University of Kerbala
College of Computer Science & Information Technology
Computer Science Department

Prediction and Classification Model of Diabetes Using Machine Learning

A Thesis

Submitted to the Council of the College of Computer Science & Information Technology / University of Kerbala in Partial Fulfillment of the Requirements for the Master Degree in Computer Science

Written by
Aya Ahmed Hashim

Supervised by

Assist. Prof. Dr. Ayad Hameed Mousa

2025 A.D. 1446 A.H.

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

(وَقُلْ رَبِّ زِدْنِي عِلْمًا)

صدقَ اللهُ العليُّ العظيم

اية (114) سورة طه

Supervisor Certification

I certify that the thesis entitled (**Prediction and Classification Model of Diabetes Using Machine Learning**) was prepared under my supervision at the department of Computer Science / College of Computer Science & Information Technology / University of Kerbala as partial fulfillment of the requirements of the degree of Master in Computer Science.

Signature:
Supervisor Name: Assist. Prof. Dr. Ayad Hameed Mousa
Date: / /2025

The Head of the Department Certification

In view of the available recommendations, I forward the thesis entitled "**Prediction and Classification Model of Diabetes Using Machine Learning**" for debate by the examination committee.

Signature:
Assist. Prof. Dr. Muhannad Kamil Abdulhameed
Head of Computer Science Department
Date: / / 2025

Certification of the Examination Committee

We hereby certify that we have studied the dissertation entitled (**Prediction and Classification Model of Diabetes Using Machine Learning**) presented by the student (**Aya Ahmed Hashim**) and examined him/her in its content and what is related to it, and that, in our opinion, it is adequate with (**Excellent**) standing as a thesis for the Master degree in Computer Science.

Signature:
Name: Mustafa Jawad Radif
Title: Assist. Prof. Dr
Date: / / 2025
(Chairman)

Signature:
Name: Hayder Mohammed Ali
Title: Assist. Prof. Dr
Date: / / 2025
(Member)

Signature:
Name: Asam Hamed Abbas
Title: Dr.
Date: / / 2025
(Member)

Signature:
Name: Ayad Hameed Mousa
Title: Assist. Prof. Dr
Date: / / 2025
(Member and Supervisor)

Approved by the Dean of the College of Computer Science & Information Technology, University of Kerbala.

Signature:
Assist. Prof. Dr. Mowafak Khadom Mohsen
Date: / / 2025
(Dean of College of Computer Science & Information Technology)

Dedication

To those who supported me at this important stage, Assist. Prof. Dr. (Ayad Hameed Mousa). To those who are always by my side, my dear parents, my husband, my daughter, my family, my friends, and everyone who supported me. Thank you very much.

Acknowledgement

In the name of Allah, the Merciful and Most Gracious. Praise and thanks are due to Allah for His guidance, mercy, and blessing, for taking care of me every step of the way and helping me complete this thesis.

My heartfelt thanks to my supervisor Assist. Prof. Dr. Ayad Hameed Mousa for his assistance and continuous support. I appreciate his encouragement, kindness, guidance, and contribution, which provided a good basis for the success of this thesis.

My deepest gratitude goes to my family and my husband for their patience, support, and encouragement.

Abstract

Diabetes is one of the most prevalent chronic diseases globally, affecting millions and leading to serious complications such as heart disease, kidney failure, and vision loss. Early detection and proper classification of diabetes types (Type 1 and Type 2) are essential for effective treatment planning and long-term disease management. Early detection and prediction of diabetes can greatly improve patient outcomes, making it a global health concern. Manual patient data analysis is a common method of traditional diagnostic techniques, but it can be laborious and prone to human error. The research problem addressed here is the inefficiency and inaccuracy of traditional diagnostic methods, which this study aims to overcome using automated, data-driven approaches. This study investigates the use of clinical and demographic data in conjunction with machine learning (ML) techniques to predict and categorize diabetes. The study uses a variety of datasets features, such as lifestyle factors, biometric data, and clinical records, to train and assess different machine learning models, like decision trees, support vector machine (SVM), K-nearest neighbors (KNN), logistic regression, random forest, bagging, voting, Naïve bayse, XGBoost(Extreme Gradient Boosting), LightGBM (Light Gradient Boosting Machine), to create predictive models.

Three datasets used in this study include two well-known datasets sourced from publicly available repositories like the Diabetes Health Indicators containing 253,680 samples, the LMCH Diabetes dataset containing 1,000 patient samples, and a real-world dataset collected from the Imam Hassan Al-Mujtaba Diabetes and Endocrinology Center in Iraq, Karbala Governorate, containing 1,596 patient samples, includes features such as age, BMI, glucose levels, blood pressure, and insulin levels. Preprocessing techniques, including handling categorical variables,

handling missing values, removing duplicates, feature scaling, addressing class imbalance and feature selection are applied to enhance model performance. The models are evaluated using confusion matrix, accuracy, precision, recall, F1-score, and AUC. According to the results, Bagging achieved an accuracy of 89.27%, while KNN achieved an accuracy of 86.9%, which are the highest on Dataset 1. For Dataset 2, the decision tree achieved an accuracy of 99.50%. Using a real-world dataset, Random Forest and LightGBM achieved the highest accuracy of 99.79% in predicting diabetes. Additionally, the voting classifier achieved an accuracy of 95.74% in identifying the type of diabetes, confirming its effectiveness in classification tasks.

This study's innovation lies in combining real-world and large-scale datasets with advanced ensemble learning methods for both prediction and type classification. The results demonstrate that machine learning offers a scalable, efficient, and highly accurate solution for early diabetes diagnosis and classification, contributing valuable insights to the field of intelligent healthcare.

Declaration Associated with this Thesis

Parts of the work discussed in this thesis have been published or accepted, and are indicated below.

- The research article titled “An Evaluation Framework for Diabetes prediction techniques Using Machine Learning” has been published in “BIO Web of Conferences”.

- The research article titled “Predicting Diabetes and Identifying its Types Using Machine Learning Algorithms” has been accepted by “IEEE Xplore”.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Description
AI	Artificial intelligent
ANN	Artificial Neural Network
AROC	Area under the Receiver Operating Characteristic
AUC	Area Under the Curve
BMI	Body Mass Index
DL	Deep Learning
GA	Genetic Algorithm
GBM	Gradient Boosting Machine
KNN	K-Nearest Neighbors
LightGBM	Light Gradient Boosting Machine
LR	Logistic Regression
ML	Machine Learning
RF	Random Forest
ROC	Receiver-Operating Characteristic Curve
SVM	Support Vector Machine
XGBoost	Extreme Gradient Boosting

CHAPTER ONE

INTRODUCTION

1.1 Overview

For years, healthcare systems used traditional methods and expert judgment to estimate diabetes and categorize patients, typically through time-consuming and costly consults and laboratory assessments[1]. Due to the global increase in the prevalence of diabetes, it has become an essential requirement to classify and forecast various types of diabetes (Type 1, Type 2, and gestational diabetes) accurately for its early diagnosis and effective treatment. Type 1 is an autoimmune condition that typically begins in childhood, while type 2 is more common in adults and is often associated with lifestyle and insulin resistance. Accurately distinguishing between these two types is crucial as health professionals not only try to reduce load on medical staff and decrease treatment costs but also target improving patient outcomes through timely interventions [2]. To overcome these challenges, researchers have concentrated on creating data-driven models that utilize historical medical data and patient health records to improve diabetes prediction and classification accuracy [3].

Artificial Intelligence (AI), especially Machine Learning (ML) algorithms, has recently opened up the scene of medical diagnostics by providing robust solutions for predicting the onset of diabetes and classifying between types of diabetes. Such algorithms, including Support Vector Machines (SVM), Random Forest, and Gradient Boosting have shown promising results in addressing complex medical datasets with high accuracy[4].

One of the essential aspects to reliable predictions is data pre-processing, which includes missing value treatment, normalization and feature selection to enhance model performance [5]. Post preparation of data, models are trained and

validated to test their prediction accuracy, helping the healthcare professional deploy these models for early diabetes detection and management. This innovation has a lot to offer for improving healthcare prevention and care and is critical for addressing the world burden of diabetes[6].

Achieving high accuracy in diabetes prediction using machine learning techniques remains a significant challenge, as current models often face limitations in generalizability and clinical applicability. Moreover, recent research on the classification of diabetes types particularly distinguishing between Type 1 and Type 2 is relatively limited, highlighting the need for further studies focused on this area[7].

1.2 Problem Statement

Diabetes is a globally prevalent, lifelong metabolic disorder. Its negative impacts seriously affect peoples' health and well-being and it always develops as a result of late diagnosis and in addition to the missing or lack of robust systems for effective control and management. These negative implications may cause damage to other body systems such as Cardiovascular system, Nervous system, Kidneys, Eyes and Feet. In this context, an attention to study the effects of diabetes and the process to manage it is important. Several research works have been done to study diabetes and tried to predict its onset using artificial intelligence power. However, these studies generally gave different results based on the AI techniques used (e.g., machine learning). It means the prediction of diabetes is still requiring more efforts in terms of implementing different techniques to concrete the prediction capability. Furthermore, traditional methods to predict and classify diabetes using manual data analysis and labor tests is still a common practice which it is prone to human error. As the advancement of AI based technologies, a research work to predict diabetes

using AI techniques like machine learning is still a challenge. Therefore, this study is conducted to predict diabetes and classify its types using machine learning techniques.

1.3 Research Aim and Objectives

The main aim of this study is to propose a model that predict and classify diabetes using machine learning techniques. To accomplish this, the following specific objectives have to be achieved.:

1. To investigate the most effective machine learning algorithms for predicting diabetes.
2. To collect real world data in addition to the one from academic for prediction purposes.
3. To design, implement, and evaluate a ML-based model for diabetic prediction and classification.

1.4 Related Work

Diabetes is a chronic metabolic disorder characterized by elevated blood glucose levels resulting from either insufficient insulin production (Type 1) or ineffective insulin utilization (Type 2). Accurate prediction and classification of diabetes types are essential for developing personalized treatment plans, as each type requires different medical management strategies. Researchers have long been developing predictive models to support early diagnosis and reduce diabetes-related complications, ultimately advancing both scientific knowledge and public health. In recent years, various studies have employed artificial intelligence techniques, including machine learning (ML)

and deep learning (DL), to enhance predictive accuracy. These approaches have shown promising results, particularly in applying classification algorithms and integrating modern computational methods to improve diagnostic precision and clinical decision-making. A summary of the important relevant studies can be detailed in the following sections:

1. The researchers developed a machine learning-based system for classifying and predicting diabetes using the National Health and Nutrition Examination Survey dataset. They used logistic regression to identify significant risk factors, and then employed four machine learning classifiers (Naïve Bayes, Decision Tree, AdaBoost, and Random Forest) using three data partitioning protocols (K2, K5, K10). They evaluated performance based on accuracy, sensitivity, and AUC. Logistic regression identified seven primary risk factors, and the combination of logistic regression and Random Forest achieved the highest accuracy of 94.25% under the K10 protocol[8] .

2. The researchers proposed a robust framework for diabetes prediction by incorporating outlier rejection, missing value imputation, data standardization, feature selection, k-fold cross-validation, and several machines learning classifiers, including KNN, Decision Trees, Random Forest, AdaBoost, Naïve Bayes, and XGBoost. They introduced an additional ensemble method to enhance prediction accuracy, weighted by the corresponding AUC. All experiments were conducted on the Pima Indian Diabetes Dataset, where their ensemble classifier achieved sensitivity, specificity, false omission rate, diagnostic odds ratio, and AUC values of 0.789, 0.934, 0.092, 66.234, and 0.950, respectively—representing a 2.00% improvement over state-of-the-art results[9].

3. The researchers examined the impact of various parameters on diabetes data to predict the presence of the disease. They noted that certain factors increased the risk of chronic diseases such as diabetes and cardiovascular conditions. They developed a prediction model using an enhanced artificial neural network (ANN) trained via the artificial backpropagation scale conjugate gradient neural network (ABP-SCGNN) algorithm. They validated the model on the Pima Indian Diabetes dataset using accuracy and mean squared error (MSE) as metrics. Among several ANN configurations, the ABP-SCGNN model with 20 neurons achieved the highest accuracy of 93% [10].

4. The researchers aimed to develop effective machine learning classifiers for detecting diabetes in clinical data. They trained multiple algorithms—including Decision Tree, Naïve Bayes, KNN, Random Forest, Gradient Boosting, Logistic Regression, and SVM—on various datasets. They applied preprocessing techniques such as label encoding and normalization to improve model accuracy. They also integrated domain knowledge through baseline feature selection methods to identify key risk factors and compared model performance across datasets. Their proposed model demonstrated improved performance, with accuracy gains ranging from 2.71% to 13.13% over previous studies[11].

5. The researchers introduced a novel prediction method, Average Weighted Objective Distance (AWOD), based on the premise that individuals require tailored predictions due to diverse health conditions and risk factors. AWOD represented a modification of Weighted Objective Distance (WOD), incorporating information gain to better weight relevant individuals. They validated their method using two open-source datasets: the Pima Indians Diabetes Dataset (Dataset 1) and Mendeley data (Dataset 2), each with 392 records. Their experimental comparisons demonstrated that AWOD achieved

93.22% accuracy on Dataset 1 and 98.95% on Dataset 2 outperforming KNN, Random Forest, SVM, and Deep Learning approaches [12].

6. The researchers developed a GA-stacking ensemble learning model to enhance diabetes risk prediction accuracy. They used a genetic algorithm (GA) to select predictive features and applied Decision Tree, CNN, and SVM as base learners. The outputs were passed to a fully connected layer for final classification. Using a dataset from Qingdao that included diverse health indicators, they demonstrated that GA integration improved prediction efficiency and accuracy. The GA-stacking model outperformed other models in generalization and performance on early-stage diabetes datasets from the UCI repository[13].

7. The researchers emphasized the importance of recognizing diabetes symptoms to ensure accurate prediction. They analyzed diabetes diagnosis datasets using a hybrid machine learning approach combining SVM and Artificial Neural Networks (ANN). They also incorporated a fuzzy logic model, which utilized model outputs as inputs for final diagnosis. The dataset was divided into training (70%) and testing (30%) sets. Their fused model utilized real-time patient data and achieved a prediction accuracy of 94.87%, outperforming previously used methodologies[14].

8. The researchers presented a framework for designing a diabetes prediction model to support clinical diagnosis. They applied Spearman correlation for feature selection and polynomial regression for imputing missing values. They implemented various supervised machine learning models, including Random Forest (RF), SVM, and a novel twice-growth deep neural network (2GDNN), optimized via grid search and repeated stratified k-fold cross-validation. Their experiments showed that 2GDNN improved accuracy, achieving 97.25% on the

PIMA Indian dataset and 97.33% on the LMCH dataset, along with strong precision, sensitivity, and F1 scores[15].

9. The researchers aimed to develop an ensemble learning (EL)-based prediction system using predictive features and clinical outcomes. They tested the proposed EL approach built using Bayesian networks and radial basis functions against five machine learning techniques: Logistic Regression, Decision Tree, SVM, KNN, and Random Forest. Their results demonstrated the superiority of the EL method, achieving a maximum accuracy of 97.11%. This framework offered clinicians a comprehensive tool for accurate diabetes diagnosis and effective patient management [16].

10. The researchers applied machine learning techniques to predict diabetes using the Pima Indian Diabetes Dataset and a Kaggle dataset. They compared Support Vector Machine, Decision Forest, Linear Regression, and ANN models. Among them, ANN achieved the highest accuracy of 98.8%, indicating its reliability for early diabetes prediction. They also explored the potential use of the model in imaging analysis, assisting with the detection of ionized lactose patterns in both diabetic and non-diabetic patients[17].

11. The researchers proposed a novel Type 2 diabetes prediction framework called Health Edge, integrated with an IoT-edge and cloud computing system. Using real-world diabetes datasets, they compared the performance of Random Forest (RF) and Logistic Regression (LR). Experimental results demonstrated that RF consistently outperformed LR, achieving 6% higher average accuracy. The framework was designed to function as a smart healthcare engine, enabling earlier diagnosis and improving patient outcomes[18].

12. The researchers developed consolidated frameworks for diabetes diagnosis in women aged 21–81 using minimal data and innovative preprocessing methods.

They applied data augmentation, attribute analysis, and missing value imputation. Using Shapley Additive Explanations (SHAP), they identified glucose, age, and BMI as key predictors. They implemented four models(Random Forest, Extra Tree, AdaBoost, and XGBoost) on a unique dataset. XGBoost and AdaBoost demonstrated the highest performance, achieving 94.67% accuracy and F1 scores of 95.27 and 95.95, respectively, validating their effectiveness in diabetes prediction[19] .

13. The researchers proposed a multiclass approach for detecting and classifying diabetes using highly imbalanced data from the Laboratory of Medical City Hospital. They prepared an imbalanced dataset for evaluation and applied three machine learning classifiers (SVM, Logistic Regression, and KNN) along with feature selection techniques (filter, wrapper, embedded) to enhance performance. They further optimized the models using 10-fold cross-validation. Notably, SVM using the top four features (filter method) achieved 96.4% accuracy, 96.8% precision, and an AUC of 0.99[20].

14. The researchers used a cleaned Kaggle dataset from 2015, which contained 253,680 survey responses from the CDC's BRFSS. They focused on the target variable “diabetes” and 21 associated features. They applied Chi-square tests to assess the relationships between variables and diabetes, followed by training various machine learning models. The CatBoost Classifier emerged as the best-performing model, yielding an accuracy of 86.6%. Permutation Feature Importance analysis for Logistic Regression revealed General Health, BMI, Age, High Blood Pressure, and High Cholesterol as the top five predictive features[21].

15. The researchers developed a hybrid diabetes risk detection model called DiabML, leveraging the Artificial Intelligence of Medical Things (AIoMT).

They combined the Binary Whale Optimization (BWO) algorithm with machine learning techniques. BWO was used for both feature selection and addressing data imbalance with SMOTE during preprocessing. Their simulations demonstrated that DiabML outperformed existing models. Specifically, it achieved an 86.1% classification accuracy using the AdaBoost classifier, highlighting its potential for effective diabetes risk prediction[22].

Table 0.1 Summary of Related Works

Authors And Year	**Classifier Used (method type)**	**Accuracy Obtained**	**Outcome**
[8] 2020	combination of LR and RF-based classifier	94.25%	Diabetes Prediction
[9] 2020	Ensemble classifier (AdaBoost, XGBoost)	95%	Diabetes Prediction
[10] 2021	artificial back propagation scaled conjugate gradient neural network (ABPSCGNN)	93%	Diabetes Prediction
[11] 2021	Decision tree (DT), Support Vector Machine (SVM).	96.81%	Diabetes Prediction
[12] 2021	average-based weighted objective distance (AWOD)	93.22% 98.95%	Diabetes Prediction
[13] 2022	GA-stacking	98.71%	Diabetes Prediction
[14] 2022	fused ML	94.87	Diabetes Prediction
[15] 2022	twice-growth deep neural network (2GDNN) model	97.25%, 97.33%	Diabetes Prediction

[16] 2022	EL based-on Bayesian networks	97.11%	Diabetes Prediction
[17] 2023	Artificial Neural Network	98.8%	Diabetes Prediction
[18] 2023	Random Forest	78.27% PIDD 97.23% Sylhet	Diabetes Prediction
[19] 2023	Xgboost and Adaboost	94.67%	Diabetes Prediction
[20] 2024	SVM	96.4%	Diabetes Prediction
[21] 2024	Cat Boost	86.6%	Diabetes Prediction
[22] 2024	AdaBoost	86.1%	Diabetes Prediction

As shown in Table 1.1, although the related studies demonstrate good performance in predicting diabetes, there are several limitations. High-accuracy diabetes prediction remains a challenge, and studies classifying diabetes subtypes are virtually nonexistent. The use of different datasets limits comparability, and most models focus primarily on accuracy without addressing data imbalances, interpretability, or clinical applicability. Future research should prioritize generalizability, transparency, and integration in healthcare settings.

1.5 Thesis Organization

The thesis comprises five chapters. Each chapter is preceded by a brief background that outlines the contributions made within the chapter and provides an overview of how its results contribute to the overall direction of the study:

Chapter 1: provides a part of a very important objective introduction for the readers regarding the different points and methodologies that supported the research, and it provides Problem Statement Summary. It also substantially introduces the most important objectives and aspirations of this study. It mainly covers the important basic contribution and things that this study will append to the scientific and medical field. This chapter also covers the most relevant preceding studies, the datasets used, the AI algorithms used, and the accuracy of the proposed model in these studies.

Chapter 2: This chapter presents the theoretical framework of the thesis, leading the reader to the advantages of artificial intelligence from its beginnings to its advanced stages so far and its reflection in the field of prediction. In addition, a comprehensive explanation of the machine learning algorithms and techniques applied to the proposed model is also provided.

Chapter 3: In this chapter it shall be possible to know by what means the proposed model. The steps required to implement the proposed model are performed. This chapter combine all modules from the stage of selection datasets to the stage of datasets pre-processing and stages of applying ML algorithms. Finally, the reader will be able to compare the results obtained from each particular usage of AI technologies and feel in his own hands what is the real accuracy achieved.

Chapter 4: One of the important points that is explained in this chapter, the obtaining of the results with the proposed model more dependent on the kind of dataset used and the machine learning method degree of efficiency, these points showed in fourth chapter of what is presented in the thesis.

Chapter 5: This chapter is a conclusion of everything mentioned in the thesis, and it focuses on what can be relied upon in future studies, the most fundamental and practical scientific advantage that the research presents.

CHAPTER TWO

THEORETICAL BACKGROUND

2.1Overview

Today, artificial intelligent (AI) represents a wide array of technologies that help machines replicate human-like thinking and do what typically requires cognitive functions like learning, reasoning, and decision-making. These fields include natural language processing, robotics, computer vision etc.

Machine Learning (ML): ML is a subfield of AI that focuses in making machines learn using statistical techniques to enhance their performance on a task through experience[23]. In this approach, models are built that learn from data. There are three types of Machine Learning:

- Supervised learning: where a model is trained on labeled data to make predictions.
- Unsupervised learning: means finding patterns in data that is not labeled.
- Reinforcement learning: It is the process of learning the best action taken from the trial-and-error process.

This study utilizes supervised machine learning with an emphasis on classification approaches, as it is well-suited for labeled medical data such as diabetes diagnosis and type. Real-valued and continuous features are used to train algorithms that are evaluated based on how accurately their outputs match known outcomes. This approach enables effective pattern recognition, supports clinical decision-making, and provides a straightforward way to assess each model's predictive performance.

2.2 Preprocessing

Preprocessing is an important step in machine learning where raw data has converted into a model-understandable format by cleaning and transforming before entering it into machine learning algorithms. Preprocessing the data involving handling categorical variables, handling missing values, removing duplicates, Feature scaling, and handling imbalanced data. By cleaning and transforming the data into a structured format, it helps in providing the model with high-quality input, thereby enhancing its performance and minimizing the chances of errors. Preprocessing is a critical step in machine learning; it plays a significant role in discovering meaningful patterns in the data and generalizing to new, unseen data[24]. In the following subparagraphs, some of the techniques used for preprocessing have highlighted according to the context of this study.

2.2.1 Handling Categorical Variables

Many machine learning algorithms need input in the form of numbers, so managing categorical variables is a key component of data preprocessing. For categorical variables (those that represent distinct groups), it is possible to use one-hot encoding (A binary column is created for each category) or label encoding (A unique integer is assigned to each category). Another method is ordinal encoding, which can be helpful when the categories have an order to them. Given that the improper handling of categorical features can result in inaccurate forecasts and misleading outcomes, the selection of categorical handling/encoding methods is highly dependent on the type of categorical data available and the algorithm being used[25].

2.2.2 Handling Missing Values

Handling missing values is one of the important data preprocessing steps in machine learning. There are many causes of missing data such as errors when collecting data or not responding to surveys. There are common techniques for treating missing values like imputation where the missing values are filled with estimated values (mean, median, or mode etc.) or go ahead and remove rows or columns having missing values in the context of this study the mean strategy will be used. Method selection for this criterion should reflect both the proportion of null values and the effect the null values might have on the output of the analysis. Handling of missing values appropriately helps ensure that the model is able to predict correctly without being influenced by incomplete information[26].

2.2.3 Removing Duplicates

Removing duplicates is a state of art during the preprocessing of data because duplicates, if exists in data they may introduce bias while it comes to modeling and make machine learning model predictions low when it comes to accuracy. Duplicate records may arise from errors in data collection, merging datasets, or data entry errors. Duplication detection and removal ensures that the information fed into the model is unique, preventing it from learning from repeated information. Standard methods include utilizing drop duplicates () in pandas or other data-cleaning methods to spot and remove duplicate rows or records according to certain conditions. Handling duplicates correctly ensures data integrity and better model performance[27].

2.2.4 Feature Scaling

Normalizing and standardizing are techniques to scale numerical features in a similar range or distribution. Normalization usually scales the data to a range (such as 0 to 1) by subtracting the min value and dividing by the range (max – min). While standardization, on the other hand, rescales each feature such that it has the properties of a standard normal distribution with a mean of 0 and a standard deviation of 1, in the context of this study standardization will be used. Model types that are affected by scale will benefit from these techniques distance-based models (like k-NN, SVM) or gradient-based models (like neural net), as feature scaling yields improved convergence and model performance[24].

2.2.5 Handling Imbalanced Data

Imbalanced class distribution is a common scenario in problem domains and is a huge challenge for Machine Learning. As a result, this can create biased models that tend to learn the majority class leading to low performance when predicting the minority class. Usually, to overcome this problem, you can apply resampling methods, i.e., oversampling and undersampling the expensive classes or using class weights to impose a penalty in favor of the lowest class during the model fitting. Furthermore, some of the sophisticated techniques such as ensemble methods and anomaly detection can be used for improving the model performance and to allow fair representation of both classes, in the context of this study RandomOverSampler technique will be used. Imbalanced data refers to one class significantly outweighing the others in a dataset. Proper handling of this issue leads to more accurate and robust models that generalize well across all classes[28].

2.2.6 Feature selection

Feature selection is an essential process in machine learning and data mining that encompasses discovering as well as selecting the most appropriate features for generating predictive models. Feature selection increases model performance in addition to tackling overfitting issues and reducing computation cost by reducing the number of input variables. Feature selection is one of the important techniques that undertake a few methods: filter, wrapper, and embedded. In the context of this study the correlation will be used which is a filter method. These approaches assess the importance of each feature according to various factors, like statistical testing's, model performance or algorithm-based evaluation. Feature selection is a crucial technique for improving interpretability of the model as well as it can generalize better on unseen data [29].

2.3 Models applied

To perform a thorough and effective comparison, ten machine learning algorithms were tested, including both individual models and ensemble techniques (which combine multiple models to achieve higher accuracy than a single one). These algorithms were selected because they represent core and widely used approaches in diabetes prediction, offering a balance between interpretability, accuracy, and computational efficiency. The chosen models (Decision Tree, SVM, KNN, Logistic Regression, Random Forest, Bagging, Voting, Naïve Bayes, XGBoost, and LightGBM) cover a broad spectrum of machine learning methods, from simple linear models to complex ensemble learners. This diversity allows for a comprehensive evaluation of which techniques perform best on the diabetes prediction and classification tasks. The theoretical background of these algorithms

will be discussed in the following sections in a way that aligns with the goals of this study and benefits the reader.

2.3.1 Decision Tree

Decision tree is one of the most used machine learning algorithms for both classification and regression and provide a model of decisions and the possible consequences, which is relatively easy to understand and interpret[30].

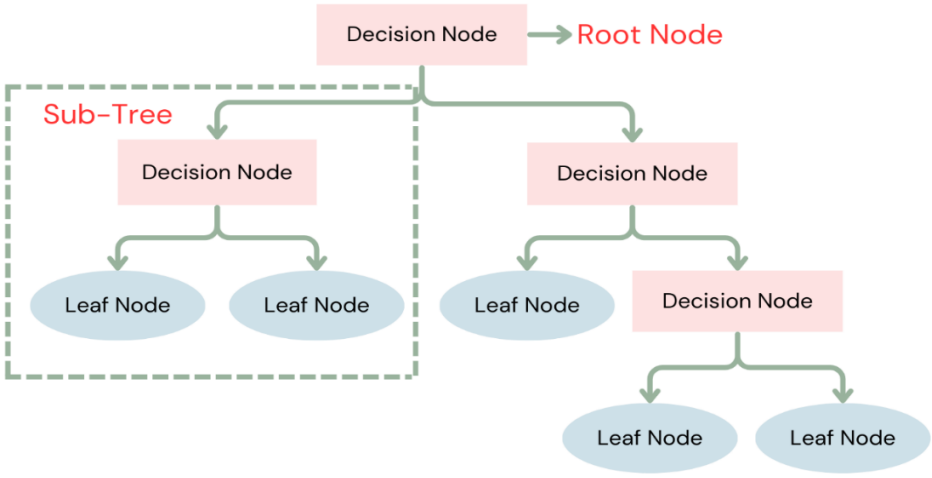

Figure 2.1 Simple Decision Tree.

In a decision tree, the data is split according to certain criteria. The name given to the root point is the root node, intermediate point which data split from called decision nodes and finally leads to provide us the output are the leaf nodes[31]. Decision trees are built top down, starting with root node, the dataset is divided into subsets using a feature that has maximum information gain or minimum Gini impurity. Important factors in the splitting process include metrics such as Gini

impurity, which measures the impurity of a dataset and lower values indicate better splits, and entropy, which determines the degree of disorder in the dataset and defines information gain. Pruning techniques (e.g., cost-complexity pruning) ⊆ may be introduced to limit parts of the tree that offer minimal predictive value (overfitting). Finally, for classification of new data, it is done by tracking a path from root node to leaf node according to the values of the specific features[32].

2.3.2 Support vector machine (SVM)

SVM is a widely adopted supervised learning algorithm known for its effectiveness in classification and regression tasks. It operates by identifying the optimal hyperplane that maximizes the margin between distinct classes, improving generalization and reducing overfitting. To handle non-linearly separable data, SVM employs various kernel functions—such as linear, polynomial, and radial basis function (RBF)—to transform data into higher-dimensional feature spaces for better separation[33]. Recent research has focused on enhancing SVM's scalability and efficiency for large datasets through kernel approximation methods and parallel computing[34]. The regularization parameter CCC in SVM plays a crucial role in balancing the trade-off between maximizing the margin and minimizing classification errors, making the model robust against noisy data [35]. Due to these advancements, SVM continues to be effective in areas such as text classification, image recognition, and bioinformatics.

2.3.3 K-Nearest Neighbors (KNN)

KNN algorithm is a supervised learning algorithm that calculates similarity between instances in the training data (known instances) and the new

data (test instances) based on the distance between them. So, the primary steps of KNN algorithm are:

1. Choose the parameter k, number of neighbors.
2. Reformulated: [One can also calculate the Euclidean distance between the new data to each point in the training set.
3. Find the nearest neighbors and their respective distances.
4. Labeling has done through k nearest neighbor
5. Final classification will be based on majority label of that neighbors

2.3.4 logistic Regression

Logistic Regression is a statistical method commonly used for binary and multi-class classification tasks owing to its simplicity and efficiency. Its high-level function applies the logistic (sigmoid) function to a linear combination of input features to these lines with n classes containing values from 0 to 1*[36]*. Regularization methods, like L1 (Lasso), L2 (Ridge), were developed to boost the performance of logistic regression and prevent overfitting, increasing its generalization on high-dimensional datasets *[37]*. Moreover, logistic regression has been combined with feature selection algorithms and ensemble techniques to address complex data distributions and imbalanced datasets in various applications, such as bioinformatics, text classification, and financial prediction*[38]*. Its interpretability and scalability provide a foundational framework for predictive modeling across many disciplines.

2.3.5 Naïve Bayes

Naive Bayes (NB) also known as Bayes' theorem is a supervised learning method. It is a widely used machine learning method that is efficient

and simple because it is an optimization- based approach assuming that all features are not related and independent to each other. NB makes the assumption that the categorization of one feature in a class has no effect on the classification of the other feature. it is a very powerful classification algorithm as it classifies the dataset according to Bayes' rule of probability[39, 40]. Using the Naïve Bayes theorem, it can be computed according to the equation:

$$\Pr(A|B) = \frac{\left(\Pr(B|A)\Pr(A)\right)}{\left(\Pr(B)\right)} \quad (1)$$

Where Pr(A) = Prior probability of class, Pr(B|A) = Likelihood probability of (B) conditional on (A), Pr(A|B) = Posterior probability of class, Pr(B) = Prior probability of (B).

2.3.6 Random Forest

Random Forest is a popular ensemble learning method that builds several decision trees and aggregates their predictions to improve accuracy and generalization. It works by constructing a bunch of decision trees at training time and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. The advantage of this approach is that it helps to prevent overfitting and enhances the generalizability of the model[41, 42]. Earlier works emphasize parallelization and online learning approaches. Moreover, some studies have investigated the combination of Random Forests with other machine learning approaches to improve accuracy and interpretability. Such advancements have cemented Random Forest as a staple functionality in machine learning, utilized across fields from bioinformatics to finance to remote sensing[43].

2.3.7 Bagging

Bagging, or bootstrap aggregation, is an ensemble method that creates multiple training datasets to benefit model creation[44]. In bagging, multiple training sets are created from the original dataset using randomly sampled (with replacement) subsets. After these different training sets are constructed, several models are trained on each resampled set in an ensemble approach. Finally, the individual predictions are combined to come to a final prediction. This effectively adds variance to the model during training, which can help combat overfitting. Bagging mainly consists of three processes: bootstrapping, parallel training, and result aggregation[45]. Some of the benefits of bagging are as follows: reduces variance by averaging multiple models for more stable predictions. It also improves accuracy on complex datasets by combining model strengths. Another benefit of bagging is its ability to avoid overfitting, which is particularly helpful when working with high-variance algorithms such as decision trees. With a balanced stability, accuracy, and robustness for predictive modeling.[46].

2.3.8 XGBoost

Gradient boosting with XGBoost stands for eXtreme Gradient Boosting, which is a fast and accurate machine learning algorithm for supervised learning challenges. Specifically, it is an efficient version of a boosting-based machine learning framework that was specifically built to boost the speed of computational progress and the precision of machine learning algorithms. XGBoost works by fitting several decision trees one after the other, where the new tree works on minimizing the prediction error of its earlier trees, thus minimizing model bias and variance[47]. Some key features of XGboost are: it

incorporates regularization techniques (L1 and L2 regularization) to help in preventing overfitting and improving the generalization of the model. It also uses a sparse aware algorithm for efficient missing value imputation, which makes PP-LDA resilient to real-world situations consisting of incomplete datasets. XGBoost also implements parallel and distributed computing so it scales well with larger datasets. During this time, the time evolution of applications has shown the objective range of MLP — the versatility of it applicable to directly in relation to a family of classification, regression, or even ranking problems, etc. With its performance being so verifiable by the success of the algorithm in machine learning competitions and a lot of practical applications in industries[48].

2.3.9 Light Gradient Boosting Machine (LGBM)

LGBM is a powerful gradient boosting framework that is based on a decision tree-based learning algorithm. Using top precooled trees with the most perfect fit, LGBM is used primarily around rating and classification during this process, multiple data improvement techniques are implemented, and the calculation is often measured (validated) by the variance after performing value partition [49]. An example equation is:

$$Y1=\text{Base Tree}(X1) - lr1*\text{Tree1}(X1) - lr1*\text{Tree2}(X1) \quad (2)$$

the value indicates how the decision tree algorithm can be used for splitting the dataset and implementation of value which shows the number of trees based on instances of dataset[50]. LGBM provides faster performance compared with traditional gradient boosting, with many unique parameters, which help enhance or modify the efficiency according to use[51]**.**

2.3.10 Voting

Voting is an ensemble learning method in machine learning that combines the decisions of multiple classifiers to achieve better results. Voting combines outputs across different models in theory utilizing strengths of each individual model while yielding superior accuracy and robustness than a single classifier[52]. There are primarily two types of voting — hard voting and soft voting. Hard voting: each model votes for a class label, the label with the most votes is selected. Soft Voting: Predict well in all models and take an average of the probability of each class and take the class with the maximum average probability as the final output. When diversity of the models is obtained this way, it is likely to obtain weak error correlation, which, in turn, improves this aggregation approach[53]. The synthesis of voting classifiers is possible through different algorithms and frameworks; hence those frameworks possess tools to build hard/soft voting ensembles. The use of voting classifiers has been shown to be very effective in several different domains like text classification, image recognition, bioinformatics, etc. and hence form an important tool in the machine learning practitioner's toolkit[54].

2.4 Evaluation Measures

Evaluation metrics are used to determine how effective a model is. There are many different ways in which a model can be judged. Multi-class classification is one of the fundamental tasks in machine learning when there are more than two classes. There are quite a few performance indicators that can be helpful while assessing and comparing various models/techniques that can be used to handle multi-class classification problems. Furthermore, these metrics supply useful feedback during the development process like when comparing the performances of two models, or tuning different parameters of a single model.

The following subsections describe many of the metrics available for performance evaluation.

2.4.1 **Confusion Matrix**

sample data table that shows item by item the rating comparison of two raters in terms of true/actual classifications and estimated classifications. The order of classes in rows and columns should be the same such that the correct classifications lie across the major diagonal from top left to bottom right it is used to quantify the accuracy for the proposed approach. The confusion matrix has four outcomes: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The efficiency can be evaluated and analyze the algorithms in the proposed model using the metrics through this matrix. These metrics are Accuracy, Precision, Recall, F1-Score and AUC-ROC (Area Under the Receiver Operating Characteristic Curve).

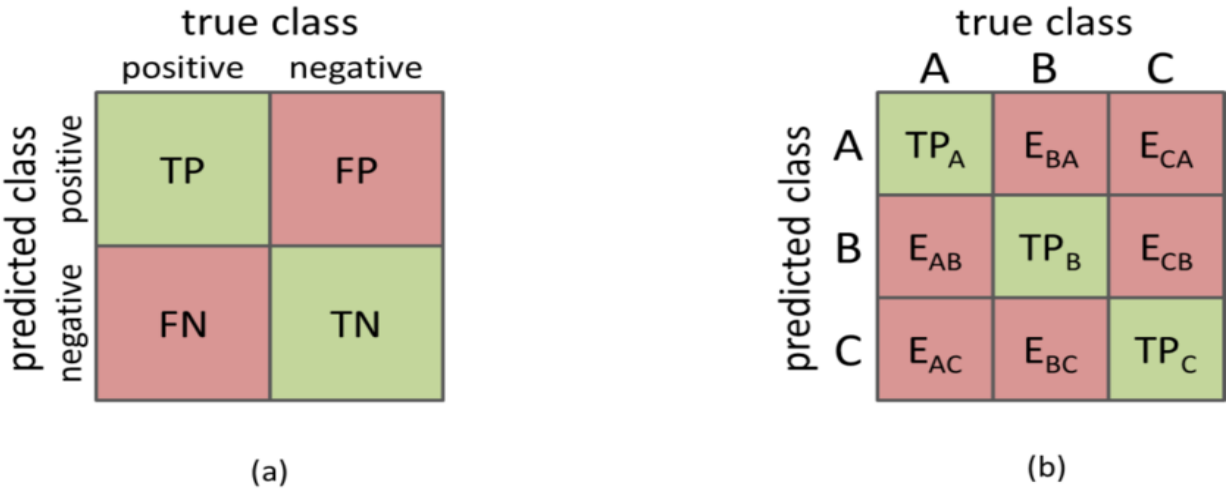

Figure 2.2 Confusion matrix. Illustration (a) is a confusion matrix for binary classification (2 x 2) and illustration (b) is a confusion matrix for multi-class classification (3 x 3)[55]

- **Accuracy**

Accuracy is the major criteria used for evaluation, representing the overall performance of the classifier, which is the ratio of the total number of accurate predictions to the total number of predictions, and is defined as follow.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \qquad (3)$$

- **Precision**

Precision is the number of positive predictions that were true divided by the total number of positive predictions.

Figure 2.3 Confusion matrix. Illustration (a) is a confusion matrix for binary classification (2 x 2) and illustration (b) is a confusion matrix for multi-class classification (3 x 3)

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (4)$$

- **Recall**

Recall is defined as the number of positive predictions to the actual total number of positive predictions.

$$\text{Re}call = \frac{TP}{TP + FN} \qquad (5)$$

- **F1-Score**

F1-Score is the weighted average of precision and recall. As a result, this score considers both false positives and false negatives.

$$\text{F1-score} = 2 * [\frac{(\text{Recall} * \text{Precision})}{\text{Recall} + \text{Precision}}] \qquad (6)$$

- **AUC-ROC**

AUC-ROC is an important metric in the field of machine learning, especially for binary classification tasks. It assesses the ability of a model to separate classes by plotting the true positive rate against the false positive rate across

multiple threshold values. AUC = 1.0 means a perfect classification, while AUC = 0.5 is equivalent to random guessing. Recently, many developments have generalized AUC maximization to multiple classes, making it useful in various fields[56].

2.4.2 K-Fold Cross Validation

K-fold cross-validation is a commonly used technique in machine learning for assessing the performance of models by dividing the dataset into K subsets, or "folds." In each iteration, the model is trained on K-1 folds and evaluated on the left fold, this process is repeated K times (each of the K folds serves as a test set once). Next, the average of all K iterations is computed, resulting in a more robust estimate of the model accuracy. Cross-validation, on the other hand, is a technique to evaluate the model and ensure it is not overfitting to a single train-test split, giving an idea of how the model will generalize to an independent dataset. K-fold cross-validation is ideally used when the datasets are not very large so that every data point is used for training and testing. Recent research further confirms its capacity to enhance the reliability and performance of our models[57].

2.4.3 Early Stopping

Early Stopping for machine learning models is a regularization technique to avoid overfitting while training the model. Instead of continuing to fit the model to the training data, it monitors the model on a validation set, and stops training once performance starts to degrade. This prevents overfitting, meaning the model must be able to generalize effectively to novel data it has never encountered previously. The concept of early stopping is a commonly used approach as part of many machine

learning applications and performs well alongside techniques such as cross-validation, where model hyper-parameters must be tuned[58].

CHAPTER THREE

THE PROPOSED MODEL

3.1Overview

This study aims to propose a framework for predicting diabetes and classifying the type of diabetes. This strategy works to predict diabetes by identifying people with the disease and people without the disease, and the type of diabetes is then determined, whether it is type 1 or type 2. This model consists of several successive steps, as shown in Figure 3.1.

The first step in building a model is to select and collect datasets. The second step is to obtain a suitable dataset ready to work on, where preprocessing is used on the data. In the third stage, the three datasets are enacting eight algorithms in this model. The fourth Stage is Track Several evaluations Metrics to measure the efficiency and accuracy of the algorithms used. The fifth and final step, based on the results obtained, the best algorithm used for this model to provide a prediction of diabetes is concluded.

In the third dataset after diabetes prediction, all the above steps are used to classify diabetic patients according to the type of disease, whether it is type 1 or type 2.

3.2 Dataset Used

This study utilized three datasets. Two were integrated into the model to enhance the accuracy and efficiency of diabetes prediction compared to existing studies. The third dataset, manually collected by the researcher, was used for both predicting diabetes and classifying its type.

The first dataset (Diabetes Health Indicators Dataset) was taken from Kaggle[59]. It is a clean data of 253,680 survey responses to CDC's BRFSS2015 (The Centers for Disease Control and Prevention) (Behavioral Risk Factor

Surveillance System). This dataset provides a wide range of health indicators related to diabetes and lifestyle, making it unique from commonly used datasets and valuable for training accurate and generalizable machine learning models. The target variable (Diabetes_012) consists of 3 classes. 0 for no diabetes or diabetes only during pregnancy, 1 for prediabetes and 2 for diabetes. There is class imbalance in this dataset. It has 21 features variables in this dataset as shown in Table 3.1 and Table 3.2.

Table 3.1 Features Description of Dataset1

No	Attribute Name	Descriptions
1	HighBP	high Blood pressure
2	HighChol	high cholesterol
3	CholCheck	cholesterol check in 5 years
4	BMI	Body Mass Index
5	Smoker	In your life, have you smoked less than 100 cigarettes?
6	Stroke	Did you have a stroke?
7	HeartDiseaseorAttack	coronary heart disease (CHD) or myocardial infarction (MI)
8	PhysActivity	Except for work, have you been physically active in the last month?
9	Fruits	Eat fruit once or more daily
10	Veggies	Eat vegetables once or more a day
11	HvyAlcoholConsump	Alcoholics Anonymous (adult men who have more than 14 drinks during the week and adult women who have more than 7 drinks during the week)
12	AnyHealthcare	Do you have any type of health coverage, such as health insurance, prepaid plans like HMO, etc.

13	NoDocbcCost	Has there been a period in the last 12 months when you needed to see a doctor but were unable to do so because of the cost?
14	GenHlth	Is your health generally? : scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
15	MentHlth	How was your mental health, including stress, depression, and emotional problems, over the past month? Scale from 1 to 30 days
16	PhysHlth	How many days during the past month has your physical health been poor, including illnesses and injuries? Scale: 1 to 30 days
17	DiffWalk	Do you complain of pain or difficulty walking or climbing stairs?
18	Sex	0 = Female 1 = Male
19	Age	Age group 13 level 1 = 18-24 9 = 60-64 13 = 80 or older
20	Education	6 levels of education 1 = Never attended school or kindergarten only 2 = Elementary 3 = Some high school grades 4 = High school graduate 5 = Technical schools 6 = College graduate
21	Income	Income Scales 1 = Less than $10,000 5 = Less than $35,000 8 = $75,000 or more

Table 3.2 A Part of the Dataset 1

	Diabetes_	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	Diseaseor	PhysActivit	Fruits	Veggies	HvyAlcoh	AnyHealth	NoDocbcC	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
2	0	1	1	1	40	1	0	0	0	0	1	0	1	0	5	18	15	1	0	9	4	3
3	0	0	0	0	25	1	0	0	1	0	0	0	0	1	3	0	0	0	0	7	6	1
4	0	1	1	1	28	0	0	0	0	1	0	0	1	1	5	30	30	1	0	9	4	8
5	0	1	0	1	27	0	0	0	1	1	1	0	1	0	2	0	0	0	0	11	3	6
6	0	1	1	1	24	0	0	0	1	1	1	0	1	0	2	3	0	0	0	11	5	4
7	0	1	1	1	25	1	0	0	1	1	1	0	1	0	2	0	2	0	1	10	6	8
8	0	1	0	1	30	1	0	0	0	0	0	0	1	0	3	0	14	0	0	9	6	7
9	0	1	1	1	25	1	0	0	1	0	1	0	1	0	3	0	0	1	0	11	4	4
10	2	1	1	1	30	1	0	1	0	1	1	0	1	0	5	30	30	1	0	9	5	1
11	0	0	0	1	24	0	0	0	0	0	1	0	1	0	2	0	0	0	1	8	4	3
12	2	0	0	1	25	1	0	0	1	1	1	0	1	0	3	0	0	0	1	13	6	8
13	0	1	1	1	34	1	0	0	0	1	1	0	1	0	3	0	30	1	0	10	5	1
14	0	0	0	1	26	1	0	0	0	0	1	0	1	0	3	0	15	0	0	7	5	7
15	2	1	1	1	28	0	0	0	0	0	1	0	1	0	4	0	0	1	0	11	4	6

The second dataset sourced from “Mendeley Data" were gathered in the Iraqi society (the Medical City Hospital laboratory and (the Specializes Center for Endocrinology and Diabetes-Al-Kindy Teaching Hospital)[60]. The diabetes dataset was constructed by extracting data from patients’ files that were taken and entered in to the database. The data include medical data and lab analysis. It consist of 1000 patients and three classes (Diabetic, Non-Diabetic, and Predicted-Diabetic) and have 13 attributes shown in Table 3.3 and Table 3.4.

Table 3.3 Features Description of Dataset2

No	Attribute Name	Descriptions
1	No. of Patient	Unique identifier for each patient
2	Gender	Biological sex of the patient (M=Male/F=Female)
3	Age	Patient’s age in years
4	Urea	Waste product in blood reflecting kidney function
6	Cr	Creatinine ratio Indicator of kidney function
7	HBA1C	Average blood sugar levels over 2-3 months
8	Chol	Cholesterol (Chol) Total cholesterol level in blood.
9	TG	Triglycerides (TG) Type of fat stored in the body.
10	HDL	"Good" cholesterol, helps remove LDL
11	LDL	"Bad" cholesterol, linked to heart disease
12	VLDL	Precursor to LDL, carries triglycerides
13	BMI	Body Mass Index

Table 3.4 A Part of the Dataset 2

1	ID	No_Pation	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	CLASS
2	502	17975	F	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24	N
3	735	34221	M	26	4.5	62	4.9	3.7	1.4	1.1	2.1	0.6	23	N
4	420	47975	F	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24	N
5	680	87656	F	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24	N
6	504	34223	M	33	7.1	46	4.9	4.9	1	0.8	2	0.4	21	N
7	634	34224	F	45	2.3	24	4	2.9	1	1	1.5	0.4	21	N
8	721	34225	F	50	2	50	4	3.6	1.3	0.9	2.1	0.6	24	N
9	421	34227	M	48	4.7	47	4	2.9	0.8	0.9	1.6	0.4	24	N
10	670	34229	M	43	2.6	67	4	3.8	0.9	2.4	3.7	1	21	N
11	759	34230	F	32	3.6	28	4	3.8	2	2.4	3.8	1	24	N
12	636	34231	F	31	4.4	55	4.2	3.6	0.7	1.7	1.6	0.3	23	N
13	788	34232	F	33	3.3	53	4	4	1.1	0.9	2.7	1	21	N
14	82	46815	F	30	3	42	4.1	4.9	1.3	1.2	3.2	0.5	22	N
15	132	34234	F	45	4.6	54	5.1	4.2	1.7	1.2	2.2	0.8	23	N

The third dataset was collected manually from the Imam Hassan al-Mujtaba Center for Diabetes and Endocrinology in Karbala, Iraq, under the medical supervision of specialists in this disease. The dataset was collected and recorded over a period of four months. This dataset was entered into the model for use in predicting diabetes. A separate set of data, containing information on those with the disease, was then used to build a model to classify the type of diabetes, whether it was Type 1 or Type 2.

As mentioned, this dataset was collected by researchers and consists of 1596 samples and contains 16 characteristics as shown in Table 3.5 and Table 3.6. This dataset contains a set of lab test results related to the disease for healthy people as well as for people with diabetes in addition to a column showing the person with diabetes from the non-diabetic (Outcome) and the last column shows the type of diabetes whether it is type 1 or 2(Type).

Table 3.5 Features Description of Dataset3

No	Attribute Name	Descriptions
1	Gender	Biological classification (Male/Female)
2	Family history	Presence of diabetes in family members (0=yes /1=no)

3	Age	Patient's age in years
4	BMI	Body weight index indicating obesity risk
5	HbA1C	Average blood sugar over 2–3 months
6	Fasting Blood Sugar	Blood glucose level after fasting.
7	Random Blood Sugar	Blood glucose level at any time.
8	Tchol	Total Cholesterol (Tchol) Overall cholesterol in blood.
9	Trig	Triglycerides (Trig): Type of fat in blood
10	Urea	Waste product indicating kidney function.
11	Creatinine	Kidney function marker.
12	TSH	Thyroid stimulating hormone (TSH) regulating thyroid function.
13	AST	Aspartate aminotransferase (AST) Liver enzyme indicating liver health
14	ALT	Alanine transaminase (ALT) Liver enzyme linked to liver function.
15	Systolic	Upper blood pressure reading.
16	Diastolic	Lower blood pressure reading.

Table 3.6 A Part of the Collected Dataset

	Gender	amily histo	Age	BMI	HbA1c	g blood gl	m blood g	Tchol	Trig	urea	Creatinine	TSH	AST	ALT	Systolic	Diastolic	Outcome	Type
2	Female	1	15	19.8	8.1	188	230	215	114	24	0.5	0.7	13	14	110	60	1	0
3	Female	0	39	25.5	9.9	278	295	162.6	130	21.69	0.6	1.66	29.1	24.33	110	70	1	1
4	Male	1	20	20.25	5.8	77	140	155.26	88.96	19	1.03	2.9	19.78	21.69	110	70	0	NULL
5	Female	0	15	29.7	10.6	287	536	103	100	NULL	NULL	1.09	NULL	NULL	110	70	1	0
6	Female	0	31	22.4	10.9	222	415	218.3	126.4	NULL	0.8	1.53	30.7	20.9	100	70	1	1
7	Male	0	12	16.06	6	100	90	118.93	86	19	0.91	6.34	20	10.52	130	90	0	NULL
8	Male	0	12	27.32	6.2	78	126	193.45	77	21	1.04	3.13	29	12.73	110	90	0	NULL
9	Male	1	20	20.84	4.8	88	130	145.62	66.49	13	0.88	2.37	NULL	NULL	120	80	0	NULL
10	Female	0	23	26.6	8.1	248	298	167.8	111.7	24.1	0.49	2.93	31.51	25.95	100	90	1	1
11	Female	0	27	19.13	6.5	83	140	154.71	77.61	15	0.6	3.03	22.09	20	100	70	0	NULL
12	Male	0	38	27.32	4.5	80	90	NULL	57.11	18	0.94	NULL	23.15	21.93	170	100	0	NULL
13	Male	0	15	12.6	16.8	125	228	NULL	NULL	23	0.5	NULL	17	37	110	80	1	0
14	Female	0	12	17.9	4	95	158	NULL	111	21	0.68	3.02	19	16.6	110	70	0	NULL
15	Male	0	12	21.58	4.5	80	145	177.13	108	22	0.71	5.26	26	20.98	130	90	0	NULL

3.2 The Proposed Model

The model shown below in Figure 3.1 was used to predict diabetes and classify its type, and 3 different types of data were used. Each data base is entered into the model separately. Next is a detailed description of the proposed model.

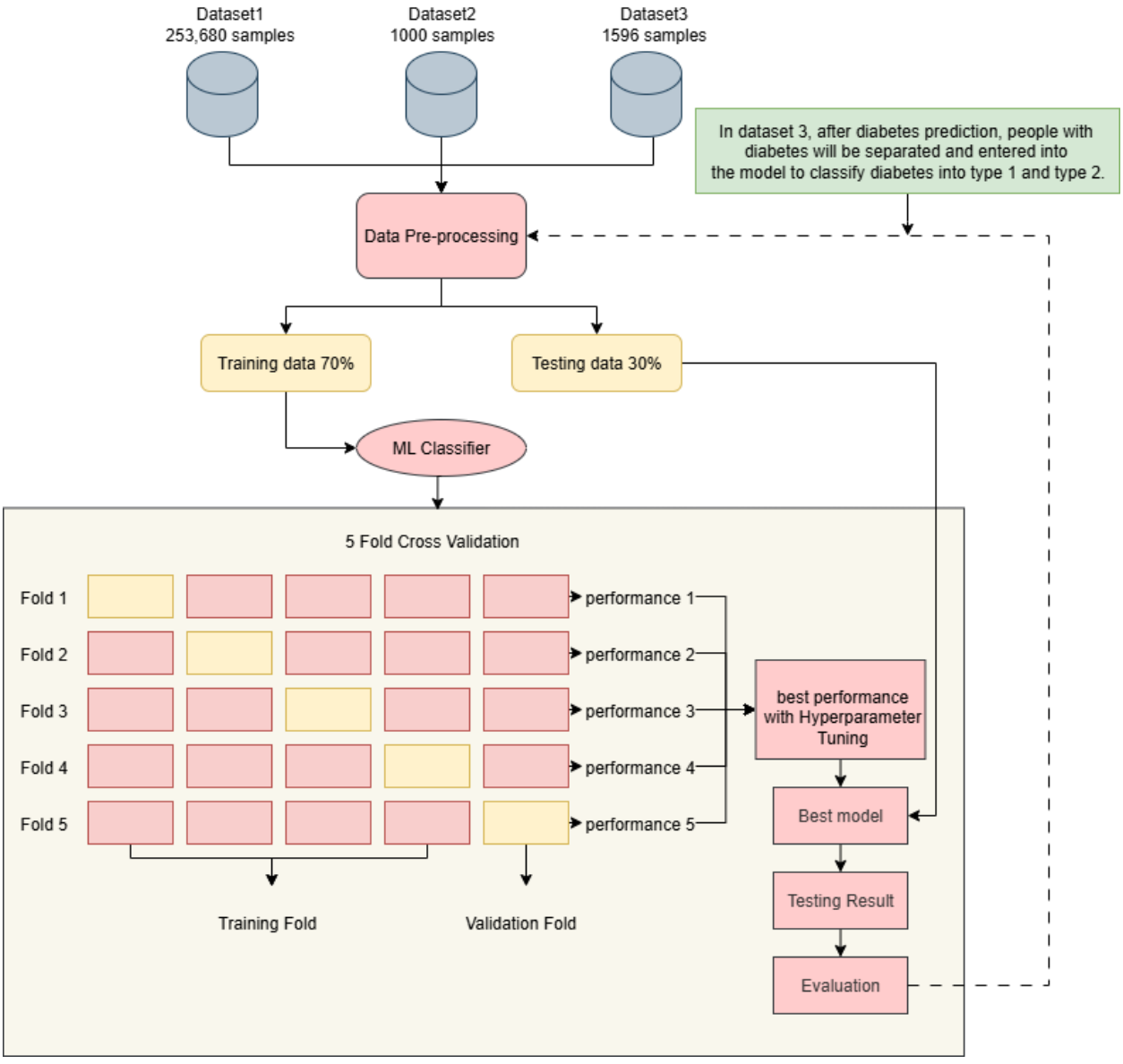

Figure 3.1 Proposed Model for diabetes prediction

3.2.1 Pre-processing

All necessary steps have been taken for data preprocessing as mentioned in chapter2 section 2.3 for all parts of the proposed model. The following are the steps involved in preprocessing:

- **Handling Categorical Variables**

Machine learning models need numeric input. On-hot encoding converts categorical data into a format that models can understand. It was used to convert categorical variable Gender (M/F) in dataset 2 and Gender (Male/Female) in dataset 3 into binary (0/1) columns.

- **Handling Missing Values**

Identifying missing values is important because they can negatively impact model performance. Understanding the extent and distribution of missing data helps determine an imputation strategy. Imputation of missing values completes the dataset, allowing models to train more efficiently. The technique used is mean imputation in dataset 3, which replaces missing values with the mean of each column. Mean imputation was used because the simple and powerful method is very efficient in compacting the overall distribution over numerical features. Other datasets do not contain missing data.

- **Removing Duplicates**

Removing Duplicates means identifying and eliminating repeated entries from your dataset, which could affect the integrity of your data and therefore the performance of your model. In our proposed model this preprocessing step was used in (dataset1 and dataset2) except dataset3 because it is free of duplicates.

- **Feature Scaling**

Feature scaling ensures that all features are equally involved in the model learning process. The technique used is StandardScaler which standardizes features by removing the mean and scaling to unit variance using Fit-Transform to compute the mean and standard deviation on the training data, and applying the scaling.

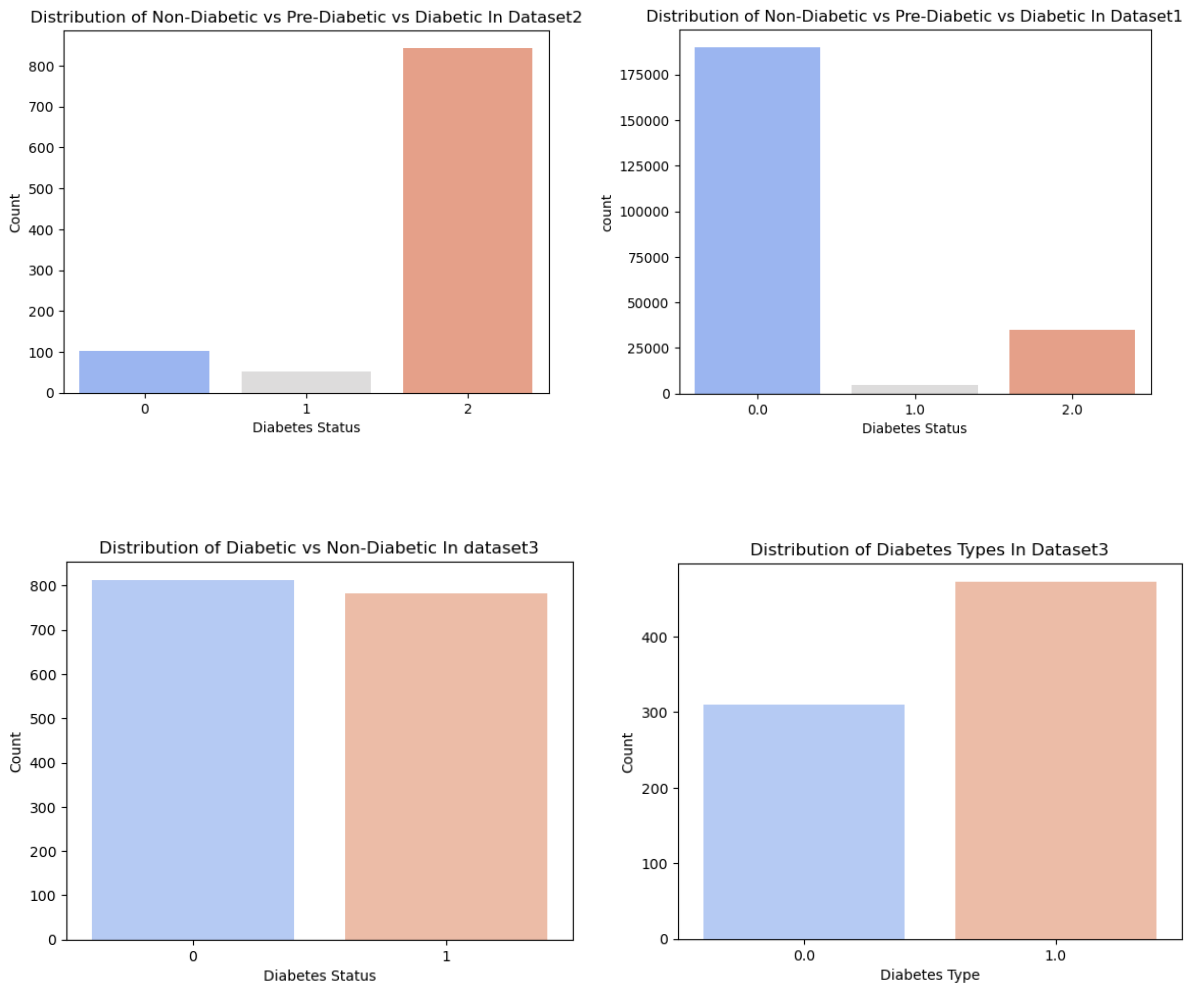

Figure 3.2 Class Distribution of The Three Datasets Used in The Proposed Model

- **Handling Imbalanced Data**

Handling imbalance data use to reduce the difference between class distribution so that models can perform better and more fair. In our proposed model (RandomOverSampler) was used for all datasets which balances an imbalanced dataset by randomly replicating instances of the minority class. This increases the number of minority samples to match the majority class, helping models learn equally from all classes you can see the class distribution in the Figure 3.2 above.

- **Feature Selection Based on Correlation**

To see how the attributes are related and to get a comprehensive view of the dataset used, a heatmap of the correlation matrix was used. This tool is provided by the Seaborn Python library. It is a two-dimensional grid with colored squares. The colors are categorized based on the power of the correlation. The stronger the correlation, the darker the color, and conversely, the weaker the correlation, the lighter the color. The range of the values in the cells of this matrix is between (-1 and 1), the closer this value is to the positive value, the stronger the correlation, and the closer it is to the negative value, the weaker the correlation between the attributes.

After calculating the correlation for all the attributes and based on these values, a number of features are selected in such a way that the correlation between them is less and the correlation with the output is greater. In dataset 1, the features selected based on correlation were ['HighBP', 'HighChol', 'CholCheck', 'BMI', 'HeartDiseaseorAttack', 'PhysActivity', 'GenHlth', 'PhysHlth', 'DiffWalk', 'Age', 'Education', 'Income'] and in dataset 2, the features selected were ['AGE', 'HbA1c', 'Chol', 'TG', 'BMI', 'BMI_Age']. For dataset 3, the features selected for predicting

diabetes were ['HbA1c', 'Random blood glucose', 'Tchol', 'urea', 'Creatinine', 'TSH', 'Diastolic']. For classifying diabetes type, the features selected were ['Age', 'BMI', 'Fasting blood glucose', 'Trig', 'TSH', 'Gender']. Extra features are removed and only the selected features are kept based on the threshold that is determined. Different threshold values were tested to analyze their impact on model performance by trying several different values on each dataset. The consequence of this was that (0.1) appeared to be the best threshold in dataset1 and dataset2. On the other hand, in the dataset3 the best results were obtained in predicting diabetes with a threshold value of (0.3), and as for classifying the disease, the best results were with a threshold value of (0.2) since the impactful features could be captured and thus irrelevant or noisy variables which widely removed from the data, making an accurate model while reducing overfitting.

In our proposed model, the correlation is calculated for all datasets. In The dataset1 the correlation heatmap in Figure 3.3 shows that there are some health features that are strongly correlated with each other, while lifestyle and demographic features show weaker correlations. These relationships indicate that certain health conditions are highly correlated (e.g., colorectal cancer and other digestive disorders, or HIV and other infectious diseases) while others have a more of an average correlation (e.g., income and age). As a result, the gap parameters of health-related features are critical in predicting the result.

The Correlations in the Dataset2, display strong relationships between key features, it looks like a good dataset for predicting diabetes, even though some variables are very little affected as shown in Figure 3.4.

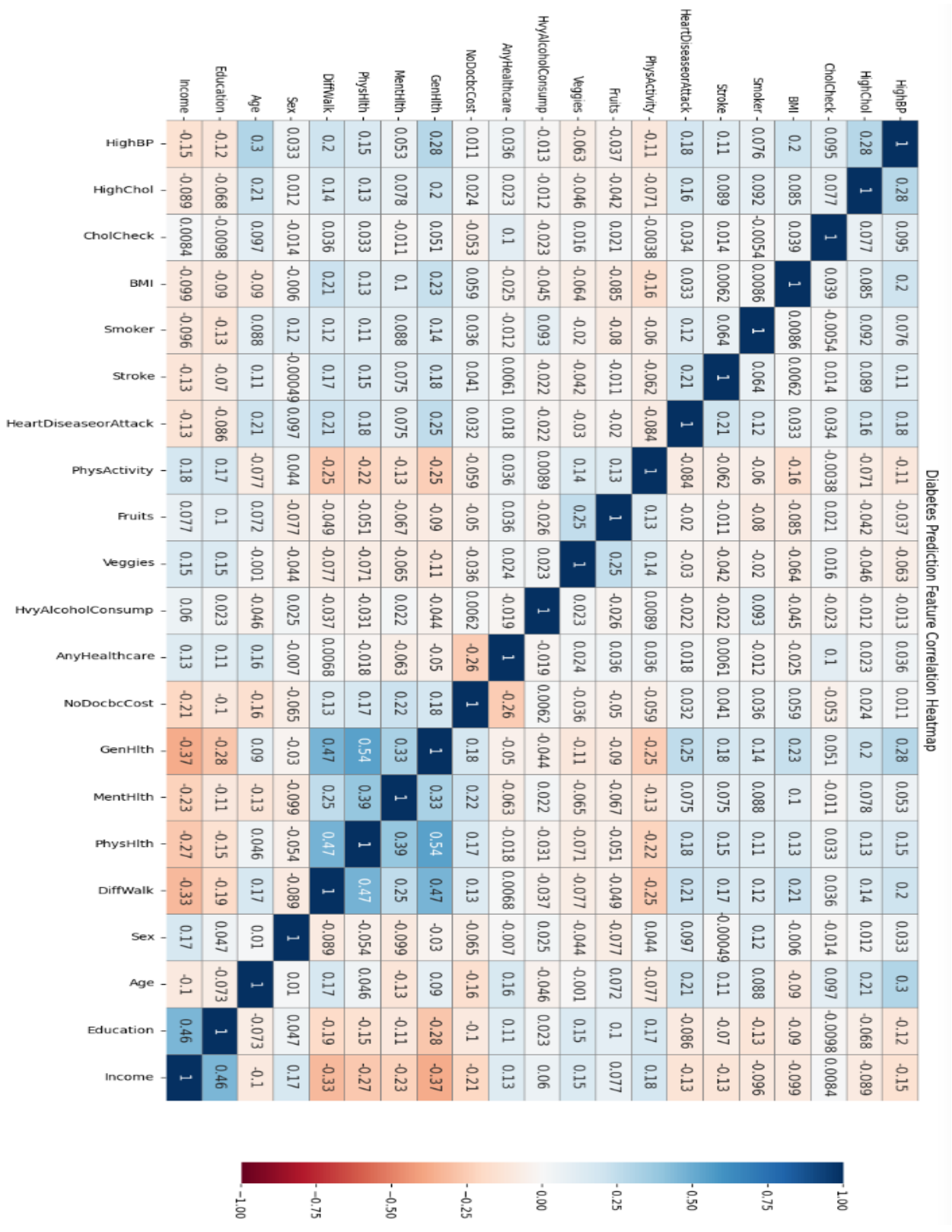

Figure 3.3 Diabetes prediction feature correlation heatmap in dataset1

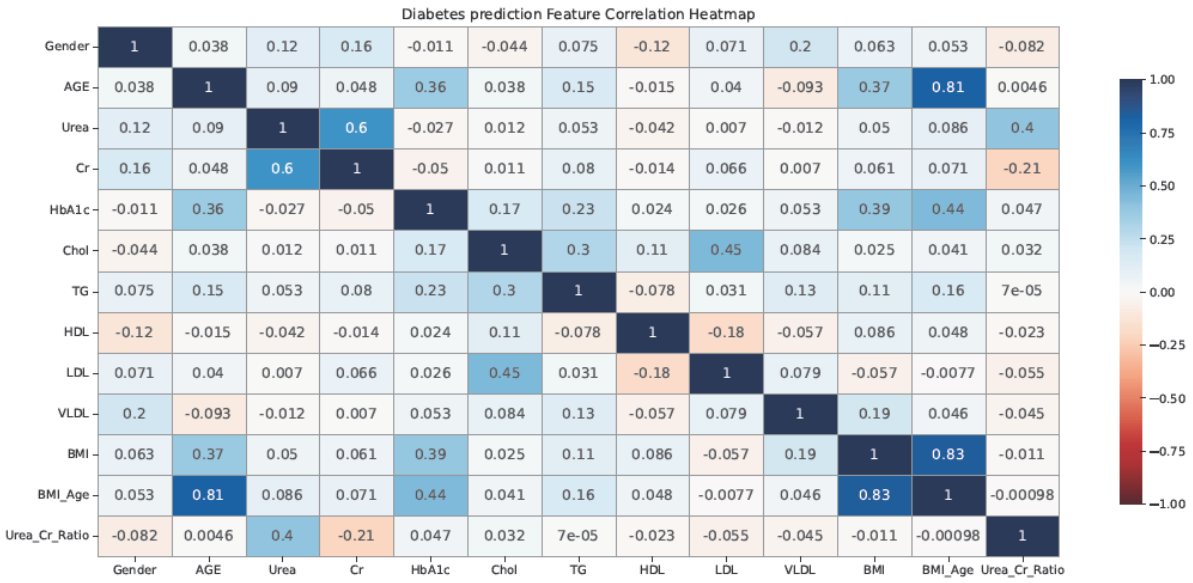

Figure 3.4 Diabetes prediction feature correlation heatmap for dataset2

In the dataset3 Various features show strong correlations such as those in the heatmap of prediction diabetes in Figure 3.5, while some features have low or non-existent correlation. It sheds light on how the different variables in the dataset are related to one another.

From the correlation heatmap in Figure 3.6 the dataset is well organized for diabetes type classification there are some include meaningful relationship between a several features, while some variables appear weak influence.

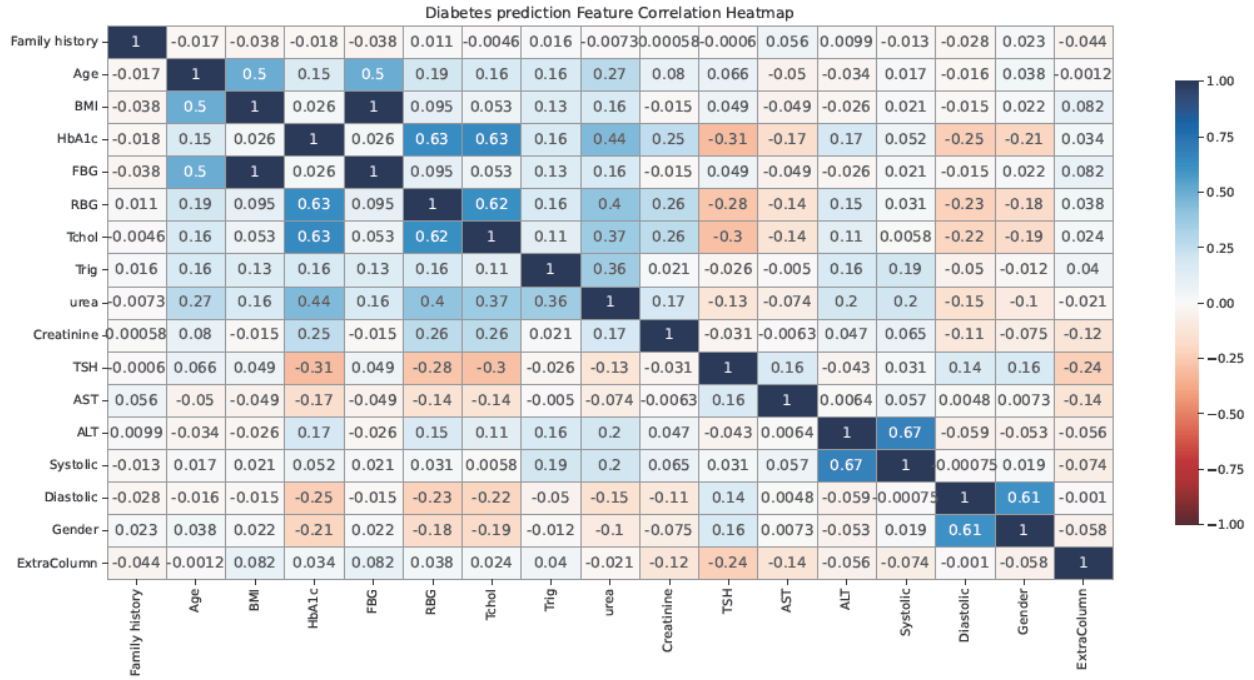

Figure 3.5 Diabetes feature correlation heatmap for dataset3

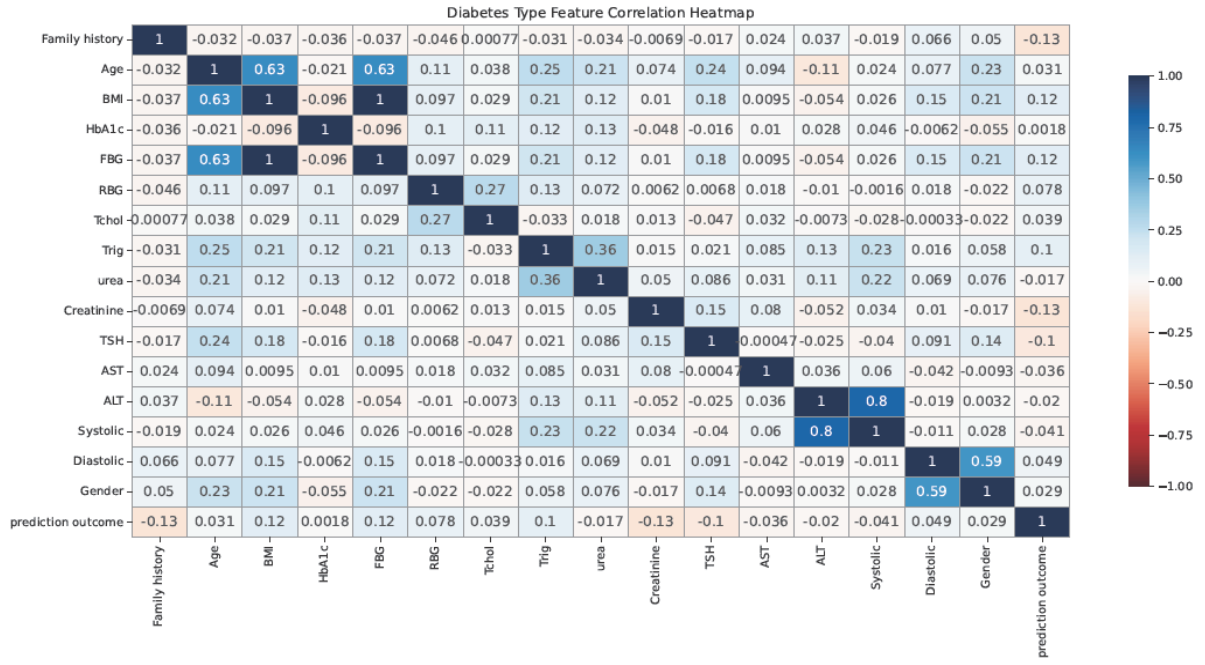

Figure 3.6 Diabetes type feature correlation heatmap for dataset3

3.2.2 Splitting the datasets

After several experiments in splitting the datasets into a training set and a testing set in order to obtain the best results and accuracy, the data splitting was determined to be 70% training and 30% testing for all datasets.

3.2.3 Models Implementation

The machine learning algorithms introduced in Chapter 2, Section 2.3 are the primary machine learning algorithms used in this study. A short yet full comparative between most of the successful, popular algorithms was made in diabetes predicting between last ones used in most studies, diverse machine learning algorithms were employed including in diabetes prediction of three datasets used was (DecisionTree, SVC, KNN,VotingClassifier, BaggingClassifier, RandomForest, XGBoost, and LightGBM), while in the second part of the dataset3 to classify diabetes types these algorithms are used (random forest, decision tree, support vector machine, naive Bayes, logistic regression, k-nearest neighbors, LightGBM and voting). The following section presents a single pseudocode for each algorithm one by one and explain how these algorithms work:

Decision Tree
Input: # x_train: Training features matrix # y_train: Training target vector # Import Decision Tree Import (Decision_Tree) # Calculate best depth

```
maxDepthRange = list (range (2, 12))
for D in maxDepthRange:
    decTree = Decision_Tree(max_depth=D)
    decTree.fit (x_train, y_train)
    score = decTree.score(x_test, y_test)
    r2.append (score)
# Plot the performance of different depths
Plot (maxDepthRange, r2)
# Initialize Decision_Tree with best depth
decTree = Decision_Tree(max_depth=best_depth)
# Fit the model on the training data
decTree.fit (x_train, y_train)
# Output:
# - Model Metrics Scores
Metrics_Score(decTree)
```

END

SVM

```
# Input:
# x_train: Training features matrix
# y_train: Training target vector
# Import SVM
Import (SVM)
# Calculate best C parameter (Regularization)
CRange = [0.1, 1, 10, 100]
for C in CRange:
```

```
    svm = SVM(C=C)
    svm.fit(x_train, y_train)
    score = svm.score(x_test, y_test)
    accuracy_scores.append(score)
# Initialize SVM with the best C
svm = SVM(C=best_C)  # best_C is the value of C that gave the highest score
# Fit the model on the training data
svm.fit(x_train, y_train)
# Output:
# - Model Metrics Scores
Metrics_Score(svm)
```

END

KNN

```
# Input:
# x_train: Training features matrix
# y_train: Training target vector
# Import KNN
Import (KNN)
# Initialize KNN with desired number of neighbors
knn = KNN(n_neighbors=3)  # You can change n_neighbors based on your
needs
# Fit the model on the training data
knn.fit(x_train, y_train)
```

```
# Output:
# - Model Metrics Scores
Metrics_Score(knn)
```

END

Logistic Regression

```
#Input:
#x_train:Training features matrix
# y_train: Training target vector
# Import Logistic Regression
Import (LogisticRegression)
# Calculate best regularization strength (C)
CRange = [0.1, 1, 10, 100]
for C in CRange:
    logReg = LogisticRegression(C=C)
    logReg.fit(x_train, y_train)
    score = logReg.score(x_test, y_test)
    accuracy_scores.append(score)
Plot(CRange, accuracy_scores)
# Initialize Logistic Regression with best C
logReg = LogisticRegression(C=best_C)  # best_C is the value of C that gave the highest score
# Fit the model on the training data
logReg.fit(x_train, y_train)
# Output:
# - Model Metrics Scores
```

```
Metrics_Score(logReg)
```

END

Naïve Bayes

```
# Input:
#x_train: Training features matrix
# y_train: Training target vector
# Import Naive Bayes
Import (NaiveBayes)
# Initialize Naive Bayes model
nb = NaiveBayes()
# Fit the model on the training data
nb.fit(x_train, y_train)
# Output:
# - Model Metrics Scores
score = nb.score(x_test, y_test)
Metrics_Score(nb)
```

END

Random Forest

```
# Input:
#x_train: Training features matrix
# y_train: Training target vector
# Import Random Forest
Import (RandomForest)
```

```
# Calculate best number of trees (n_estimators)
nTreesRange = list(range(10, 101, 10))
for n in nTreesRange:
    rf = RandomForest(n_estimators=n)
    rf.fit(x_train, y_train)
    score = rf.score(x_test, y_test)
    accuracy_scores.append(score)
Plot(nTreesRange, accuracy_scores)
# Initialize RandomForest with best number of trees
rf = RandomForest(n_estimators=best_n)  # best_n is the number of trees with the highest score
# Fit the model on the training data
rf.fit(x_train, y_train)
# Output:
# - Model Metrics Scores
Metrics_Score(rf)
```

END

Bagging

```
# Input:
#x_train: Training features matrix
# y_train: Training target vector
# Import Bagging and Decision Tree
Import (Bagging)
Import (Decision_Tree)
```

```
# Initialize Bagging with Decision Tree as base classifier
bagging_model = Bagging(base_estimator=Decision_Tree(), n_estimators=50)
# 50 trees

# Fit the model on the training data
bagging_model.fit(x_train, y_train)

# Output:
# - Model Metrics Scores
Metrics_Score(bagging_model)
```

END

XGBoost

```
# Input:
#x_train: Training features matrix
# y_train: Training target vector
# Import XGBoost
Import (XGBoost)

# Hyperparameter tuning for XGBoost
best_score = -float('inf')
best_params = {}
for depth in range(3, 15):
    for lr in [0.01, 0.05, 0.1, 0.2]:
        for n_estimators in [50, 100, 200]:
```

```
        xgb = XGBoost(max_depth=depth, learning_rate=lr,
n_estimators=n_estimators)
        xgb.fit(x_train, y_train)
        score = xgb.score(x_test, y_test)
        if score > best_score:
            best_score = score
            best_params = {'max_depth': depth, 'learning_rate': lr, 'n_estimators':
n_estimators}
# Initialize XGBoost with best parameters
xgb_best = XGBoost(**best_params)
xgb_best.fit(x_train, y_train)
# Output:
# - Model Metrics Scores
Metrics_Score(xgb_best)
```

END

LightGBM

```
# Input:
#x_train: Training features matrix
# y_train: Training target vector
# Import LightGBM
Import (LightGBM)
# Tune n_estimators and learning_rate
best_score = 0
for n in [50, 100, 200]:
```

```
    for lr in [0.01, 0.05, 0.1]:
        model = LightGBM(n_estimators=n, learning_rate=lr)
        model.fit(x_train, y_train)
        score = model.score(x_test, y_test)
        if score > best_score:
            best_score = score
            best_params = {'n_estimators': n, 'learning_rate': lr}
# Initialize model with best params and fit
best_model = LightGBM(n_estimators=best_params['n_estimators'],
learning_rate=best_params['learning_rate'])
best_model.fit(x_train, y_train)
# Output:
# - Model Metrics Scores
Metrics_Score(best_model)
```

END

Voting

```
# Input:
# x_train: Training features matrix
# y_train: Training target vector
# Import classifiers and VotingClassifier
Import (LogisticRegression, SVC, RandomForestClassifier, VotingClassifier)


# Initialize individual classifiers
clf1 = LogisticRegression()
```

```
clf2 = SVC()
clf3 = RandomForestClassifier()

# Initialize Voting Classifier
voting_clf = VotingClassifier(estimators=[('lr', clf1), ('svc', clf2), ('rf', clf3)],
voting='hard')

# Fit the Voting Classifier on the training data
voting_clf.fit(x_train, y_train)

# Output:
# - Model Metrics Scores
Metrics_Score(voting_clf)
```

END

3.3 Evaluation Performance

One of the important concepts in machine learning is performance evaluation, which involves assessing the effectiveness of the model to be able to generalize on new data unseen from the model in training.

3.3.1 Confusion Matrix

This study utilizes the confusion matrix and evaluation metrics like accuracy, precision, recall, and F1 score that explained in chapter 2 section 2.5.1

3.3.2 Early Stopping Using K-Fold Cross Validation

This technique prevents overfitting by stopping the training process when the model performance on the training data stops improving and giving better results, where the method of its work was explained in the second chapter in the section 2.5.2. In our model, after several tests 5 folds was used to get good results in all datasets. The early stopping technique was also applied, which stops training when there is no improvement in performance over a specified number of epochs.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Overview

In this chapter, the outcomes are presented of the three datasets obtained from the suggested model to predict if an individual possesses diabetes and classify the type of diabetes (either Type 1 or Type 2) in dataet3. Different machine learning algorithms were trained and tested with the results of each one shown either for the prediction or the classification. By using sophisticated machine learning methods and feature selection algorithms to improve accuracy and efficiency, the research illustrates the success of the proposed method applied to both challenges.

4.2 Results of the Proposed Model

4.2.1 Result of the First Dataset

Of all these classification machine learning models conducted on the dataset in Table 4.1.

Table 4.1 Diabetes Prediction Scores in Dataset1

ML Used	Accuracy	Precision	Recall	F1 Score	AUC
Decision Tree	0.5267	0.5289	0.5267	0.5276	0.7199
SVM	0.6757	0.6674	0.6877	0.6763	0.7849
KNN	**0.8691**	**0.8779**	**0.8691**	**0.8653**	**0.9546**
Voting	0.6317	0.6364	0.6317	0.6303	Nan
Bagging	**0.8927**	**0.8923**	**0.8927**	**0.8921**	**0.9757**
Random Forest	0.5386	0.5358	0.5386	0.5287	0.7385
XGboost	0.6658	0.6667	0.6658	0.6659	0.8386
LightGBM	0.8015	0.8002	0.8015	0.7990	0.9274

Bagging based on Decision Tree performs best with an accuracy of 0.8927, precision of 0.8923, recall of 0.8927, F1 score of 0.8921, and AUC of 0.9757 and the confusion matrix shown in figure 4.1 and AUC-ROC plot in Figure 4.2. With an accuracy of 0.8691 and an AUC of 0.9546, K-Nearest Neighbors (KNN) follows closely behind, making tow of our models with strong predictive capabilities and the confusion matrix shown in figure 4.3 and AUC-ROC plot in Figure 4.4.

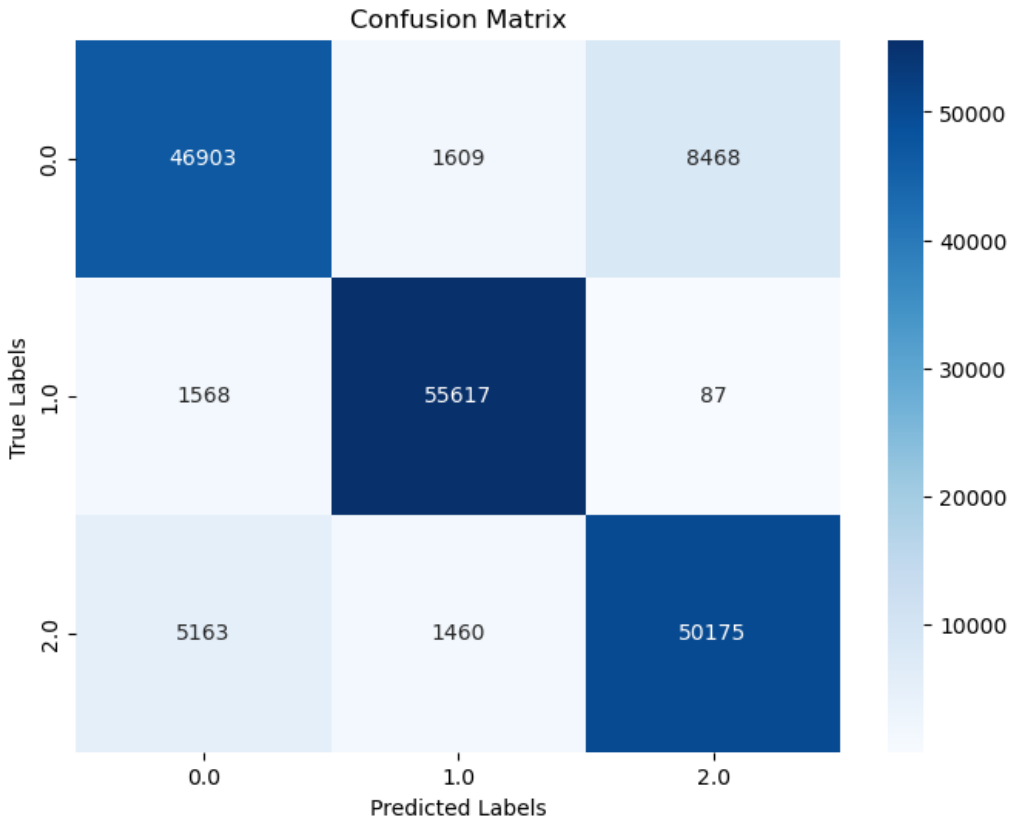

Figure 4.1Confusion matrix of Bagging in Dataset1

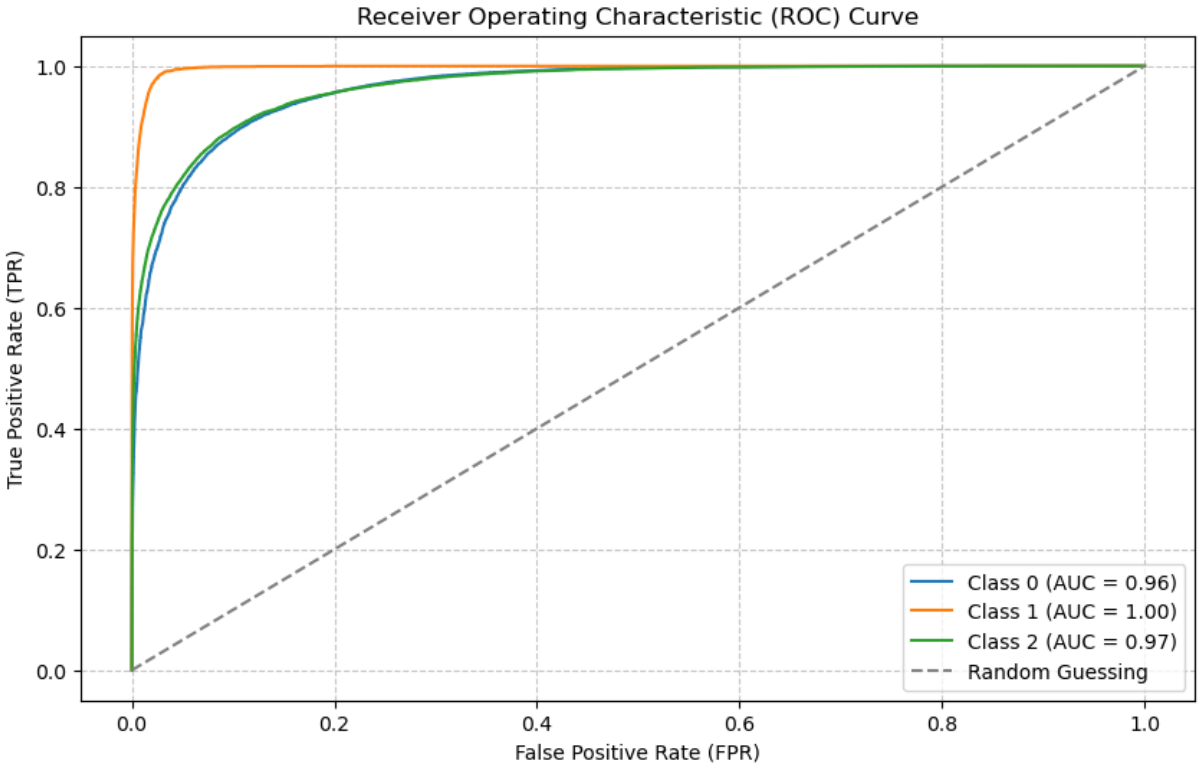

Figure 4.2 ROC Curve of Bagging in dataset1

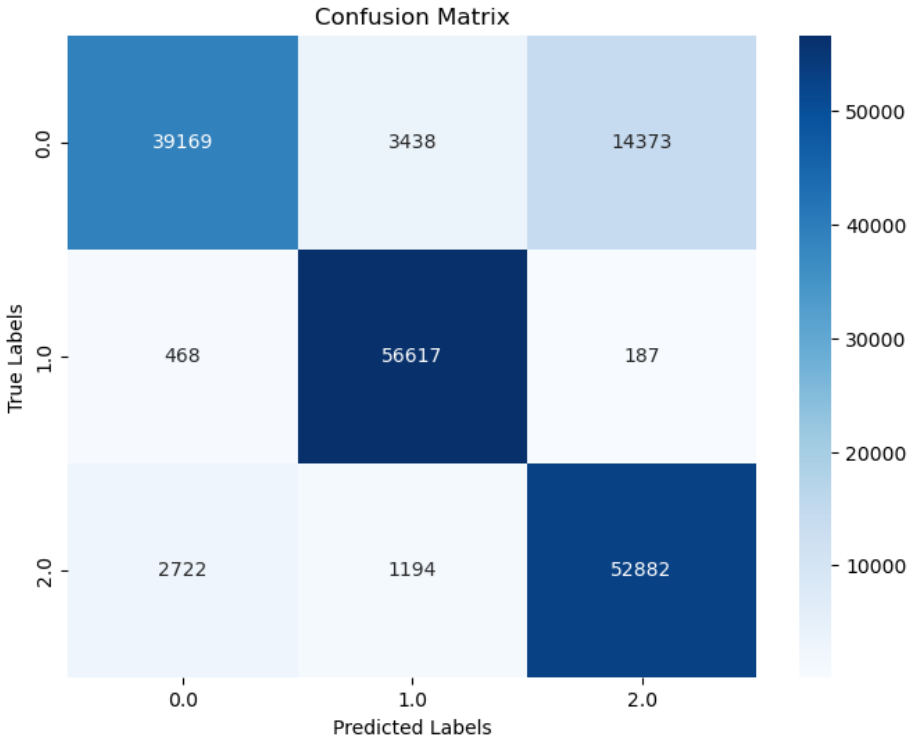

Figure 4.31Confusion matrix of KNN in Dataset1

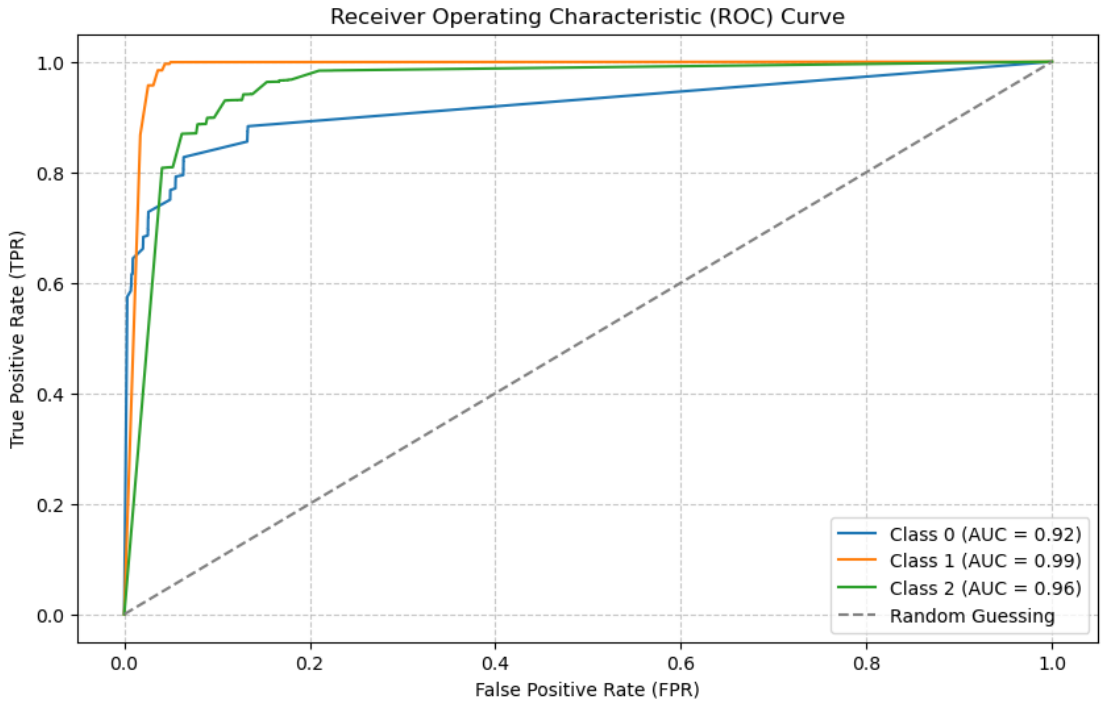

Figure 4.4 ROC Curve of KNN in dataset1

LightGBM does well and scores 0.8015 for accuracy and 0.9274 for AUC, beating XGBoost at 0.6658 for accuracy and 0.8386 for AUC. Of the models that performed poorly, Decision Tree and Random Forest had the weakest performance with an accuracy of 0.5267 and 0.5386, respectively and AUCs under 0.74 given that these are classification-based models, they did not work outside of their normal scope of administrative undersampling. The SVM models perform moderately, returning an accuracy of 0.6757 and an AUC of 0.7849, where in comparison the Voting classifier returned an accuracy of 0.6317 but no AUC value. Interestingly, these results emphasize the power of the ensemble-based methods, notably Bagging and KNN, which not only outperform the individual algorithms but also reaffirm the crucial role that ensemble learning plays in making accurate and generalizable predictions.

To verify the accuracy and effectiveness of the model on Dataset 1, the results of previous studies was compared based on their accuracy in predicting diabetes, as

shown in the table below. This table includes two external studies: the first study[21] used the CatBoost algorithm and achieved an accuracy of 86.6%, while the second study[22] used the AdaBoost algorithm, achieving an accuracy of 86.1%. Comparing these results, our results demonstrate better accuracy, with KNN achieving an accuracy of 86.9%, while Bagging achieved the highest accuracy of 89.2%, indicating improved effectiveness compared to reference methods.

Table 4.2 Comparison of The Results of Previous Studies with Dataset 1

ML Model	Classifier used	Accuracy
[21]	Cat boost	86.6%
[22]	Ada boost	86.1%
Our model	KNN	86.9%
Our model	bagging	89.2%

4.2.2 Result of the Second Dataset

In the comparative analysis of the various machine learning models in Table 4.3, it can be observed that tree-based and ensemble approaches proved to be better in terms of their classification ability.

The best scores were attained by the Decision Tree classifier, which managed to achieve a score of 0.9950 in terms of accuracy, and 0.9994 in AUC, pointing towards the ability to classify data with near-perfect accuracy. The confusion matrix of this model shown in figure 4.5 and AUC-ROC plot in Figure 4.6.

Table 4.3 Diabetes Prediction Scores in Dataset2

ML USED	Accuracy	Precision	Recall	F1 score	AUC
Decision Tree	**0.9950**	**0.9952**	**0.9950**	**0.9951**	**0.9994**
SVM	0.9500	0.9497	0.9500	0.9497	0.9795
KNN	0.9300	0.9355	0.9300	0.9322	0.9820
Voting	0.9800	0.9812	0.9800	0.9804	Nan
Bagging	0.9150	0.9260	0.9150	0.8911	0.9795
Random Forest	0.9850	0.9854	0.9850	0.9852	0.9959
XGboost	0.9750	0.9506	0.9623	0.9563	0.9967
LightGBM	0.9650	0.9647	0.9650	0.9647	0.9942

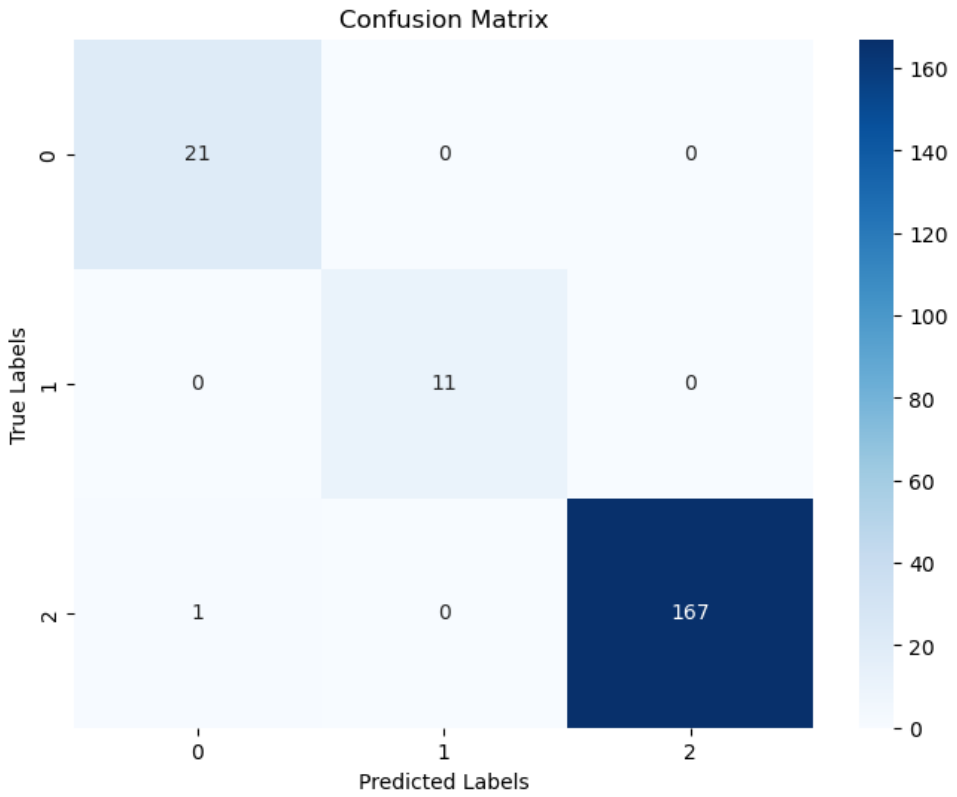

Figure 4.5 Confusion matrix of Decision Tree in Dataset2

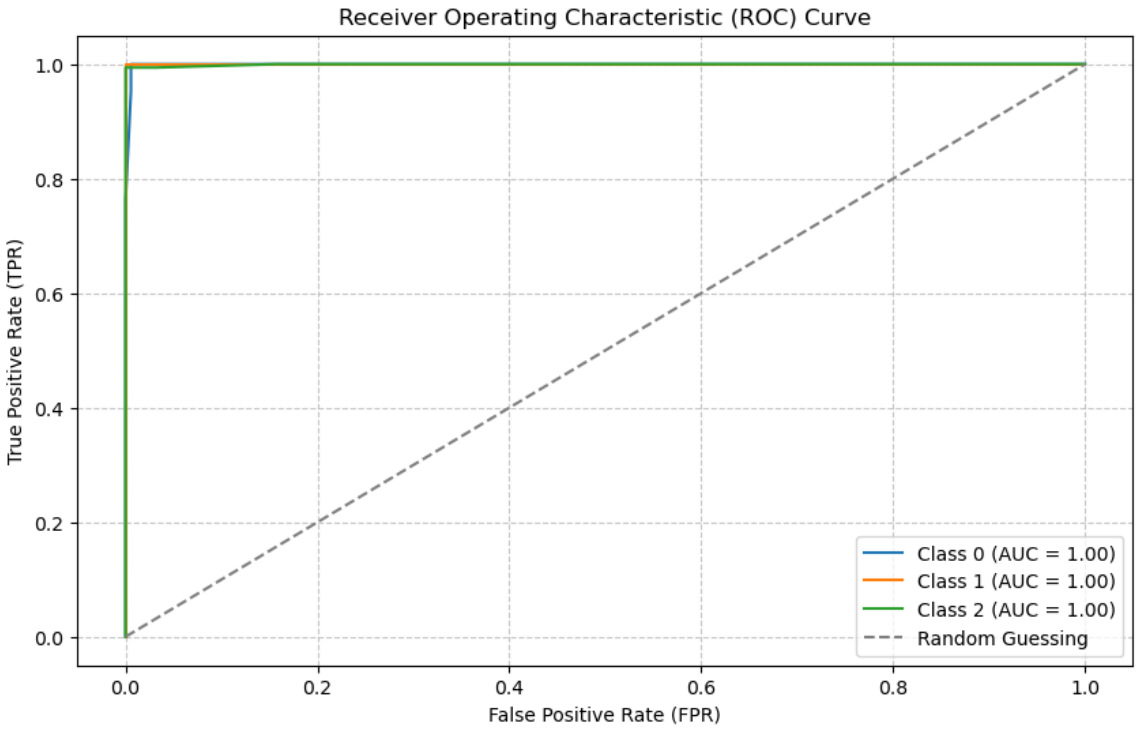

Figure 4.6 ROC Curve of Decision Tree in dataset2

Random Forest, XGBoost, and LightGBM are also conveniently included which also gave us impressive accuracies of 0.9850, 0.9750, and 0.9650 respectively, as well as high AUCs, which suggests further confirmation of the applicability of ensemble and boosting methods. The Voting Classifier performed strongly, with an accuracy of 0.9800, demonstrating the effectiveness of model aggregation, but reported no AUC. Bagging Classifier achieved modest performance with 0.9150 accuracy and 0.9795 area under curve (AUC), indicating that reducing variance did not work as well as boosting methods did. On the other hand, SVM and KNN had a poor performance with the lowest accuracy of 0.9300 obtained via KNN, suggesting their incapacity to cope properly with the complexity of the dataset. These results highlight the strong superiority of boosting methods (XGBoost and LightGBM) in achieving better classification performance and generalization which make them the best candidates for predictive modeling while the traditional classifiers such as SVM and KNN showed their weakness.

After obtaining the results, they should be compared with the results of previous studies to confirm the effectiveness of the proposed model and its ability to predict diabetes. Table 4.4 shows the various classifiers that used dataset 2. Reference [12] used the AWOD (weighted average-based objective distance) method, achieving an accuracy of 98.95%. Reference [15] used the 2GDNN model with an accuracy of 97.33%, while Reference[20] applied the SVM classifier and achieved an accuracy of 96.4%. In comparison, our model, which used a decision tree, achieved the highest accuracy of 99.51%, demonstrating superior performance to the previously mentioned methods.

Table 4.4 Comparison of The Results of Previous Studies with Dataset2

ML Model	Classifier used	Accuracy
[12]	average-based weighted objective distance (AWOD)	98.95%
[15]	twice-growth deep neural network (2GDNN) model	97.33%
[20]	SVM	96.4%
Our model	Decision tree	99.51%

4.2.3 Result of the Third Dataset

The third dataset contains two parts. For the first part, which is about predicting diabetes, it was shown that the comparative analysis of multiple machine learning models in table 4.4.

Table 4.5Diabetes Prediction Scores in Dataset3

Ml Used	Accuracy	Precision	Recall	F1 Score	AUC
Decision Tree	99.5825	99.5825	99.5825	99.5825	99.5077
SVC	98.9562	98.9764	98.9562	98.9554	99.9860
KNN	98.9562	98.6354	98.9562	98.9557	99.7621
Bagging	99.5825	99.5825	99.5825	99.5825	99.9956
Voting	99.3737	99.3810	99.3737	99.3734	Nan
Random Forest	**99.7912**	**99.7921**	**99.7912**	**99.7912**	**99.9983**
XGboost	99.5825	99.9748	99.1150	99.5556	99.9860
LightGBM	**99.7912**	**99.7921**	**99.7912**	**99.7912**	**99.9948**

showcases two main findings; mainly, the superiority of ensemble-based methods like Random Forest and LightGBM with the highest accuracy (99.7912), precision (99.7921), recall (99.7912), and F1-score (99.7912), as well as AUC values equal to 99.9983 and 99.9948 respectively, revealing their robustness and generalization capabilities. The confusion matrix of this model shown in figure 4.7 and AUC-ROC plot in Figure 4.8.

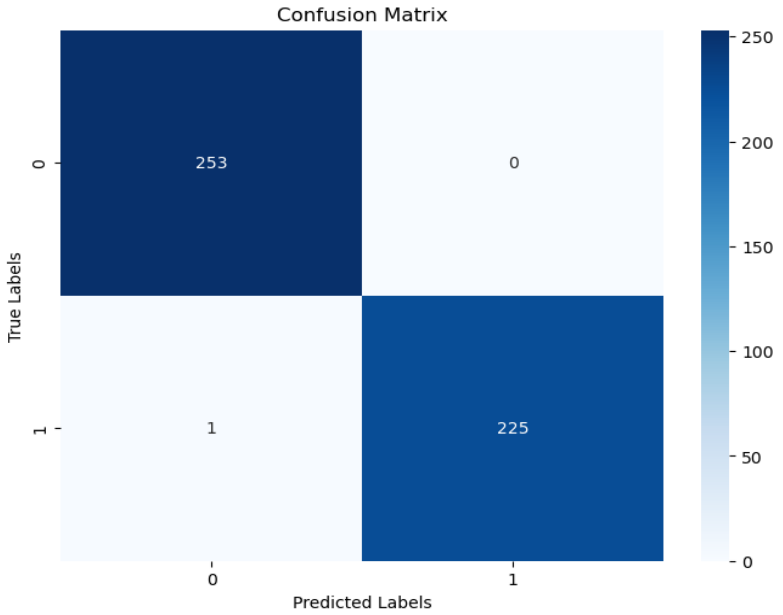

Figure 4.7 Confusion matrix of Random Forest and LightGBM in Dataset3 for Diabetes Prediction

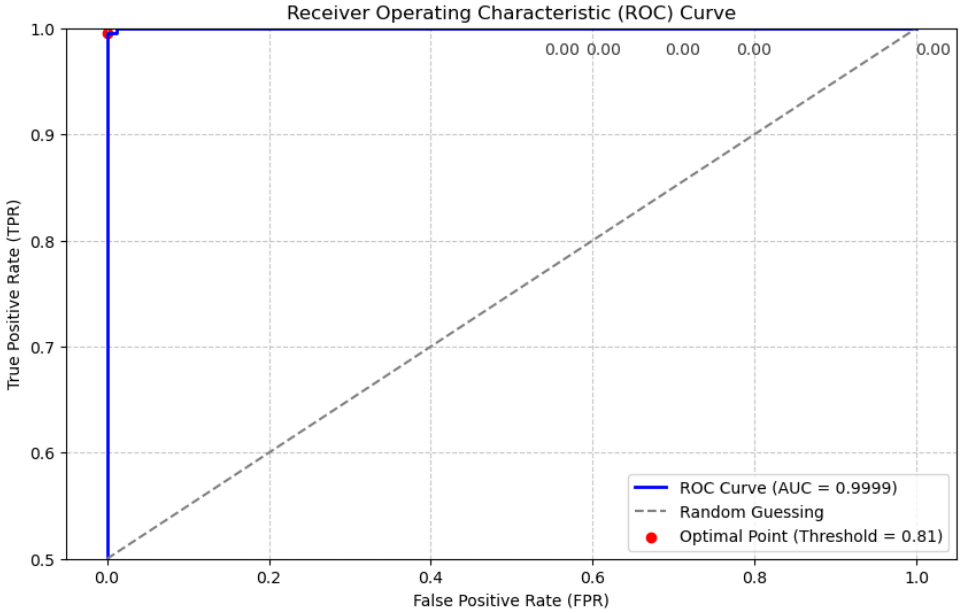

Figure 4.8 ROC Curve of Random Forest and LightGBM in Dataset3 for Diabetes Prediction

The XGBoost accuracy was the second-best at 99.5825 with precision, recall and AUC of 99.9748, 99.1150 and 99.9860 respectively, making it a competitive alternative. Both the Bagging classifier and Decision Tree classifier showed the same performance with an accuracy of 99.5825 and AUC values of

99.9956 and 99.5077, indicating that both of these classifiers are suitable for classification tasks. The classifier Voting reached an accuracy of 99.3737, but an AUC value is not reported, indicating that the ensemble strategies may have paid off, but does not provide any AUC evaluation. The performance of SVC and KNN was lower, with accuracy at 98.9562 and AUC values at 99.9860 and 99.7621 respectively, suggesting that although they are effective, they maybe poorly trained for more complex classification tasks than tree-based models. From an overall perspective, these results further confirm that ensemble models, especially Random Forest, LightGBM and XGBoost, are the most robust in terms of predictive capacity and generalization, and therefore fit for high-stakes classification tasks.

As shown in Table 4.5, evaluation of the various machine learning models shows that the Voting (Decision Tree, Random Forest and LightGBM) classifier provides the highest accuracy (95.74%), precision (95.91%), recall (95.74%) and F1-score (95.71%), reflecting its strength in predictive power. The confusion matrix of this model shown in figure 4.9 and AUC-ROC plot in Figure 4.10.

Table 4.6 Diabetes Type Classification Scores in Dataset3

Ml Used	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	94.89	94.91	94.89	94.87	95.5724
Decision Tree	93.62	93.61	93.62	93.61	93.6265
Logistic Regression	93.19	93.19	93.19	93.19	94.9377
Naive B	90.21	90.38	90.21	90.26	93.3718
SVC	92.77	92.83	92.77	92.71	95.6902
KNN	92.34	95.78	92.34	92.27	92.6915
Voting	**95.74**	**95.91**	**95.74**	**95.71**	**95.1961**
LightGBM	94.89	95.04	94.89	94.85	94.7933

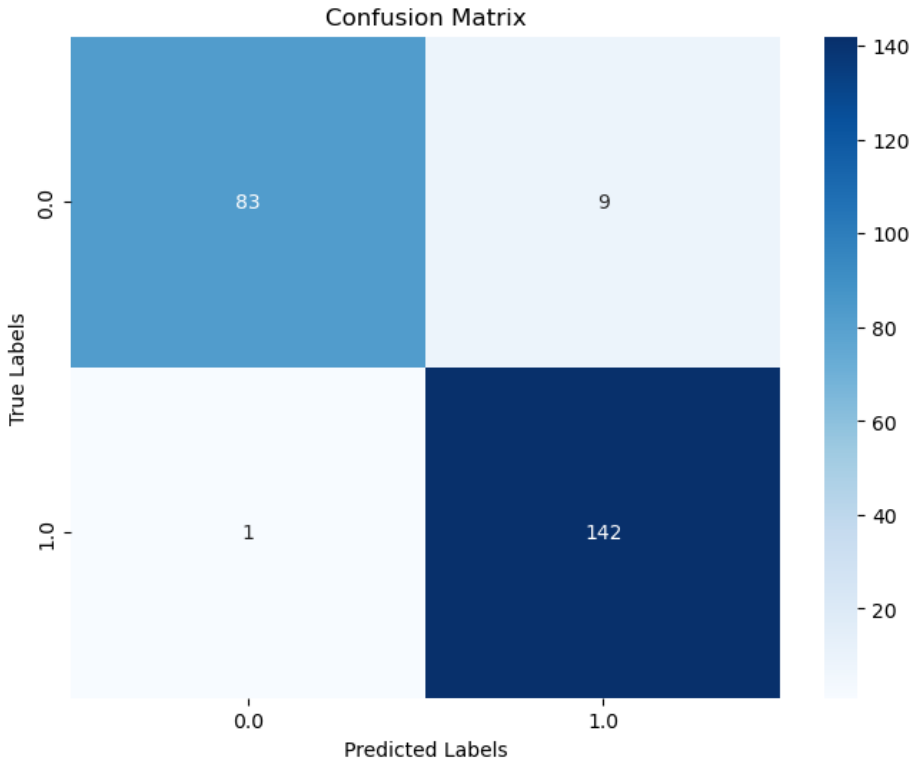

Figure 4.9 Confusion matrix of Voting in Dataset3 for Diabetes Type Classification

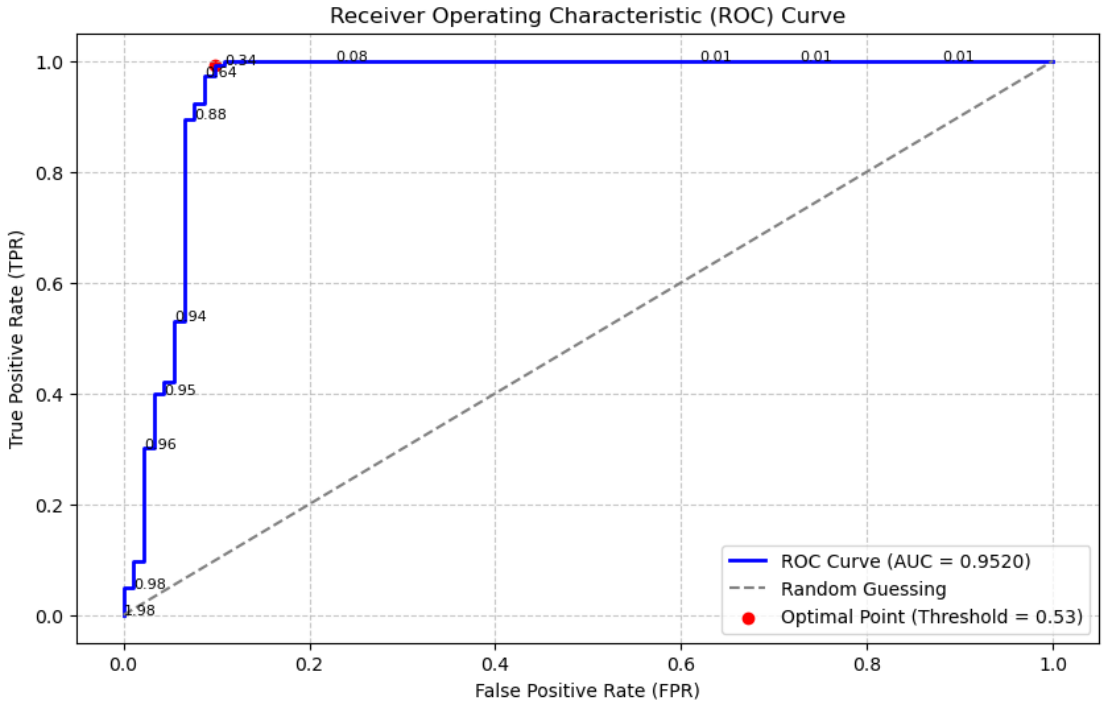

Figure 4.10 ROC Curve of Voting in Dataset3 for Diabetes Type Classification

Close after those are Random Forest and LightGBM, both achieving an accuracy of 94.89% — with LightGBM having a slightly higher precision (95.04%) than Random Forest (94.91%). At the same time, Random Forest provides the best AUC value of 95.5724, indicating that random forest is the best prediction model among them. Additionally, the Support Vector Classifier (SVC) also gets a AUC of 95.6902, slightly above the one achieved with LightGBM (94.7933) and Logistic Regression (94.9377). The traditional models display similar performance, with Decision Tree and Logistic Regression achieving an accuracy of 93.62% and 93.19% respectively. Although Naïve Bayes has the lowest accuracy as mentioned above (90.21%), its AUC score is tolerably at 93.3718. The KNN algorithm, on the other hand, shows a lower level of recall (92.34%) and F1 (92.27%), but its precision (95.78%) is much better, which also suggests that the algorithm has been able to minimize the rate of false positive records. The ensemble methods particularly Voting and Random Forest outperform, reaffirming the strength of ensemble methods for classification.

4.3 Discussion

Ensemble models play a significant role in improving the accuracy and reliability of diabetes prediction by combining the strengths of multiple individual classifiers. Techniques such as random forest, bagging, and voting classifiers combine predictions from diverse algorithms to reduce variance, limit bias, and enhance generalization. This approach is particularly effective in dealing with the complex, nonlinear relationships often found in health-related datasets. In diabetes prediction, ensemble models have consistently demonstrated higher accuracy compared to individual models, benefiting from a wider range of patterns and decision boundaries. Their robustness against overfitting and improved predictive

performance makes them particularly valuable in clinical applications where accuracy is critical.

The significantly higher accuracy observed in the laboratory-based datasets (Datasets 2 and 3) compared to the lifestyle-based dataset (Dataset 1) can be attributed to the nature of the features used in each. Laboratory-based features, such as blood glucose levels and glycated hemoglobin (HbA1c), are direct and clinically reliable indicators of diabetes, indicating a strong association with the disease, enabling more accurate model predictions. In contrast, lifestyle-related features, such as nutrition, physical activity, and body measurements, are indirect indicators that may contribute to diabetes risk but do not provide conclusive evidence of the condition, which weakens model performance. As a result, models trained on objective clinical data tend to outperform those based solely on lifestyle factors.

Lifestyle-based traits provide a comprehensive view of a person's behaviors and environment that may predispose them to diabetes. However, these traits are indirect and often self-reported, making them less accurate than clinical measures. However, they are still valuable for early risk assessment, particularly in public health and prevention contexts.

CHAPTER FIVE

CONCLUSION AND FUTURE WORK

5.1 Overview

This chapter will address the conclusion and future work developments to identify the fundamental of the research and pave the path for future researches.

5.2 Conclusion

This thesis demonstrated how to improve diabetes prediction and classification by applying various machine learning algorithms, contributing to more efficient healthcare delivery and enabling timely diagnosis. Given the global burden of diabetes, early detection remains critical to reducing complications and improving patient outcomes. The study used a combination of lifestyle and medical characteristics across multiple real-world and public datasets to train and evaluate several machine learning classifiers. Through extensive experiments, it was concluded that model performance is strongly influenced by both the chosen algorithm and the quality of the input features. Notably, ensemble models, such as Bagging achieved an accuracy of 89.27%, while KNN achieved an accuracy of 86.9% on Dataset 1. On Dataset 2, the decision tree achieved an accuracy of 99.50%. Using a real-world dataset, the Random Forest and LightGBM algorithms achieved the highest accuracy of 99.79% in predicting diabetes. Additionally, the voting classifier achieved an accuracy of 95.74% in identifying the type of diabetes. These results underscore the importance of machine learning techniques in dealing with complex patterns in clinical data. Additionally, regularization played a key role in reducing overfitting and enhancing model generalization on unseen data.

The study also offers several theoretical and practical contributions. Theoretically, machine learning applications in healthcare are advanced by calibrating multiple models—such as SVM, decision tree, KNN, and ensemble methods—while highlighting the interpretability of key indicators such as BMI and glucose levels. A robust evaluation framework based on accuracy, F1 score, and AUC enhanced the generalizability and reliability of the results across diverse populations. Practically, the models support early diagnosis and personalized risk assessment, providing a cost-effective screening solution that automates diabetes prediction using routine clinical data. The developed framework not only provides reproducibility but also serves as a benchmark for future machine learning-based medical research. Ultimately, the research objectives were achieved: identifying effective algorithms, integrating academic and real-world data, and designing and evaluating a machine learning model capable of predicting diabetes and classifying its types. This study underscores the growing potential of machine learning as a decision support tool in clinical settings and reflects its pivotal role in the future of chronic disease management.

This study faced several limitations, including limited sample sizes and the absence or inconsistency of data in real-world datasets. This required extensive preprocessing, especially since the data was paper-based, requiring re-formatting for use in the model. The use of static data limited the ability to detect temporal health changes. Furthermore, the results may not generalize well to underrepresented populations, and immediate clinical validation remains a future goal.

5.3 Future Work

This study provided a foundation for subsequent studies. This study can be enhanced and built upon in many areas, such as:

1. Utilizing larger, more diverse datasets, and adding features, such as genetic and lifestyle information, would allow for better generalization and performance of the models.

2. Future research should investigate a wider range of algorithms, including ensemble approaches and deep learning models, to improve prediction performance.

3. Add time-series models, like LSTMs, to take temporal changes into account, thereby improving predictions and tracking diabetes progression in patients.

4. Developing an application to provide immediate assessment of diabetes and disease risk, early alerts, and lifestyle guidance, supporting early intervention and integration with healthcare systems

REFERENCES

[1] K. Roy *et al.*, "An enhanced machine learning framework for type 2 diabetes classification using imbalanced data with missing values," vol. 2021, no. 1, p. 9953314, 2021.

[2] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. J. J. o. h. e. Sherazi, "Machine learning based diabetes classification and prediction for healthcare applications," vol. 2021, no. 1, p. 9930985, 2021.

[3] E. Dritsas and M. J. S. Trigka, "Data-driven machine-learning methods for diabetes risk prediction," vol. 22, no. 14, p. 5304, 2022.

[4] J. Xie and Q. J. I. T. o. B. E. Wang, "Benchmarking machine learning algorithms on blood glucose prediction for type I diabetes in comparison with classical time-series models," vol. 67, no. 11, pp. 3101-3124, 2020.

[5] T. Zhu, L. Kuang, J. Daniels, P. Herrero, K. Li, and P. J. I. I. o. T. J. Georgiou, "IoMT-enabled real-time blood glucose prediction with deep learning and edge computing," vol. 10, no. 5, pp. 3706-3719, 2022.

[6] R. Saxena, S. Sharma, and M. Gupta, "Analysis of machine learning algorithms in diabetes mellitus prediction," in *Journal of Physics: Conference Series*, 2021, vol. 1921, no. 1, p. 012073: IOP Publishing.

[7] T. De *et al.*, "Diabetes Prediction using Machine Learning," *IJARCCE*, vol. 13, 03/20 2024.

[8] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," *Health Information Science and Systems*, vol. 8, no. 1, p. 7, 2020/01/03 2020.

[9] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76516-76531, 2020.

[10] M. M. Bukhari, B. F. Alkhamees, S. Hussain, A. Gumaei, A. Assiri, and S. S. Ullah, "An improved artificial neural network model for effective diabetes prediction," *Complexity*, vol. 2021, no. 1, p. 5525271, 2021.

[11] N. Ahmed *et al.*, "Machine learning based diabetes prediction and development of smart web application," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 229-241, 2021.

[12] P. Nuankaew, S. Chaising, and P. Temdee, "Average Weighted Objective Distance-Based Method for Type 2 Diabetes Prediction," *IEEE Access*, vol. 9, pp. 137015-137028, 2021.

[13] Y. Tan, H. Chen, J. Zhang, R. Tang, and P. Liu, "Early risk prediction of diabetes based on GA-stacking," *Applied Sciences,* vol. 12, no. 2, p. 632, 2022.
[14] U. Ahmed *et al.*, "Prediction of diabetes empowered with fused machine learning," *IEEE Access,* vol. 10, pp. 8529-8538, 2022.
[15] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Computer Methods and Programs in Biomedicine,* vol. 220, p. 106773, 2022.
[16] T. Mahesh *et al.*, "Blended ensemble learning prediction model for strengthening diagnosis and treatment of chronic diabetes disease," *Computational Intelligence and Neuroscience,* vol. 2022, no. 1, p. 4451792, 2022.
[17] M. Zarar and Y. Wang, "Early Stage Diabetes Prediction by Approach Using Machine Learning Techniques," 2023.
[18] A. Hennebelle, H. Materwala, and L. Ismail, "HealthEdge: a machine learning-based smart healthcare framework for prediction of type 2 diabetes in an integrated IoT, edge, and cloud computing system," *Procedia Computer Science,* vol. 220, pp. 331-338, 2023.
[19] C. J. Ejiyi *et al.*, "A robust predictive diagnosis model for diabetes mellitus using Shapley-incorporated machine learning algorithms," *Healthcare Analytics,* vol. 3, p. 100166, 2023.
[20] M. A. Sahid, M. U. H. Babar, and M. P. J. P. o. Uddin, "Predictive modeling of multi-class diabetes mellitus using machine learning and filtering iraqi diabetes data dynamics," vol. 19, no. 5, p. e0300785, 2024.
[21] X. J. A. Ren and C. Engineering, "Predictions of diabetes through machine learning models based on the health indicators dataset," vol. 32, pp. 216-222, 2024.
[22] V. Hayyolalam and Ö. J. a. p. a. Özkasap, "DiabML: AI-assisted diabetes diagnosis method with meta-heuristic-based feature selection," 2024.
[23] G. S. Nadella, S. Satish, K. Meduri, and S. S. Meduri, "A systematic literature review of advancements, challenges and future directions of AI and ML in healthcare," *International Journal of Machine Learning for Sustainable Development,* vol. 5, no. 3, pp. 115-130, 2023.
[24] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," *Frontiers in Energy Research,* vol. 9, 03/29 2021.
[25] D. Breskuvienė and G. Dzemyda, "Categorical Feature Encoding Techniques for Improved Classifier Performance when Dealing with

Imbalanced Data of Fraudulent Transactions," *INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL*, vol. 18, 05/09 2023.
[26] D. Das, M. Nayak, and D. S. Pani, "Missing Value Imputation-A Review," *International Journal of Computer Sciences and Engineering*, vol. 7, pp. 548-558, 04/30 2019.
[27] A. El-Shimi, R. Kalach, A. Kumar, A. Oltean, J. li, and S. Sengupta, *Primary data deduplication-large scale study and system design*. 2012, pp. 26-26.
[28] H. He and E. A. Garcia, "Learning from Imbalanced Data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 21, pp. 1263-1284, 10/01 2009.
[29] I. M. Guyon and A. J. J. M. L. R. Elisseeff, "An Introduction to Variable and Feature Selection," vol. 3, pp. 1157-1182, 2003.
[30] O.Yamini and D. G. V. R. Babu, "A Review on Classification of Various Types of Decision Trees with Merits and Demerits," *International Journal For Multidisciplinary Research,* 2023.
[31] R. Saxena, S. K. Sharma, M. Gupta, and G. Sampada, "A novel approach for feature selection and classification of diabetes mellitus: machine learning methods," *Computational Intelligence and Neuroscience,* vol. 2022, no. 1, p. 3820360, 2022.
[32] O. C. Njoku, "Decision Trees and Their Application for Classification and Regression Problems," 2019.
[33] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *2010 IEEE computer society conference on computer vision and pattern recognition*, 2010, pp. 3360-3367: IEEE.
[34] J. Tang, S. Alelyani, H. J. D. c. A. Liu, and applications, "Feature selection for classification: A review," p. 37, 2014.
[35] A. Ben-Hur and J. J. D. m. t. f. t. l. s. Weston, "A user's guide to support vector machines," pp. 223-239, 2010.
[36] S. Sun, C. Luo, and J. J. I. f. Chen, "A review of natural language processing techniques for opinion mining systems," vol. 36, pp. 10-25, 2017.
[37] M. Abdar *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," vol. 76, pp. 243-297, 2021.
[38] S. Suleiman, M. Sani Burodo, and I. Suleman, "Credit Scoring using Principal Components Analysis-based Binary Logistic Regression," *The Journal of Scientific and Engineering Research,* vol. 2017, pp. 99-110, 01/03 2018.

[39] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Computer Science,* vol. 132, pp. 1578-1585, 2018/01/01/ 2018.
[40] F. Alaa Khaleel and A. M. Al-Bakry, "Diagnosis of diabetes using machine learning algorithms," *Materials Today: Proceedings,* vol. 80, pp. 3200-3203, 2023/01/01/ 2023.
[41] R. Genuer, J.-M. Poggi, C. Tuleau-Malot, and N. J. B. D. R. Villa-Vialaneix, "Random forests for big data," vol. 9, pp. 28-46, 2017.
[42] G. Biau and E. J. T. Scornet, "A random forest guided tour," vol. 25, pp. 197-227, 2016.
[43] K. Fawagreh, M. M. Gaber, E. J. S. S. Elyan, and C. E. A. O. A. Journal, "Random forests: from early developments to recent advancements," vol. 2, no. 1, pp. 602-609, 2014.
[44] M. N. Algedawy, "Detecting Diabetes Mellitus using Machine Learning Ensemble," *International Journal of Computer Systems (ISSN: 2394-1065),* vol. 3, no. 12, 2016.
[45] S. M. Ganie and M. B. Malik, "An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators," *Healthcare Analytics,* vol. 2, p. 100092, 2022/11/01/ 2022.
[46] A. P. C and A. K. G, "Ensemble Machine Learning Approach for Detecting and Predicting Diabetes Mellitus Using Bagging and Stacking," *2023 Fourth International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE),* pp. 1-6, 2023.
[47] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785-794.
[48] C. Bentéjac, A. Csörgő, and G. J. A. I. R. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," vol. 54, pp. 1937-1967, 2021.
[49] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *Journal of Big data,* vol. 6, no. 1, pp. 1-19, 2019.
[50] B. S. Ahamed, M. S. Arya, and A. O. V. Nancy, "Diabetes mellitus disease prediction using machine learning classifiers with oversampling and feature augmentation," *Advances in Human-Computer Interaction,* vol. 2022, no. 1, p. 9220560, 2022.
[51] B. S. J. T. J. o. C. Ahamed and M. Education, "Prediction of type-2 diabetes using the LGBM classifier methods and techniques," vol. 12, no. 12, pp. 223-231, 2021.

[52] D. Riccio, F. Maturo, E. J. S. Romano, and Computing, "Supervised learning via ensembles of diverse functional representations: the functional voting classifier," vol. 34, no. 6, p. 191, 2024.
[53] C. Cornelio, M. Donini, A. Loreggia, M. S. Pini, F. J. A. A. Rossi, and M.-A. Systems, "Voting with random classifiers (VORACE): theoretical and experimental analysis," vol. 35, no. 2, p. 22, 2021.
[54] R. A. Lobo and M. E. Valle, "Ensemble of binary classifiers combined using recurrent correlation associative memories," in *Brazilian Conference on Intelligent Systems*, 2020, pp. 442-455: Springer.
[55] C. Bliem, "Estimation of the (re-)integration likelihood into the Austrian labour market: a random forest approach (Christian Bliem, 2020)," 2020.
[56] Z. Yang, Q. Xu, S. Bao, X. Cao, and Q. Huang, "Learning With Multiclass AUC: Theory and Algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7747-7763, 2022.
[57] L.-B. Sweet, C. Müller, M. Anand, and J. Zscheischler, "Cross-Validation Strategy Impacts the Performance and Interpretation of Machine Learning Models," *Artificial Intelligence for the Earth Systems*, vol. 2, pp. 1-35, 07/10 2023.
[58] Y. Yao, L. Rosasco, and A. J. C. A. Caponnetto, "On early stopping in gradient descent learning," vol. 26, no. 2, pp. 289-315, 2007.
[59] Diabetes Health Indicators Dataset [Online]. Available: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset
[60] A. Rashid. Diabetes Dataset [Online]. Available: https://data.mendeley.com/datasets/wj9rwkp9c2/1

الخلاصة

يُعد داء السكري أحد أكثر الأمراض المزمنة انتشارًا على مستوى العالم، إذ يصيب الملايين ويؤدي إلى مضاعفات خطيرة مثل أمراض القلب والفشل الكلوي وفقدان البصر. يُعد الكشف المبكر والتصنيف الصحيح لأنواع داء السكري (النوع الأول والنوع الثاني) أمرًا ضروريًا لتخطيط العلاج الفعال وإدارة المرض على المدى الطويل. ويمكن للكشف المبكر والتنبؤ بمرض السكري أن يُحسّن بشكل كبير نتائج المرضى، مما يجعله مصدر قلق صحي عالمي. يُعد تحليل بيانات المرضى يدويًا طريقة شائعة في تقنيات التشخيص التقليدية، ولكنه قد يكون شاقًا وعرضة للخطأ البشري. تتمثل مشكلة البحث التي يتناولها هذا البحث في عدم كفاءة ودقة طرق التشخيص التقليدية، والتي تهدف هذه الدراسة إلى التغلب عليها باستخدام مناهج آلية قائمة على البيانات. تبحث هذه الدراسة في استخدام البيانات السريرية والديموغرافية جنبًا إلى جنب مع تقنيات التعلم الآلي (ML) للتنبؤ بمرض السكري وتصنيفه. تستخدم الدراسة مجموعة متنوعة من ميزات مجموعات البيانات، مثل عوامل نمط الحياة والبيانات الحيوية والسجلات السريرية، لتدريب وتقييم نماذج التعلم الآلي المختلفة، مثل decision trees, Support vector machine (SVM), K-nearest neighbors (KNN), logistic regression, random forest, bagging, voting, Naïve bayse, XGBoost(Extreme Gradient Boosting), LightGBM (Light Gradient Boosting Machine) ، لإنشاء نماذج تنبؤية. تتضمن مجموعات البيانات الثلاث المستخدمة في هذه الدراسة مجموعتي بيانات معروفتين تم الحصول عليهما من مستودعات متاحة للجمهور مثل مؤشرات صحة السكري التي تحتوي على 253680 عينة، ومجموعة بيانات LMCH للسكري التي تحتوي على 1000 عينة مريض، ومجموعة بيانات واقعية تم جمعها من مركز الإمام حسن المجتبى للسكري والغدد الصماء في العراق، محافظة كربلاء، والتي تحتوي على 1596 عينة مريض، وتشمل ميزات مثل العمر ومؤشر كتلة الجسم ومستويات الجلوكوز وضغط الدم ومستويات الأنسولين. تُطبّق تقنيات المعالجة المسبقة، بما في ذلك التعامل مع المتغيرات الفئوية، والتعامل مع القيم المفقودة، وإزالة التكرارات، وقياس الميزات، ومعالجة اختلال توازن الفئات، واختيار الميزات، لتحسين أداء النموذج. تُقيّم النماذج باستخدام مصفوفة الارتباك، وaccuracy وprecision وrecall و F1-score و AUC . ووفقًا للنتائج، حقق bagging دقة بنسبة 89.27%، بينما حقق KNN دقة بنسبة 86.9%، وهي أعلى دقة في مجموعة البيانات 1. أما بالنسبة لمجموعة البيانات 2، فقد حققت decision trees دقة بنسبة 99.50%. باستخدام مجموعة بيانات واقعية، حققت random forest وLightGBM أعلى دقة بنسبة 99.79% في التنبؤ بمرض السكري. بالإضافة إلى ذلك، حقق مصنف voting دقة بنسبة 95.74% في تحديد نوع مرض السكري، مما يؤكد فعاليته في مهام التصنيف.

جامعة كربلاء
كلية علوم الحاسوب وتكنولوجيا المعلومات
قسم علوم الحاسوب

نموذج التنبؤ والتصنيف لمرض السكري باستخدام التعلم الآلي

رسالة ماجستير
مقدمة الى مجلس كلية علوم الحاسوب وتكنولوجيا المعلومات / جامعة كربلاء وهي جزء من متطلبات نيل درجة الماجستير في علوم الحاسوب

كتبت بواسطة
اية احمد هاشم الموسوي

بإشراف
أ.م.د.اياد حميد موسى

1446هـ 2025 م