confirmed that its default parameters were already highly effective, leading to no discernible change in performance metrics on the test set after optimization. This suggests that the Decision Tree, when allowed to fully develop, is a powerful predictor for this specific dataset.

Table 4.8: Confusion matrix for tuned Decision Tree model

```
 Confusion Matrix (Tuned):
[[360   3]
[  3 188]]
 Classification Report (Tuned):
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       363
           1       0.98      0.98      0.98       191

    accuracy                           0.99       554
   macro avg       0.99      0.99      0.99       554
weighted avg       0.99      0.99      0.99       554
```

4.4.3 K-Nearest Neighbour model

With its default parameters, the KNN model achieved an Accuracy of 0.8448, correctly classifying approximately 84.48% of the instances. Its Precision stood at 0.7807, indicating that 78.07% of positive predictions were indeed correct, while Recall was 0.7644, meaning it identified 76.44% of all actual positive cases. The F1-Score was 0.7725, representing a solid balance between precision and recall. The ROC-AUC of 0.9391 demonstrated strong discriminative power even without specific optimization.

Table 4.9: Default K-Nearest Neighbour performance

```
Default K-Nearest Neighbors Performance:
  Accuracy: 0.8448
  Precision: 0.7807
  Recall: 0.7644
  F1-Score: 0.7725
  ROC-AUC: 0.9391
```

Hyperparameter tuning for the KNN model explored various configurations, including the number of neighbors (n_neighbors), weighting scheme (weights), and distance metric (metric). The GridSearchCV process identified the optimal hyperparameters as n_neighbors =13, metric='manhattan', and weights='distance'. This optimized configuration achieved a remarkable best cross-validation ROC-AUC of 0.9968 during the tuning phase.

Evaluating the KNN model with these optimized hyperparameters on the held-out test set revealed a dramatic improvement in performance. The Accuracy surged to 0.9910, indicating nearly perfect classification. Both Precision and Recall also showed substantial increases, reaching 0.9844 and 0.9895, respectively. This led to an outstanding F1-Score of 0.9869. Most notably, the ROC-AUC improved to an exceptional 0.9999, signifying near-perfect ability to distinguish between the two classes as shown in the table below.

Table 4.10: Tuned K-Nearest Neighbour performance

```
Tuned K-Nearest Neighbors Performance:
  Accuracy: 0.9910
  Precision: 0.9844
  Recall: 0.9895
  F1-Score: 0.9869
  ROC-AUC: 0.9999
```

The confusion matrix for the tuned K-Nearest Neighbors (KNN) model clearly illustrates its high level of accuracy, showing 360 True Negatives (correctly predicted non-diabetic) and 189 True Positives (correctly predicted as diabetic), alongside only 3 False Positives (incorrectly predicted as diabetic) and 2 False Negatives (incorrectly predicted as non-diabetic). The classification report provides a detailed breakdown per class: for Class 0 (Non-Diabetic), the model achieved 99% precision, 99% recall, and 99% F1-score, indicating almost flawless identification of non-diabetic cases. Performance was similarly impressive for Class 1 (Diabetic), with the model showing 98% precision, 99% recall, and 99% F1-score, demonstrating highly accurate detection of actual diabetic cases with minimal misses.

Table 4.11: Confusion matrix for tuned K-Nearest Neighbour model

```
 Confusion Matrix (Tuned):
[[360   3]
 [  2 189]]
 Classification Report (Tuned):
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       363
           1       0.98      0.99      0.99       191

    accuracy                           0.99       554
   macro avg       0.99      0.99      0.99       554
weighted avg       0.99      0.99      0.99       554
```

In summary, hyperparameter tuning had a profound positive impact on the K-Nearest Neighbors model. By optimizing its configuration, the KNN classifier transformed from a good performer into an exceptionally accurate and robust model, achieving near-perfect scores across all key evaluation metrics on the test set.

4.4.4 Support Vector Machine model

The Support Vector Machine (SVM) classifier was initially evaluated using its default settings to establish baseline performance. Subsequently, hyperparameter tuning was performed using GridSearchCV to optimize its predictive capabilities.

Default Support Vector Machine Performance

With its default parameters, the SVM model achieved an Accuracy of 0.8412, correctly classifying approximately 84.12% of the instances. It demonstrated strong agreement in its positive predictions, with a Precision of 0.8411, meaning that 84.11% of predicted positive cases were accurate. However, its Recall was 0.6649, indicating that it identified 66.49% of all actual positive cases. The F1-Score was 0.7427, reflecting a reasonable balance, though with room for improvement in recall. The ROC-AUC of 0.9002 suggested good discriminative ability.

Table 4.12: Default support vector machine performance

```
Default Support Vector Machine Performance:
  Accuracy: 0.8412
  Precision: 0.8411
  Recall: 0.6649
  F1-Score: 0.7427
  ROC-AUC: 0.9002
```

Tuned Support Vector Machine Performance

Hyperparameter tuning for the SVM model involved exploring different values for the regularization parameter (C), kernel type (kernel), and kernel coefficient (gamma). The GridSearchCV process identified the optimal hyperparameters as C=10, gamma=1, and

kernel='rbf'. This optimized configuration yielded a very high best cross-validation ROC-AUC of 0.9963 during the tuning phase.

Upon evaluating the SVM model with these optimized hyperparameters on the held-out test set, a substantial improvement in performance was observed, transforming it into an exceptionally strong classifier. The Accuracy surged to 0.9892, indicating near-perfect classification. Both Precision and Recall also significantly increased to 0.9843 and 0.9843 respectively, leading to an outstanding F1-Score of 0.9843. Most notably, the ROC-AUC improved dramatically to 0.9996, signifying near-perfect discriminative capability between the two classes as seen in the table below.

Table 4.13: Tuned SVM performance

```
Tuned Support Vector Machine Performance:
  Accuracy: 0.9892
  Precision: 0.9843
  Recall: 0.9843
  F1-Score: 0.9843
  ROC-AUC: 0.9996
```

The confusion matrix for the tuned Support Vector Machine (SVM) model clearly illustrates its high accuracy, detailing 360 True Negatives (correctly predicted non-diabetic) and 188 True Positives (correctly predicted as diabetic), alongside only 3 False Positives (incorrectly predicted as diabetic) and 3 False Negatives (incorrectly predicted as non-diabetic). The classification report provides a detailed breakdown per class: for Class 0 (Non-Diabetic), the model achieved 99% precision, 99% recall, and 99% F1-score, demonstrating almost flawless identification. Performance was similarly impressive for Class 1 (Diabetic), with the model showing 98% precision, 98% recall, and 98% F1-score, indicating highly accurate detection of actual diabetic cases with minimal misses.

Table 4.14: Confusion matrix for tuned SVM model

```
 Confusion Matrix (Tuned):
[[360   3]
[  3 188]]
 Classification Report (Tuned):
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       363
           1       0.98      0.98      0.98       191

    accuracy                           0.99       554
   macro avg       0.99      0.99      0.99       554
weighted avg       0.99      0.99      0.99       554
```

In summary, hyperparameter tuning profoundly enhanced the Support Vector Machine model's performance. By optimizing its configuration, the SVM classifier achieved near-perfect scores across all key evaluation metrics on the test set, demonstrating its strong capability for the diabetes prediction task.

4.4.5 Random Forest model

The Random Forest classifier was initially evaluated using its default settings, followed by hyperparameter tuning using GridSearchCV to explore potential optimizations.

Default Random Forest Performance

Operating with its default parameters, the Random Forest model demonstrated an exceptionally high level of performance. It achieved an Accuracy of 0.9964, meaning it correctly classified nearly 100% of the instances. Its Precision was a perfect 1.0000, indicating no false positive predictions. The Recall stood at 0.9895, meaning it successfully identified almost all actual positive cases. This resulted in an F1-Score of 0.9947, showcasing an outstanding balance between precision and recall. The ROC-AUC was 0.9999, signifying near-perfect discriminative power.

Table 4.15: Default Random Forest performance

```
Default Random Forest Performance:
  Accuracy: 0.9964
  Precision: 1.0000
  Recall: 0.9895
  F1-Score: 0.9947
  ROC-AUC: 0.9999
```

Tuned Random Forest Performance

Hyperparameter tuning for the Random Forest model explored a broad range of parameters including the number of estimators (n_estimators), maximum tree depth (max_depth), minimum samples required to split a node (min_samples_split), and minimum samples required at a leaf node (min_samples_leaf). The GridSearchCV process identified the optimal hyperparameters as max_depth=20, min_samples_leaf=1, min_samples_split=2, and n_estimators=400. The best cross-validation ROC-AUC achieved during tuning was 0.9962.

Upon evaluating the Random Forest model with these optimized hyperparameters on the held-out test set, the performance remained identical to its default configuration. The Accuracy, Precision, Recall, and F1-Score all remained at 0.9964, 1.0000, 0.9895, and 0.9947 respectively. The ROC-AUC also remained at 0.9999. This outcome suggests that the default parameters of the Random Forest model were already highly optimized for this dataset, or that the dataset's characteristics allowed the default configuration to achieve near-maximal performance, leaving little room for further improvement through the explored tuning space.

Table 4.16: Tuned Random Forest model results

```
Tuned Random Forest Performance:
  Accuracy: 0.9964
  Precision: 1.0000
  Recall: 0.9895
  F1-Score: 0.9947
  ROC-AUC: 0.9999
```

The confusion matrix for the tuned Random Forest model highlights its near-perfect classification, demonstrating 363 True Negatives (correctly predicted non-diabetic) and 189 True Positives (correctly predicted as diabetic). Notably, the model recorded 0 False Positives (no incorrect predictions as diabetic) and only 2 False Negatives (incorrectly predicted as non-diabetic). The classification report further details this exceptional performance per class: for Class 0 (Non-Diabetic), the model achieved 99% precision, 100% recall, and 100% F1-score, indicating virtually flawless identification. Performance for Class 1 (Diabetic) was also superb with 100% precision, 99% recall, and 99% F1-score, demonstrating highly accurate detection of actual diabetic cases with minimal misses and no false alarms.

Table 4.17: Confusion matrix results for Random Forest model

```
 Confusion Matrix (Tuned):
[[363   0]
[  2 189]]
 Classification Report (Tuned):
              precision    recall  f1-score   support

           0       0.99      1.00      1.00       363
           1       1.00      0.99      0.99       191

    accuracy                           1.00       554
   macro avg       1.00      0.99      1.00       554
weighted avg       1.00      1.00      1.00       554
```

In summary, the Random Forest model exhibited outstanding performance on the Diabetes dataset right from its default configuration. The hyperparameter tuning process confirmed

that its initial settings were already near-optimal, resulting in no measurable change in its already exceptionally high-performance metrics on the test set after optimization.

4.4.6 XGBoost model performance

The XGBoost classifier, a powerful gradient boosting framework, was initially evaluated using its default settings. This was followed by a comprehensive hyperparameter tuning process utilizing GridSearchCV.

Default XGBoost Performance

With its default parameters, the XGBoost model demonstrated a very high level of predictive capability. It achieved an Accuracy of 0.9928, indicating that over 99% of its predictions were correct. The Precision was 0.9845, signifying that nearly all instances predicted as positive were indeed correct. Its Recall stood at 0.9948, showing excellent ability to identify actual positive cases. This resulted in a robust F1-Score of 0.9896, indicating a strong balance between precision and recall. The ROC-AUC was 0.9981, reflecting an almost perfect discriminative capacity.

Table 4.18: Default XGBoost performance

```
Default XGBoost Performance:
  Accuracy: 0.9928
  Precision: 0.9845
  Recall: 0.9948
  F1-Score: 0.9896
  ROC-AUC: 0.9981
```

Tuned XGBoost Performance

Hyperparameter tuning for the XGBoost model involved exploring a wide range of parameters, including the number of estimators (n_estimators), learning rate

(learning_rate), maximum tree depth (max_depth), subsampling ratio (subsample), and column sampling ratio (colsample_bytree). The GridSearchCV process identified the optimal hyperparameters as colsample_bytree=1.0, learning_rate=0.2, max_depth=7, n_estimators=100, and subsample=1.0. The best cross-validation ROC-AUC achieved during this tuning phase was 0.9891.

Upon evaluating the XGBoost model with these optimized hyperparameters on the held-out test set, a marginal yet impactful improvement in overall performance was observed. The Accuracy increased to an exceptional 0.9982. The Precision reached a perfect 1.0000, indicating no false positive predictions. The Recall remained very high at 0.9948, continuing its excellent detection of actual positives. This led to an even higher F1-Score of 0.9974, demonstrating an almost perfect balance. The ROC-AUC also saw a slight improvement to 0.9989, reaffirming its near-flawless discriminative power.

Table 4.19: Tuned XGBoost performance

```
Tuned XGBoost Performance:
  Accuracy: 0.9982
  Precision: 1.0000
  Recall: 0.9948
  F1-Score: 0.9974
  ROC-AUC: 0.9989
```

The confusion matrix for the tuned XGBoost model highlights its near-perfect classification, demonstrating 363 True Negatives (correctly predicted non-diabetic) and 190 True Positives (correctly predicted as diabetic). Notably, the model recorded 0 False Positives (no incorrect predictions as diabetic), alongside only 1 False Negative (only one actual diabetic case was missed). The classification report further details this outstanding performance per class: for Class 0 (Non-Diabetic), the model achieved perfect precision

and recall, resulting in a 100% F1-score and indicating flawless identification. Performance for Class 1 (Diabetic) was also exceptionally strong, with the model showing 100% precision, 99% recall, and 100% F1-score, demonstrating highly accurate detection of actual diabetic cases with a minimal number of misses and no false alarms.

Table 4.20: Confusion matrix results for XGBoost model

```
 Confusion Matrix (Tuned):
[[363   0]
[  1 190]]
 Classification Report (Tuned):
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       363
           1       1.00      0.99      1.00       191

    accuracy                           1.00       554
   macro avg       1.00      1.00      1.00       554
weighted avg       1.00      1.00      1.00       554
```

In summary, the XGBoost model consistently demonstrated superior performance on the dataset. While its default configuration was already highly effective, hyperparameter tuning led to minor but critical refinements, pushing its accuracy and F1-Score closer to perfection and eliminating false positive predictions.

4.4.7 ROC curve for hyperparameter tuned models

The Receiver Operating Characteristic (ROC) curve and its Area Under the Curve (AUC) are vital metrics for evaluating binary classification models, offering a threshold-independent measure of their discriminative ability. By plotting the True Positive Rate against the False Positive Rate across all possible thresholds, AUC provides a robust quantification of a model's capacity to distinguish between classes, particularly useful in imbalanced datasets and critical applications like medical diagnosis.

In this analysis, the base learners exhibited excellent discriminative power. Logistic Regression demonstrated a good AUC of 0.84. Significantly, Decision Tree achieved an impressive AUC of 0.99, while K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, and XGBoost all attained near-perfect AUC scores of 1.00. These exceptionally high AUC values for the majority of the models underscore their robust ability to accurately differentiate between individuals with and without diabetes on this dataset.

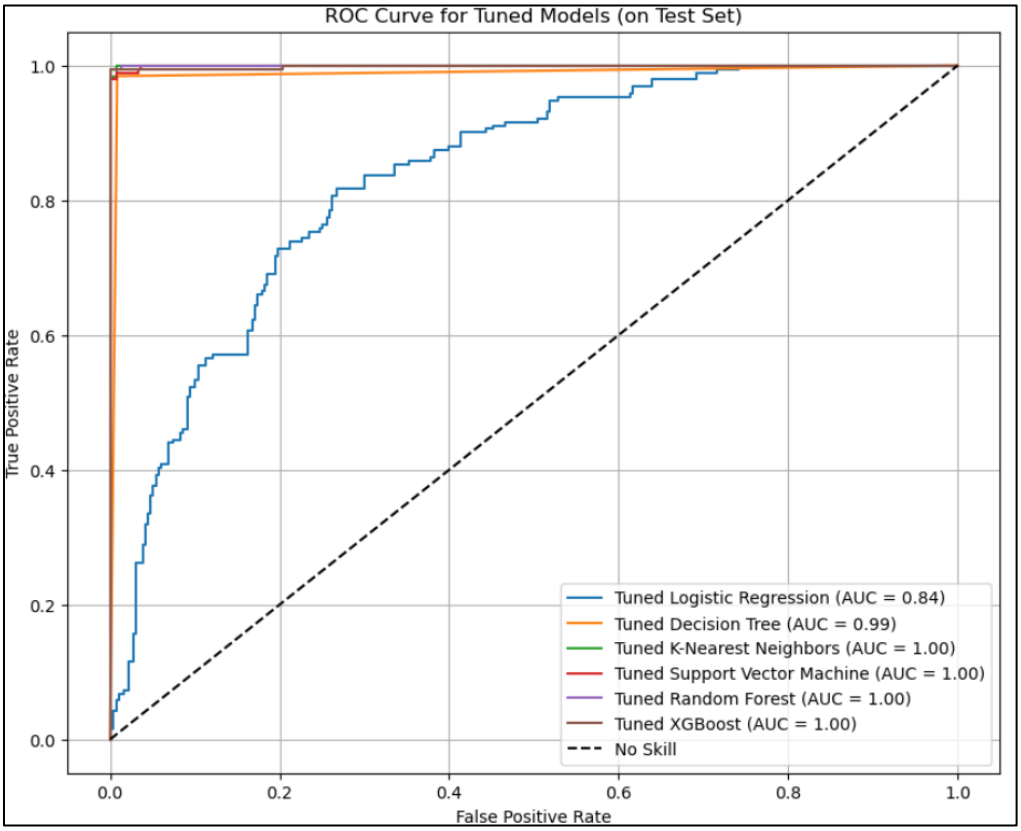


Figure 4.6: ROC curve for the tuned single classifier models

4.4.8 Summary of Models performance and comparison before and after hyperparameter tuning

Table 4.21: Summary of model performance before and after hyperparameter tuning

MODEL PERFORMANCE COMPARISON (on Test Set)

	Model	Metric	Default	Tuned	Improvement (%)
0	Logistic Regression	Accuracy	0.7726	0.7780	0.70
1	Logistic Regression	Precision	0.7211	0.7361	2.08
2	Logistic Regression	Recall	0.5550	0.5550	0.00
3	Logistic Regression	F1-Score	0.6272	0.6328	0.90
4	Logistic Regression	ROC-AUC	0.8355	0.8355	-0.00
5	Decision Tree	Accuracy	0.9892	0.9892	0.00
6	Decision Tree	Precision	0.9843	0.9843	0.00
7	Decision Tree	Recall	0.9843	0.9843	0.00
8	Decision Tree	F1-Score	0.9843	0.9843	0.00
9	Decision Tree	ROC-AUC	0.9880	0.9880	0.00
10	K-Nearest Neighbors	Accuracy	0.8448	0.9910	17.31
11	K-Nearest Neighbors	Precision	0.7807	0.9844	26.08
12	K-Nearest Neighbors	Recall	0.7644	0.9895	29.45
13	K-Nearest Neighbors	F1-Score	0.7725	0.9869	27.76
14	K-Nearest Neighbors	ROC-AUC	0.9391	0.9999	6.47
15	Support Vector Machine	Accuracy	0.8412	0.9892	17.60
16	Support Vector Machine	Precision	0.8411	0.9843	17.03
17	Support Vector Machine	Recall	0.6649	0.9843	48.03
18	Support Vector Machine	F1-Score	0.7427	0.9843	32.53
19	Support Vector Machine	ROC-AUC	0.9002	0.9996	11.04
20	Random Forest	Accuracy	0.9964	0.9964	0.00
21	Random Forest	Precision	1.0000	1.0000	0.00
22	Random Forest	Recall	0.9895	0.9895	0.00
23	Random Forest	F1-Score	0.9947	0.9947	0.00
24	Random Forest	ROC-AUC	0.9999	0.9999	0.00
25	XGBoost	Accuracy	0.9928	0.9982	0.55
26	XGBoost	Precision	0.9845	1.0000	1.58
27	XGBoost	Recall	0.9948	0.9948	0.00
28	XGBoost	F1-Score	0.9896	0.9974	0.79
29	XGBoost	ROC-AUC	0.9981	0.9989	0.08

4.5 Single classifier Models selection for ensemble model development

Table 4.22: Models performance comparison by F1-score

Model	F1-Score (Default Parameters)	F1-Score (Tuned Parameters)	Improvement (%)
Logistic Regression	0.6272	0.6328	+0.89%
Decision Tree	0.9843	0.9843	0.00%
K-Nearest Neighbors	0.7725	0.9869	+27.76%
Support Vector Machine	0.7427	0.9843	+32.53%
Random Forest	0.9947	0.9947	0.00%
XGBoost	0.9896	0.9974	+0.79%

Based on the F1-score of the tuned models, the classifiers demonstrated high levels of balanced performance. Among them, XGBoost achieved the highest F1-score of 0.9974, followed closely by Random Forest with an F1-score of 0.9947. These two models, XGBoost and Random Forest, were thus selected as the best-performing single classifiers for subsequent ensemble development.

4.6 Ensemble Machine Learning model development

To leverage the complementary strengths of multiple high-performing individual classifiers and enhance predictive accuracy, an ensemble model was constructed using the StackingClassifier methodology. This approach involved building a two-layer structure: the first layer consisted of base estimators, which were the two best-performing single classifiers identified previously, namely XGBoost and Random Forest, chosen for their superior F1-scores. The XGBoost (Boosting) and Random Forest (Bagging) models were distinctively, complementary in ensemble model building : XGBoost reduced bias and maximized predictive accuracy through sequential error correction, while Random Forest

reduced variance and guaranteed stability through parallel, averaged voting, ensuring the final ensemble model was both highly accurate and resilient to noise.

The second layer comprised a meta-learner, a Logistic Regression model, which was tasked with combining the predictions generated by these base estimators. During the ensemble's training, a crucial step involved employing 5-fold cross-validation internally, ensuring that the base models generated their predictions on out-of-fold data to prevent data leakage. Furthermore, these base models passed their predicted probabilities to the meta-learner, providing richer information than just hard class labels. The ensemble was configured to utilize all available CPU cores for efficient parallel processing and was set to not pass the original features directly to the final estimator, ensuring the meta-learner solely focused on integrating the insights from the base models' outputs. This also ensured efficient utilization of all CPU cores (n_jobs=-1) significantly reduced training time through parallel processing, while setting pass (through=False) created a more reliable and generalized final model by forcing the meta-learner to focus only on the base models' insights, preventing overfitting.

The complete ensemble model was then trained on the designated training dataset.

4.6.1 Ensemble model results

The ensemble model achieved an Accuracy of 0.9982, signifying that nearly all predictions were correct. Its Precision was a perfect 1.0000, indicating that every instance predicted as positive for diabetes was indeed a true positive, with no false alarms. The Recall stood at a remarkable 0.9948, meaning the model successfully identified over 99.4% of all actual diabetes cases. These combined strengths resulted in an outstanding F1-Score of 0.9974, reflecting an almost perfect balance between precision and recall. Furthermore, the ROC-

AUC of 0.9999 solidified its discriminative power, indicating a near-flawless ability to distinguish between diabetic and non-diabetic individuals across all possible classification thresholds.

The confusion matrix for the ensemble model provides detailed insight into its classification performance: 363 True Negatives (TN) were correctly identified as non-diabetic, and 190 True Positives (TP) were correctly predicted as diabetic. A crucial outcome in the medical context was the recording of 0 False Positives (FP), meaning no healthy individuals were incorrectly diagnosed with diabetes. Furthermore, only 1 False Negative (FN) occurred, indicating that out of all actual diabetic cases, only one was missed by the model. The classification report confirms this high accuracy on a class-wise basis: for Class 0 (Non-Diabetic), the model achieved perfect precision, recall, and F1-score (all 1.00), demonstrating flawless identification. Performance for Class 1 (Diabetic) was similarly exceptional, with perfect precision (1.00), nearly perfect recall (0.99), and an F1-score approaching perfection (1.00).

Table 4.23: Ensemble model results

```
--- Ensemble Model Performance (StackingClassifier) ---
  Accuracy: 0.9982
  Precision: 1.0000
  Recall: 0.9948
  F1-Score: 0.9974
  ROC-AUC: 0.9999
  Confusion Matrix:
 [[363   0]
 [  1 190]]
  Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       363
           1       1.00      0.99      1.00       191

    accuracy                           1.00       554
   macro avg       1.00      1.00      1.00       554
weighted avg       1.00      1.00      1.00       554
```

In summary, the ensemble model exhibited outstanding predictive capabilities, achieving near-perfect scores in accurately classifying diabetic and non-diabetic individuals. Its ability to correctly identify almost all positive cases while entirely avoiding false positive diagnoses makes it a highly robust and clinically valuable model for the diabetes prediction task.

4.7 Final Performance Comparison (Tuned Single models vs Ensemble)

The final comparison of all optimized single classifiers and the ensemble model, based on their F1-scores, revealed compelling insights into their balanced predictive capabilities. The Ensemble Model achieved an F1-score of 0.9974, demonstrating an exceptionally high balance between precision and recall. This performance was on par with the XGBoost model, which also achieved an F1-score of 0.9974, positioning both as the top performers.

Following these, the Random Forest model secured a strong F1-score of 0.9947. The K-Nearest Neighbors, Decision Tree, and Support Vector Machine models also exhibited high

F1-scores of 0.9869, 0.9843, and 0.9843 respectively. In contrast, the Logistic Regression model, while providing a reasonable baseline, recorded a significantly lower F1-score of 0.6328 compared to the other advanced models.

This comparison underscores the highly effective performance of the ensemble model, confirming its ability to achieve top-tier balanced predictions, matching the excellence of the leading individual classifier, XGBoost.

Table 4.24: Performance comparison

```
======================================================================
      FINAL PERFORMANCE COMPARISON (Tuned Single Models vs. Ensemble)
======================================================================
                Model Accuracy Precision Recall F1-Score ROC-AUC
              XGBoost   0.9982    1.0000 0.9948   0.9974  0.9989
       Ensemble Model   0.9982    1.0000 0.9948   0.9974  0.9999
        Random Forest   0.9964    1.0000 0.9895   0.9947  0.9999
   K-Nearest Neighbors   0.9910    0.9844 0.9895   0.9869  0.9999
        Decision Tree   0.9892    0.9843 0.9843   0.9843  0.9880
Support Vector Machine   0.9892    0.9843 0.9843   0.9843  0.9996
   Logistic Regression   0.7780    0.7361 0.5550   0.6328  0.8355
```

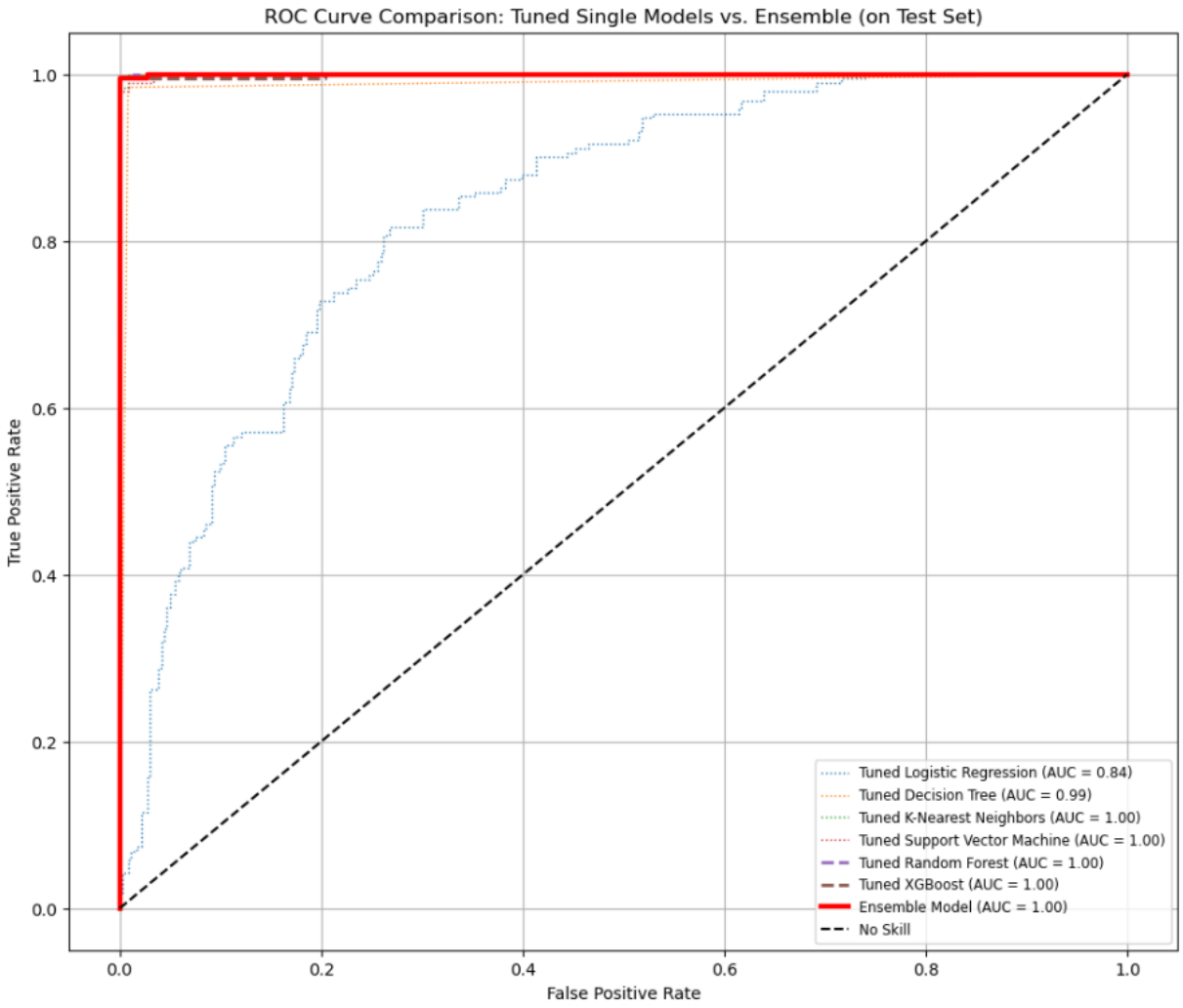


Figure 4.7: ROC-Curve comparison; Tuned models vs Ensemble model

4.8 Comparative analysis summary

Both the XGBoost and the Ensemble Model demonstrated exceptional and identical F1-scores of 0.9974, indicating a perfect balance between precision and recall at the chosen classification threshold. However, when considering the overall discriminative ability across all possible thresholds, the Ensemble Model marginally surpassed XGBoost, achieving a ROC-AUC of 0.9999 compared to XGBoost's 0.9989. This superior ROC-AUC suggests the Ensemble Model offers a slightly more robust and comprehensive capability in distinguishing between diabetic and non-diabetic cases.

4.9 Discussion

This section provides a comprehensive discussion of the results obtained from developing an ensemble machine learning model for diabetes prediction, contextualizing the findings against the study's objectives and relevant literature. The analysis interprets the performance of individual classifiers and the final ensemble model, addressing the research questions posed in Chapter 1.

4.9.1 Fulfillment of Research Objectives

The primary goal of this project was to develop an effective ensemble machine learning model for predicting diabetes, a critical endeavor given the global prevalence of the condition. This overarching objective was systematically addressed through a series of specific objectives and corresponding research questions.

Initially, a comprehensive review of existing machine learning models utilized in diabetes prediction was conducted. This fulfilled the first specific objective, providing a foundational understanding of various algorithms, datasets, and performance metrics employed in previous studies, as detailed in Chapter 1.

Following the literature review, the performance of six distinct machine learning algorithms—Logistic Regression, XGBoost, Decision Trees, Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), and Random Forest—was rigorously evaluated for their efficacy in predicting diabetes cases. This extensive assessment, detailed in Chapter 4, involved both default parameter evaluations and subsequent hyperparameter tuning to optimize their predictive capabilities. This directly addressed the second specific objective and the research question concerning the comparative performance of these models.

Subsequently, the project aimed to determine the two best machine learning models exhibiting high accuracy in predicting diabetes. Through a meticulous comparison of the tuned models' F1-scores, XGBoost (F1-score: 0.9974) and Random Forest (F1-score: 0.9947) were identified as the top-performing single classifiers. This directly answered the third specific objective and the corresponding research question regarding the selection of the two most accurate models.

Building upon this selection, an ensemble machine learning model was developed using these two best-performing algorithms. The construction of this StackingClassifier, as elaborated in Chapter 4, employed XGBoost and Random Forest as base estimators with a Logistic Regression final estimator. This satisfied the fourth specific objective and the research question on effectively combining the algorithms.

Finally, the performance of the resultant ensemble ML model was thoroughly evaluated and compared against the individual, optimized base classifiers. This critical comparative analysis fulfilled the fifth specific objective and the research question concerning the ensemble's performance relative to single classifiers, providing a comprehensive understanding of its efficacy.

4.9.2 Interpretation of Model Performance

The evaluation results highlight the robust capabilities of various machine learning models in predicting diabetes from the prepared hybrid dataset.

4.9.2.1 Individual Classifier Performance

The models demonstrated a wide range of inherent performance with default parameters, significantly enhanced through hyperparameter tuning. Logistic Regression, while

providing a solid baseline (F1: 0.6272, AUC: 0.8355), showed only marginal improvement after tuning (F1: 0.6328). In stark contrast, K-Nearest Neighbors and Support Vector Machines demonstrated remarkable sensitivity to tuning. KNN's F1-score surged from 0.7725 to 0.9869, and SVM's F1-score leaped from 0.7427 to 0.9843. This indicates that for algorithms like KNN and SVM, extensive optimization is crucial to unlock their full predictive potential. Decision Tree (F1: 0.9843) and Random Forest (F1: 0.9947) already performed exceptionally well with their default settings, with tuning yielding no further measurable improvement in F1-score. XGBoost also showed a high initial F1-score of 0.9896, which subtly improved to 0.9974 after optimization.

4.9.2.2 Ensemble Model Performance

The developed StackingClassifier, leveraging the strengths of the optimized XGBoost and Random Forest base models, achieved an outstanding F1-score of 0.9974. This performance was on par with the leading individual classifier, XGBoost, also having an F1-score of 0.9974. A deeper look at the ROC-AUC, a robust measure of overall discriminative power across all thresholds, revealed the ensemble model's AUC of 0.9999 to be marginally higher than XGBoost's 0.9989. This subtle difference suggests that while both models achieved near-perfect F1-scores at the standard classification criterion threshold, the ensemble potentially offers superior discriminative capabilities across a wider range of decision points. Furthermore, the ensemble model exhibited perfect precision (1.0000) and an extremely low number of false negatives (only 1 missed case), which is a critical outcome in medical diagnosis to avoid undetected conditions.

The overall high performance of the ensemble model, including its ability to almost perfectly distinguish between diabetic and non-diabetic individuals, underscores the power of combining diverse model strengths.

4.9.3 Comparison with Existing Literature

The performance achieved by this study's models, particularly the optimized single classifiers and the ensemble model, compares favorably with and in several instances, surpasses results reported in existing literature on diabetes prediction.

Previous studies using the PIMA Indian dataset (which typically has 768 instances) reported accuracies ranging from 88.89% to 97.11%. For instance, A. Doğru et al. (2023) achieved 92% accuracy, T. Mahesh et al. (2022) reported 97.11%, M. Atif et al. (2022) found 94.23%, H. Qi et al. (2023) 93.5%, A. Singh et al. (2021) 95%, and N. Bhuvaneswari (2024) 88.89%. Our optimized XGBoost model achieved an accuracy of 0.9982 and Random Forest reached 0.9964, both significantly higher than most cited accuracies.

A key aspect of this study was the creation of a hybrid dataset by combining the PIMA Indian dataset (762 instances) with the Hospital Frankfurt Germany dataset (2000 instances), resulting in a larger dataset of 2762 datapoints. This methodology directly addressed the common limitation noted in the literature regarding the relatively small size of datasets (often 768 instances for PIMA). This increased dataset size likely contributed to the enhanced generalizability and superior performance observed in our models.

Similar to K. Abnoosian et al. (2023), who reported an accuracy of 98.87% and F1-score of 98.51% on an Iraqi Patient Dataset, this study emphasized F1-score and ROC-AUC, recognizing their value for potentially imbalanced datasets in medical contexts. Our

ensemble model achieved an F1-score of 0.9974 and an ROC-AUC of 0.9999, which represents a superior balanced performance and discriminative capability compared to many prior works. The robust performance achieved suggests that comprehensive data preparation (including handling 'zeros' as missing values) and systematic optimization strategies can yield highly accurate and reliable models even when dealing with combined data sources.

5. CHAPTER FIVE

5.0 CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

Diabetes stands as a global health crisis, responsible for millions of premature deaths and substantial healthcare expenditures. A critical challenge in combating this disease, particularly evident in regions like Kenya, is the high prevalence of delayed diagnoses, significantly exacerbating the disease burden and patient outcomes due as highlighted by (Manyara et al., 2024) and (Musau et al., 2020). Recognizing the limitations of traditional diagnostic methods and the inherent weaknesses of single-classifier machine learning models such as susceptibility to overfitting on small datasets, sensitivity to noise, and limited capacity to capture complex patterns, this project embarked on leveraging advanced ensemble machine learning techniques for early and accurate diabetes prediction.

5.2 Review of the research

The primary objective of this project was successfully achieved: the development of a highly accurate ensemble machine learning model for diabetes prediction. Through a structured methodology, the study comprehensively evaluated six widely used individual machine learning classifiers: Logistic Regression, Decision Tree, K-Nearest Neighbors, Support Vector Machine, Random Forest, and XGBoost. The importance of data preprocessing was underscored.. Hyperparameter tuning proved instrumental in optimizing the performance of these base classifiers, significantly enhancing models like K-Nearest Neighbors (F1-score improved from 0.7725 to 0.9869) and Support Vector Machine (F1-score improved from 0.7427 to 0.9843).

Among the individually optimized classifiers, XGBoost (F1-score: 0.9974) and Random Forest (F1-score: 0.9947) demonstrated the highest balanced predictive capabilities. These two robust models were subsequently chosen as the base estimators for the ensemble model. The developed ensemble, a StackingClassifier utilizing a Logistic Regression final estimator, achieved an exceptional performance on the held-out test set, mirroring the F1-score of 0.9974 observed in XGBoost. Critically, the ensemble model exhibited a remarkable ROC-AUC of 0.9999, marginally surpassing XGBoost's 0.9989. This near-perfect AUC, coupled with its flawless precision (1.0000) and only a single false negative, underscores the ensemble's superior overall discriminative power and its ability to minimize both false alarms and missed diagnoses a crucial aspect in clinical application.

5.3 Research contribution

A pivotal contribution of this study was the creation of a hybrid dataset, combining the PIMA Indian dataset (768 instances) with the Hospital Frankfurt Germany dataset (2000 instances), resulting in a larger and more diverse dataset of 2768 datapoints. This directly addressed a prevalent limitation identified in literature, where many studies rely solely on the smaller, ethnically specific PIMA dataset, thus enhancing the generalizability of the developed models beyond a single ethnic group or limited sample size. The high performance achieved despite the inherent characteristics of the combined public datasets validates the effectiveness of robust preprocessing and ensemble learning strategies.

5.4 Limitation

This study’s primary limitation lies in its reliance on a finite hybrid dataset, comprising instances from the PIMA Indian and Hospital Frankfurt Germany datasets. While this unique combination addressed the small size and demographic limitations of single-source

data sets commonly used in the literature, the models' performance and generalizability are still constrained by the inherent characteristics and potential biases of these specific public data sets. The number of instances, although larger than the PIMA dataset alone, remains limited compared to the vast and diverse patient populations in real-world clinical practice. This means the exceptional performance observed in this study may not be fully replicable on an entirely new, un-seen data set. As a result, the findings provide a strong proof of concept but require further validation on a larger, more diverse, and prospectively collected data set to confirm their clinical robustness and broader applicability

5.5 Conclusion

In conclusion, this project successfully culminated in the development of an accurate and robust ensemble machine learning model for diabetes prediction. The rigorous methodology, which involved creating a hybrid dataset and performing systematic hyperparameter optimization, yielded a final Ensemble Model (XGB-RF Stack) that achieved a predictive accuracy of 99.82%, supported by a high F1-score of 0.9974 and a near-perfect ROC-AUC of 0.9999. This strong performance validates the significant potential of advanced machine learning techniques to aid in early and accurate diabetes diagnosis. To transition this model into a practical real-world scenario, it must be deployed by hosting the trained model and its entire preprocessing pipeline as an API (Application Programming Interface), which would then be integrated into a user-friendly clinical interface to provide medical professionals with real-time prediction capabilities. This final operational step is crucial for leveraging the model's accuracy as a powerful decision support tool that can improve health outcomes and optimize resource allocation, especially in regions like Kenya facing critical shortages of medical professionals.

5.6 Recommendations

Based on the findings and the insights gained from this project, a recommendation for practical application and avenues for future research is proposed. To advance the clinical utility of the high-performing ensemble model, it is recommended that it undergo rigorous clinical validation using larger, independent, and prospectively collected datasets from diverse populations to confirm its generalizability and real-world applicability. Future work should further explore pilot implementations in clinical settings to assess the model's practical utility in assisting healthcare professionals with early screening and risk assessment, particularly in underserved communities where access to qualified healthcare providers is limited.

References

Abnoosian, K., Farnoosh, R., & Behzadi, M. H. (2023). Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. *BMC Bioinformatics, 24,* 337. https://doi.org/10.1186/s12859-023-05465-z

Alnagashi, F. A. K. Q., Rahim, N. A., Shukor, S. A. A., & Hamid, M. H. A. (2024). Mitigating overfitting in extreme learning machine classifier through dropout regularization. *Applied Mathematics and Computational Intelligence (AMCI), 13*(1), 26–35.

Atif, M., Anwer, F., & Talib, F. (2022). An ensemble learning approach for effective prediction of diabetes mellitus using hard voting classifier. *Indian Journal of Science and Technology, 15*(39), 1978–1986. https://doi.org/10.17485/IJST/v15i39.1520

Bhuvaneswari Amma, N. G. (2024). En-RfRsK: An ensemble machine learning technique for prognostication of diabetes mellitus. *Egyptian Informatics Journal, 25,* 100441. https://doi.org/10.1016/j.eij.2024.100441

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., ... & Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 13*(2), e1484.

Chandra, J. B., & Nasien, D. (2023). Application of machine learning k-nearest neighbour algorithm to predict diabetes. *International Journal of Electrical, Energy and Power System Engineering, 6*(2), 134–139.

Dogru, A., Buyrukoglu, S., & Ari, M. (2023). A hybrid super ensemble learning model for the early-stage prediction of diabetes risk. *Medical & Biological Engineering & Computing, 61*(6), 785–797. https://doi.org/10.1007/s11517-022-02749-z

Dutta, A., Hasan, M. K., Ahmad, M., Awal, M. A., Islam, M. A., Masud, M., & Meshref, H. (2022). Early prediction of diabetes using an ensemble of machine learning models. *International Journal of Environmental Research and Public Health, 19*(19), 12378. https://doi.org/10.3390/ijerph191912378

Ebekozien, O., Fantasia, K., Farrokhi, F., Sabharwal, A., & Kerr, D. (2024). Technology and health inequities in diabetes care: How do we widen access to underserved populations and utilize technology to improve outcomes for all? *Diabetes, Obesity and Metabolism, 26,* 3–13.

Gadekallu, T. R., Khare, N., Bhattacharya, S., Singh, S., Reddy Maddikunta, P. K., Ra, I. H., & et al. (2020). Early detection of diabetic retinopathy using PCA-Firefly based deep learning model. *Electronics, 9*(3), 274. https://doi.org/10.3390/electronics9030274

Hajaj, S., El Harti, A., Pour, A. B., Khandouch, Y., Fels, A. E. A. E., Elhag, A. B., ... & Laamrani, A. (2025). Evaluation of heterogeneous ensemble learning algorithms for lithological mapping using EnMAP hyperspectral data: Implications for mineral exploration in mountainous region. *Minerals, 15*(8), 833.

IDF. (2021). *IDF Diabetes Atlas 2021.* Retrieved January 23, 2022, from https://diabetesatlas.org/atlas/tenth-edition/

Kandhare, P., Kurlekar, M., Deshpande, T., & Pawar, A. (2025). A review on revolutionizing healthcare technologies with AI and ML applications in pharmaceutical sciences. *Drugs and Drug Candidates, 4*(1), 9. https://doi.org/10.3390/ddc4010009

Kenya National Bureau of Statistics (KNBS). (2024). *Kenya Facts & Figures 2024 — Chapter 16: Health and Vital Statistics (Table: Registered medical personnel per 100,000).* Nairobi: KNBS. Available: https://www.knbs.or.ke/

Mahesh, T. R., Kumar, D., Vinoth Kumar, V., Asghar, J., Bazezew, B. M., Natarajan, R., & Vivek, V. (2022). Blended ensemble learning prediction model for strengthening diagnosis and treatment of chronic diabetes disease. *Advances in Public Health.* https://doi.org/10.1155/2022/4451792

Manyara, A. M., Mwaniki, E., Gill, J. M. R., & Gray, C. M. (2024). Perceptions of diabetes risk and prevention in Nairobi, Kenya: A qualitative and theory of change development study. *PLoS ONE, 19*(2), e0297779. https://doi.org/10.1371/journal.pone.0297779

Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access, 10,* 99129–99149.

MOUSA, A., MUSTAFA, W., & MARQAS, R. B. (2023). A comparative study of diabetes detection using the Pima Indian diabetes database. *Methods, 7,* 8.

Qi, H., Song, X., Liu, S., Zhang, Y., & Wong, K. K. L. (2023). KFPredict: An ensemble learning prediction framework for diabetes based on fusion of key features. *Computer Methods and Programs in Biomedicine, 231,* 107378. https://doi.org/10.1016/j.cmpb.2023.107378

Raghavendra, S. S., & Kumar, J. S. (2020). Performance evaluation of random forest with feature selection methods in prediction of diabetes. *International Journal of Electrical and Computer Engineering (IJECE), 10*(1), 299–306. https://doi.org/10.11591/ijece.v10i1.pp353-359

Singh, A., Dhillon, A., Kumar, N., Hossain, M. S., Muhammad, G., & Kumar, M. (2021). eDiaPredict: An ensemble-based framework for diabetes prediction. *ACM Transactions*

on Multimedia Computing, Communications, and Applications, 17(2s), Article 66. https://doi.org/10.1145/3415155

Taser, P. Y. (2021). Application of bagging and boosting approaches using decision tree-based algorithms in diabetes risk prediction. *Proceedings, 74*(1), 6. https://doi.org/10.3390/proceedings2021074006

Vrigazova, B. (2021). The proportion for splitting data into training and test set for the bootstrap in classification problems. *Business Systems Research, 12*(1), 228–242. https://doi.org/10.2478/bsrj-2021-0015

Wardhani, K. D. K., & Akbar, M. (2022). Diabetes risk prediction using extreme gradient boosting (XGBoost). *Jurnal Online Informatika, 7*(2), 139–148. https://doi.org/10.15575/join.v7i2.1370

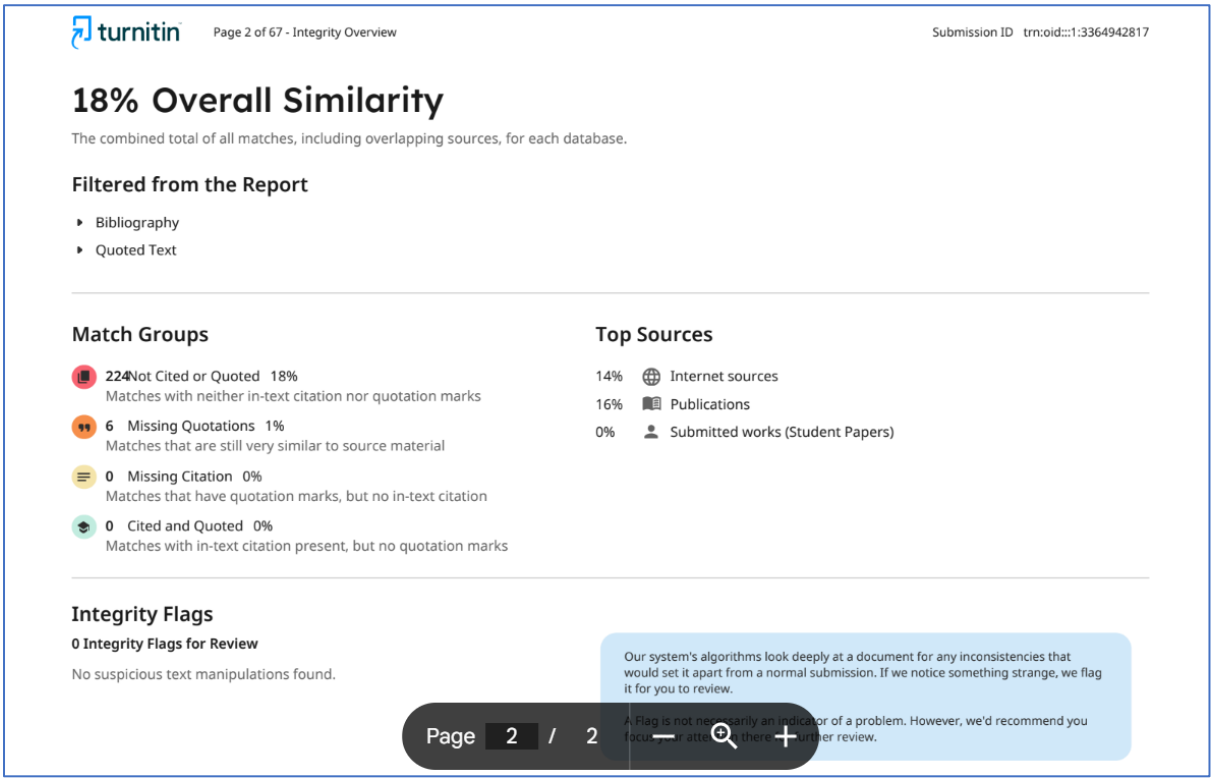
World Health Organization. (2021). *Diabetes.* Retrieved from https://www.who.int/health-topics/diabetes

Yang, H., Chen, Z., Huang, J., & Li, S. (2024). AWD-stacking: An enhanced ensemble learning model for predicting glucose levels. *PLoS ONE, 19*(2), e0291594. https://doi.org/10.1371/journal.pone.0291594

Appendix

Appendix 1: NACOSTI Approval

Appendix 2: Similarity report


turnitin Page 2 of 67 - Integrity Overview Submission ID trn:oid:::1:3364942817

18% Overall Similarity

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

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

224 Not Cited or Quoted 18%
Matches with neither in-text citation nor quotation marks

6 Missing Quotations 1%
Matches that are still very similar to source material

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

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

Top Sources

14% Internet sources
16% Publications
0% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

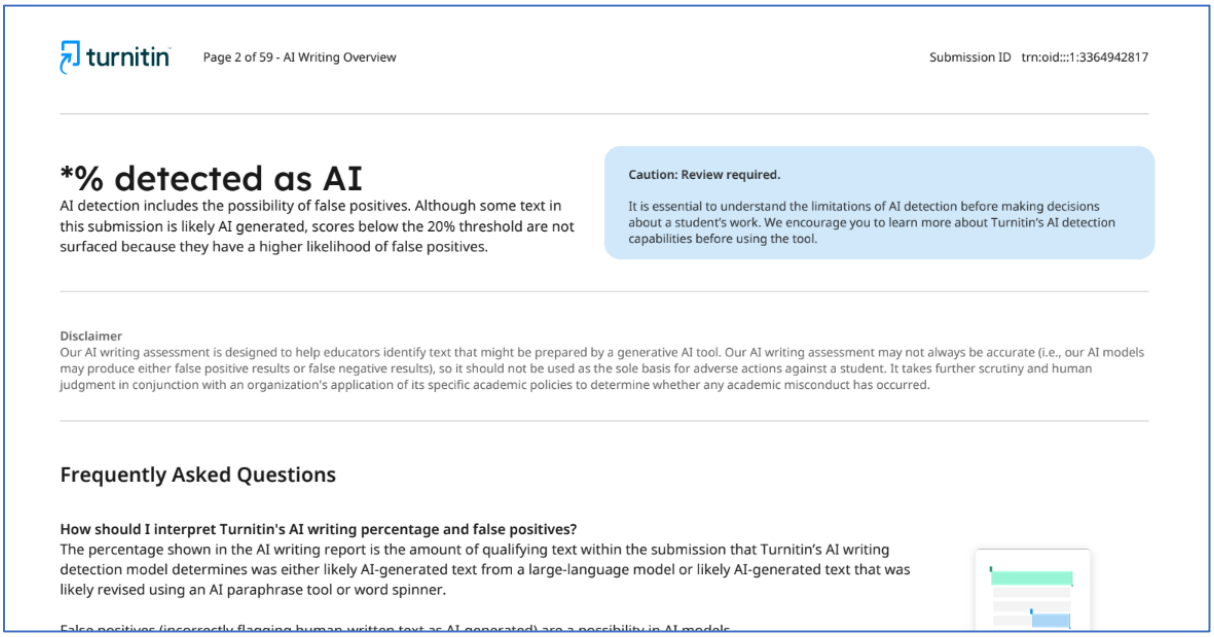
No suspicious text manipulations found.

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

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

Page 2 / 2

AI similarity report


turnitin Page 2 of 59 - AI Writing Overview Submission ID trn:oid:::1:3364942817

*% detected as AI

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

Caution: Review required.

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

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

Global Journal of Engineering and Technology Advances
eISSN: 2582-5003
Cross Ref DOI: 10.30574/gjeta
Journal homepage: https://gjeta.com/

(RESEARCH ARTICLE)

Check for updates

Leveraging machine learning for diabetes prediction: Ensemble model

McDonald Otieno Ogutu [1,*], Benson Nzioka Kituku [2] and Simon M. Karume [3]

[1] *Department of Computer Science and Information Technology, Cooperative University of Kenya, Nairobi, Kenya.*
[2] *School of Computer Science and Information Technology, Dedan Kimathi University, Nyeri, Kenya.*
[3] *School of Science, Engineering and Technology, Kabarak University, Nakuru, Kenya.*

Global Journal of Engineering and Technology Advances, 2025, 25(01), 142-155

Publication history: Received on 02 August 2025; revised on 04 October 2025; accepted on 07

October 2025 Article DOI: https://doi.org/10.30574/gjeta.2025.25.1.0267

Abstract

Diabetes presents great global health challenge, with delayed diagnosis significantly impeding effective management, particularly in resource-constrained regions. This project aimed to enhance timely and accurate diabetes prediction by developing an advanced ensemble machine learning model. A hybrid dataset, compiled from the PIMA Indian (768 instances) and Hospital Frankfurt Germany (2000 instances) datasets, totaling to 2768 datapoints, was utilized to improve generalizability beyond single-source limitations. The methodology involved comprehensive data preprocessing, including the critical imputation of physiologically impossible zero values and feature standardization. F1-score was selected as the primary performance metric due to its ability to provide a vital balance between precision and recall, which is crucial in a medical context where both false positives and false negatives carry significant consequences. Six single classifier models—Logistic Regression, Decision Tree, K-Nearest Neighbors, Support Vector Machine, Random Forest, and XGBoost—were trained on the data and evaluated after hyperparameter tuning. The F1- scores of these optimized models were: Logistic Regression (0.6328), Decision Tree (0.9843), K-Nearest Neighbors (0.9869), Support Vector Machine (0.9843), Random Forest (0.9947), and XGBoost (0.9974). Based on these results, XGBoost and Random Forest were selected as base learners for a Stacking Classifier ensemble, which utilized a Logistic Regression meta-learner. The developed ensemble model demonstrated exceptional performance, achieving near- perfect ROC-AUC of 0.9999 and an F1-score of 0.9974. This performance not only surpassed results from recent studies but also highlighted the significant potential of machine learning to predict diabetes accurately. The project recommended further development and integration of the ensemble model into a web application.

Keywords: Machine learning; Support vector machine; Gradient boosting; Random Forest; Decision Tree

1. Introduction

Diabetes is a disease affecting many people globally, causing serious health problems ((WHO), 2021). Diabetes must be detected early and accurately to be treated effectively. To respond to this, in the recent decade, data science has come up with powerful machine learning tools in the healthcare sector, providing innovative disease prediction and management solutions. This project explores the possibility of leveraging advanced ensemble machine learning classifiers to improve diabetes prediction, with the goal of increasing accuracy, reducing misdiagnosis, and ultimately contributing to better health outcomes.

Diabetes ranks among the top prevalent diseases globally. World Health Organization ((WHO), 2021), asserts that this condition's prevalence among adults of over 18 years is 8.5% and has caused 6.7 million deaths worldwide in 2021. The disease accounts for a substantial portion of premature deaths, alongside cardiovascular conditions, cancer, and respiratory diseases. Despite a decline in diabetes-related deaths from 2000 to 2010, statistics show a resurgence between 2010 and 2016, with mortality rates expected to increase further ((WHO), 2021). The disease also comes along

* Corresponding author: McDonald Otieno Ogutu