# An Explainable AI-Based Machine Learning Framework for Diabetes Prediction

## Author Name

Li Xirui

## Author Student ID

3069898

## Supervisor

Dr. Ifede Parfait Tebe

*April 2025*

*(Word count: 8540 )*

**Dissertation submitted in partial fulfilment for the degree of Bachelor of Science with Honours in Data Science**

BSc (Honours) Data Science

Stirling College-Chengdu University

# Abstract

The increasing prevalence of diabetes in recent years necessitates more accurate, robust and interpretable prediction techniques to reinforce early detection and management of the disease. Current machine learning (ML) approaches used to predict the disease are promising but mostly limited by several factors such as high complexity and imbalance of medical datasets, and lack of prediction outcomes interpretability. To address those limitations, this study proposes a novel ML framework for diabetes prediction performance improvement and for medical doctors and patients to understand prediction outcomes.

The study proposes an ensemble model that integrates two ML base models, with the objective of improving prediction performance, robustness, and generalisation. Additionally, an explainable artificial intelligence (XAI) technique is applied to the framework to provide explainability of the model's prediction outcome, making it transparent and understandable.

To achieve the objectives, four ML base models, namely Random Forest Classifier, AdBoost Classifier, XGBoost Classifier, and CatBoost Classifier were first, individually trained and tested on a publicly available diabetes dataset containing 100,000 instances and 8 features. The models performances were evaluated with metrics such as precision, F1-score, recall, and accuracy, and the results on test set showed a superior performance of CatBoost followed by XGBoost, AdaBoost and Random Forest in that order. CatBoost and XGBoost were then selected to build the ensemble model, using stacking method and Logistic Regression as meta classifier. The proposed ensemble model achieved a test accuracy of 88.93%, outperforming the base models, and it demonstrated its effectiveness of handling complex medical datasets. The model also showed its ability to handle class distribution with no significant bias. Moreover, SHapley Additive exPlanations (SHAP) was used as the XAI, and the results from its application revealed that features such as HbA1c level and blood glucose are the most critical features with a substantial impact on the risk of developing the disease.

Overall, this study proposes an ML-based approach to improve the prediction performance and understanding of diabetes by applying an AI technique to make the prediction results fairly interpretable and hence, enhance the usability and trust of AI in the disease diagnosis.

# Attestation

I understand the nature of plagiarism, and I am aware of the University's academic integrity policy.

I certify that this dissertation reports original work by me during my University project except for the following:

*The dataset used in this project was taken from Kaggle made available by Mohammed Mustafa [9].*

*The original core of the code was taken from an open online repository and was further enhanced and modified according to this project's objectives. Therefore, the major parts of the new code were written by me.*

*Repository Link: https://github.com/fatimaAfzaal/Multiple-Ensemble-models-Diabetes-Prediction-Project-*

*Figure 2 was taken from another source whose reference is included in the figure caption text.*

**Signature:** *Li Xirui* **Date:** 2025/4/11

# Acknowledgements

I would like to express my sincere gratitude to all those who supported and accompanied me during this study.

First and foremost, I want to thank my family. No one loves me more deeply or supports me more unconditionally than you do. Your love has always been my greatest source of confidence, and it continues to be the driving force that motivates me to keep moving forward.

I would like to express my heartfelt gratitude to my supervior, Dr. Ifede Parfait Tebe. He is both rigorous and highly knowledgeable. From the early stages of the project, he consistently encouraged me to review and refine my work. When I encountered difficulties in making progress, he guided me step by step through academic writing styles and techniques. Also, throughout the process of modifying my code, he patiently offered insightful suggestions, which helped me reconstruct my understanding from the ground up. It is no exaggeration to say that the successful completion of this project would not have been possible without Dr. Parfait's dedicated support. Throughout my academic journey, I have met many teachers, but none has impacted me as profoundly as Dr. Parfait.

My gratitude also extends to all the other teachers from Stirling College, Chengdu University, for their dedicated teaching and guidance. I am especially thankful for their high standards and earnest instruction, which have inspired me to adopt a more rigorous approach to my studies. From them, I have not only acquired academic knowledge but also learned how to learn and how to work effectively.

Lastly, I would like to express my heartfelt thanks to my boyfriend, Mr. Xiong Yijian. Thank you for your unwavering patience during my busiest times, and for always offering sincere advice when I felt lost. Your presence is, and will always be, a warm lamp for me in the darkest of nights.

# Table of Contents

# List of Figures

# List of Tables

# 1 Introduction

Diabetes is one of the chronic diseases affecting many people in the world. Predicting it early is then crucial to avoid eventual complications from it [1]. Traditional methods used for the prediction are mostly based on manual interpretation of clinical data, which makes the process slow and even prone to human error. Consequently, machine learning (ML) models have been introduced as efficient tools to improve the disease prediction speed and accuracy. [2]. This study proposes a ML framework that combines two models and an explainable artificial intelligence (XAI) technique. The work proposes a novel ensemble model to improve the prediction performance of diabetes and offer possibilities for physicians and patients to better understand the prediction outcomes.

## 1.1 Background and Context

Known as one of the deadliest diseases nowadays, diabetes has been classified by the World Health Organization (WHO) as one of their top priorities in their plan for chronic disease prevention [1]. In 2019, about 1.5 million death cases from diabetes were estimated by the organisation, and the number increases every year [1]. In order to avoid complications associated with the disease and decrease the mortality rate, WHO and governments continue reinforcing and finding best methods for early prediction and diagnosis.

As technology evolves, AI and ML emerged as promising tools for the prediction and diagnosis of the disease. These cutting-edge technologies provide more accurate and efficient disease prediction and then make diabetes early diagnosis a reality [2]. Current prediction studies mainly focus on ML and deep learning (DL) models and algorithms, and the results are so far encouraging [3], [4], [5]. The models and algorithms have been used to analyse diabetes patients' data, including their demographics, biomarkers and medical history. For examples, models like Support Vector Machines, Random Forest, and Gradient Boosting have been widely used for structured data patterns identification to improve prediction and diagnosis accuracy [5]. DL, specifically neural networks, has also been showing promising results from high-dimensional data such as time-series glucose levels and medical images [4]. Overall, those models help in diabetes prediction for early detection and diagnosis, and help enhance decision-making about patients' conditions.

However, despite the promising and encouraging results from current research works using ML and DL approaches to predict the disease, a few problems still need to be addressed for a more effective prediction. Such problems are identified in the following points.

- Complex Medical Datasets: Diabetes is a very complex disease with multiple effects, and most of the works using traditional ML models face some challenges in training and optimising the models when the dataset is complex. This because these types of datasets often miss some values and also involve high degrees of noisy features, which affects the prediction accuracy [6].
- Imbalanced Medical Data: The concept of ensemble and hybrid models have been developed as more promising and robust approaches to significantly outperform the base models. Such models integrate multiple base together to enhance their individual strengths to provide higher prediction accuracy [7]. However, these types of models mostly require large and high-quality datasets, and in the context of diabetes prediction, they are less effective when the data are not balanced [7].
- Lack of Model Explainability: Most of the base models used to predict diabetes lack explainability, which impedes the adoption of ML-driven solutions in healthcare settings [8]. Particularly in crucial domains like medical diagnostics, clinicians and patients need understand how a model arrives at its decisions[8]. The emsemble and hybrid models are also limited by their complexity, which can lead to challenges in

explainability [8]. Even though the developments in techniques such as SHapley Additive exPlanation (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) have enabled the interpretation of ML models decisions [8], the integration of such techniques into ensemble and hybrid models remains underexplored, particularly for diabetes prediction.

Considering the above listed problems, it urges for proposing more robust and efficient ML based approaches that can provide more precise information in early-stage prediction to improve patients care. There is a need to develop frameworks that can balance prediction performance with transparency.This research work has then proposed an esemble model that combined two base models and an explainable AI technique to further enhance the performance of diabetes prediction performance by handling imbalanced data, and to offer possibilities for physicians and patients to understand how decisions are made.

## 1.2 Scope and Objectives

The scope of this project lies on developing an ensemble machine learning model with explainable artificial intelligence for more precise and efficient diabetes prediction with results explainability. The project used a publicly available dataset from Kaggle with one hundred thousands (100,000) observations and eight (8) features. The features include heart disease, hypertension, smoking history, gender, age, body mass index (BMI), blood glucose level, and HbA1c level [9]. Even though patients' treatment and monitoring are out of the scope of the study, the proposed framework is designed to support large and diverse datasets with strong generalisation capabilities.

The main objectives of the research focus on the following:

- To develop an ensemble ML model that integrates two base models for improved diabetes prediction performance.
- To evaluate the proposed ensemble model to ensure that it provides superior diagnostic support.
- To incorporate an explainable AI (XAI) technique within the proposed model, ensuring that the predictions made by the system can be understood by healthcare professionals and patients.
- To validate the explainability of the proposed framework, making sure it offers clear insights into how predictions are derived, thus increasing trustworthiness and transparency.

Figure 1 depicts a block diagram of the proposed framework. The figure includes the main steps such as input, base models performance evaluation and comparison, best models selection, selected models fusion, ensemble model, classification, XAI, and output.

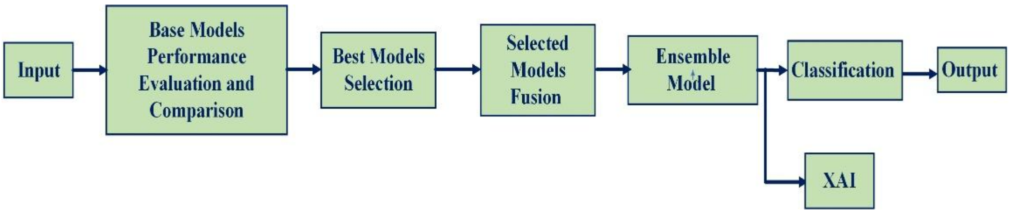


**Figure 1. Block Diagram of the Proposed Framework**

The project has great significance as it is beneficial for potential diabetes patients, healthcare professionals, and society. Indeed, early prediction of the disease using the proposed framework can help potential patients to take preventive actions to avoid complications. It will also allow healthcare professionals to take early action to avoid complications and provide efficient treatment. Early and accurate prediction can also decrease the mortality rate from the disease and then reduce tremendous pressure from society. Furthermore, the research can enhance the trust and usability of AI in diabetes disease diagnosis process by ensuring that the decisions are fairly interpretable and understood by clinicians and patients.

## 1.3 Achievements

The project successfully developed a framework that integrate two ML models as an ensemble model and an XAI. The ensemble model was built using stacking method, and SHAP was used as the XAI technique. The framework achieved a test accuracy of 88.93%, outperforming the base models, and demonstrates its effectiveness of handling complex medical datasets. With its ability to handle class distribution with no significant bias, the model also shows its stability potential in imbalanced data scenarios, ensuring its applicability to diverse clinical dataset. Moreover, the results from SHAP reveal that features such as HbA1c level and blood glucose are the most critical features in the model prediction outcome, and then have a substantial impact on the risk of developing diabetes. This is followed by age and BMI which also have significant impact on the model prediction outcome. Hypertension and heart disease have limited direct impact on the model prediction output.

## 1.4 Overview of Dissertation

The remaining parts of the dissertation are structured as follows:

Chapter 2 provides the research' state-of-the art, including brief overviews of diabetes and machine learning with some of its common models used to predict diabetes. It also introduces the concept of XAI, and finally provides a review of some existing research works on diabetes prediction using machine learning.

Chapter 3 provides the research methodology adopted in details, including the dataset selection and preprocessing, the base models selection, training, and testing. It also describes the proposed framework building and evaluation, including the ensemble model building and the integration of the XAI technique.

Chapter 4 presents the experimental results, discusses the results, and compares the base models and the proposed ensemble model performances. It also explains the results from the XAI integration.

Chapter 5 concludes the study by providing the findings summary, critical evaluation of the achievements, discussion of the work's limitations with directions for future work, and critical reflection on the project.

# 2 State-of-The-Art

This chapter provides a brief overview of diabetes. The disease is briefly explained with its causes and effects. A classification of its different types is also provided. The chapter also provides a brief overview of machine learning and some of its common models used to predict diabetes. The concept of explainable artificial intelligence (XAI) is also introduced. Finally, the chaper provides a review of some existing research works on diabetes prediction using machine learning. Limitations of those works are also pointed out to pave the way for this research.

## 2.1 Diabetes and its Classification

Diabetes is a chronic metabolic condition caused by the human body's inability to produce enough insulin or to adequately use the insulin it produces [4]. Insulin is a human hormone produced by the pancreas, allowing the bloodstream glucose to enter the body cells for energy use. A person is then diagnosed with diabetes if their blood sugar also known as blood glucose is too high. That is, if the level exceeds the normal value. This results from the pancreas' inability to carry out its function in its entirety in the human body [4].

Diabetes is generally categorised or classified into four types: type 1, type 2, gestational diabetes, and related disorders [1], [4]. Type 1 is referred to as insulin dependent and usually affects people under the age of 30 or children. Symptoms of this type include fatigue, increased thirst, frequent urination, weight loss, etc [1], [4]. Type 2 is the most common type worldwide and is referred to as non-insulin-dependent. It is more common in people over 65 and is hence, generally considered as adult-onset diabetes. Its symptoms are similar to those of type 1, but are generally less pronounced or not pronounced at all. As result, this can be undiagnosed for many years and lead to complications [1], [4]. Gestational diabetes is a temporary hyperglycemia condition that generally affects pregnant women and exposes them to the risk of type 2. It also exposes them to the risk of complications during delivery. Symptoms of this type of diabetes are often not noticeable, but the following symptoms may occur sometimes: fatigue, nausea, frequent urination, vaginal infections, etc [1], [4]. Specific associated conditions include long-term damage to organs such as heart, kidneys, eyes, blood arteries, etc., caused by diabetes [1], [4]. An overview of the classification of diabetes into its different types is provided in figure 2.

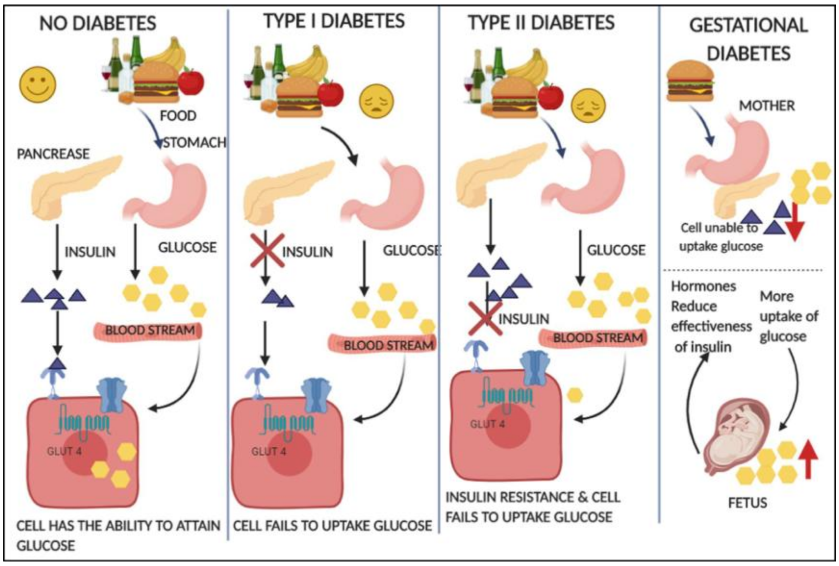


**Figure 2. Classification of Diabetes into Type 1, Type 2, and Gestational Diabetes** [3]

## 2.2 Machine Learning

Machine Learning is a subset of artificial intelligence which consists of using a set of tools to make predictions from data. In other words, machine learning develops computer models or algorithms that learn from data and make predictions or decisions.These models are developed to perform more accurately and effectively over time, depending on their applications and as they process more data [10]. This ability of the models to learn from data and improve over time makes ML a powerful and versatile tool making it the driving force behind the exponential advancements in technology we witness today.

There are three main types of ML, including reinforcement learning, supervised learning, and unsupervised learning. Supervised learning predicts future outcomes based on past data with known labels. Unsupervised learning uncovers patterns in data without known labels to classify future outcomes. Reinforcement learning uses more complex mathematical tools to be trained on a reward signal and predicts an action. There is also a specialised form of ML called deep learning, which utilises neural networks with many layers to solve complex problems [10], [11].

In recent years, ML has been widely applied in numerous fields, including finance, agriculture, robotics, transportation, sports, computer vision, natural language processing, pattern and speech recognition, healthcare, etc. In the healthcare field, trained models are used to predict to predict some disease, and some of the most common models used to predict diabetes disease are provided as follows [6], [7], [10], [11].

- Random Forest Classifier: it is a model that involves multiple base models with basic unit being the decision tree.
- AdaBoost Classifier: it is a model that focuses more on previously misclassified data points in order to pay more attention to them in subsequent training.
- Gradient Boosting Classifier: it is a model that combines multiple regressors or classifiers to form a stronger one to enhance prediction performance.
- Bagging Classifier: It is a model that builds several models by training them on arbitrary data subsets to minimise variance and increase stability.
- Extra Trees Classifier: it is a model that uses multiple decision trees to minimise overfitting and hence, enhance prediction accuracy**.**
- XGBoost Classifier: it is a model that involves multiple base models to compensate for the prediction errors from them and enhance the overall prediction accuracy.
- CatBoost Classifier: It is a model with high accuracy and speed that can instantly handle categorical variables without requiring one-hot encoding.
- Passive Aggressive Classifier: It is a model that is efficient for large-scale learning problems.

## 2.3 Explainable Artificial Intelligence (XAI)

Explainable AI, also known as interpretable AI is an AI technique used to help human to have an oversight and understanding of the decision made from a ML-based predictions. It deduces ML model predictions and offers meaningful explanations for the obtained results to make them clear and trustful for human beings [8], [12].

Incorporating XAI with prediction-based ML models has several benefits. It provides a better understanding of the rationale behind prediction results and helps make better-informed decisions. It also helps identify potential errors or biases in the trained models, and hence leads to more accurate and fair outcomes [8]. All those benefits result from some principles behind XAI technique as depicted by figure 3.

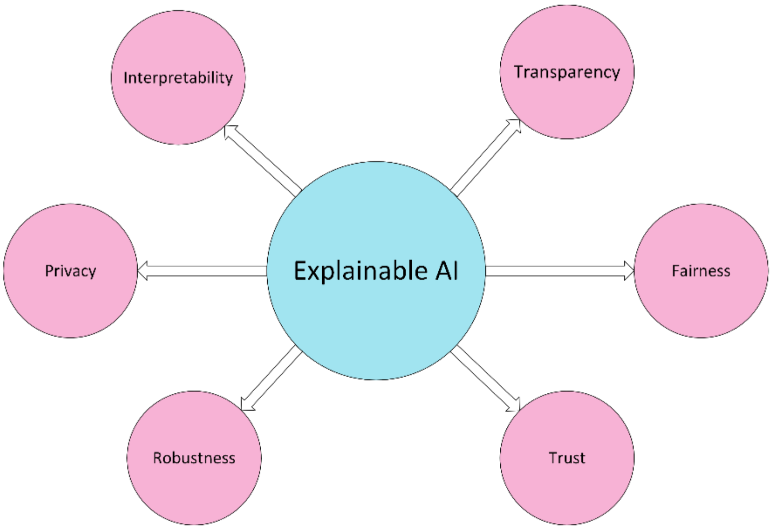


**Figure 3. Explainable AI Principles**

The most common XAI techniques used in predicting diseases in general, and diabetes in particular are SHAP) and LIME. The values from SHAP technique indicate how each data feature influences final prediction, and how significant it is when compared to others. LIME technique selects specific parts of the model's prediction that we desire to understand generates datasets for those parts [8], [12].

## 2.4 Diabetes Prediction Using Machine Learning

### 2.4.1 Related Works

Several studies have explored the use of ML models for diabetes prediction. Some of the most current ones are summarised as following.

In [13], the authors introduced a novel approach to diabetes prediction by using a weighted ensemble of ML classifiers. The resulted ensemble model improves the prediction performance by achieving an accuracy of 73.5%.

Ashisha et al. [14] used Random Forest and Extra Trees classifier for diabetes prediction based on clinical data from Sree Guru Hospital. Their results show higher accuracy for the Extra Trees classifier with a value of 96%.

Several ML algorithms, such as Random Forest, XGBoost, CatBoost, Gradient Boosting, and LightGBM were deployed in [15] to predict diabetes. The work results show the superiority of XGBoost performance over other models with an accuracy of 94%.

In [16], a novel BP-XGBoost-RF ensembled model is proposed for diabetes prediction. The authors combined the three base models to minimise the risk for missed diagnosis and improve the accuracy of the prediction. Their results show 95% and 94% for accuracy and F1 score, respectively.

Alzubaidi et al. [17] proposed a novel stacking ensemble model for diabetes mellitus prediction, combining Random Forest and Logistic Regression as base learners and XGBoost as a meta-learner. The model achieved an accuracy of 83%.

A hybrid ML approach for diabetes prediction using XGBoost and AdaBoost is proposed in [18]. The proposed model outperformed some individual models by achieving an accuracy of 85%.

Similar to [18], the authors of  proposed a hybrid ML model that combines ensemble techniques such as Random Forest, Bagging, and deep learning methods like Neural Networks to improve diabetes prediction. Their hybrid approach demonstrates an accuracy of 80%, surpassing individual models.

Another hybrid model combining XGBoost and Extreme Machine Learning (EML) for diabetes prediction has been proposed in [20]. By leveraging the strengths of both algorithms, the hybrid model outperforms traditional models like Random Forest and Logistic Regression with an accuracy of 92.85%.

In [21], a diabetes prediction based on hybrid ML model is proposed. The proposed hybrid model integrated Random Forest, Support Vector Machine, and Decision Tree, and outperformed individual models by achieving an accuracy of 90.1%.

A framework combining AdaBoost and XGBoost classifiers is proposed in [22] for diabetes prediction. The framework was enhanced by data preprocessing techniques and outperformed individual classifiers by achieving an accuracy of 90.4%.

### 2.4.2 Limitations from the Related Works

Despite their significance progress in applying ML to predict diabetes, the above-mentioned works and the majority of the existing related works in general, have some challenges and limitations. One of the foremost challenges lies on the quality of data used. The works are limited by datasets that are not large enough, not balanced, and not free from biases [13], [17], [21], [22]. Also, despite the ensemble and hybrid models' robustness and high accuracy, improving the models' ability to adapt to complex datasets and ensure consistent performance across diverse scenarios is needed [14], [15], [16]. Moreover, while efforts have been made to enhance explainability, achieving a balance between complexity and transparency remains a challenge [18], [19], [20].

Overall, existing works based on ML to predict diabetes present a gap in developing models that are both highly accurate and explainable, especially when dealing with complex, multi-dimensional medical datasets. Hence, this work aims to offer a solution to the gap by proposing an ensemble model that integrates two ML models to improve predictive performance while incorporating an explainable AI technique to ensure that the decisions are fairly interpretable and understood by clinicians and patients, thereby enhancing the trust and usability of AI in the disease diagnosis process.

# 3 Methodology

This chapter presents the research methodology, including the data collection, and preprocessing, the ML models used, the models training processes, the ensemble model building, evaluation and explainability. The methodology is firstly summarised in a flowchart presented in figure 4 before the details of each step. Also, prior to the methodology details, the different software and libraries used for the project coding and documentation are provided.

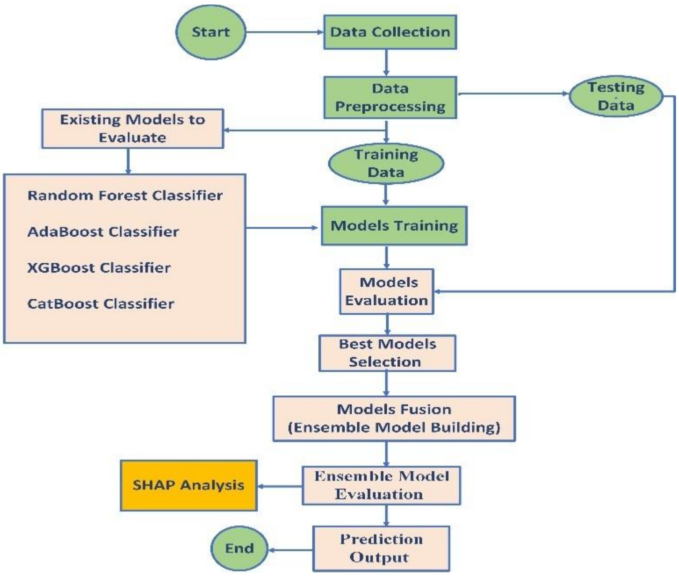


**Figure 4. Flowchart of the Proposed Framework**

## 3.1 Software and Libraries Used

The development, experimentation, and documentation of this project are done on a Lenovo Savior Y9000P laptop computer, intel core i9, with windows 11 as the opertation system. The programming language used for the project codes is Python(version 3.9.19), and the codes were written and executed in the integrated development environment (IDE) in Visual Studio Code (VScode) for fast startup speed, low resource consumption, and smooth operation.

The following libraries were also used in the codes:

**Pandas:** This open-source Python library was used in the project codes to efficiently handle the dataset by allowing easy reading of the CSV data, convenient storage, and data operations. It was also used to offer rich data analysis functions, such as correlation analysis and aggregation.

**Scikit-learn:** Abbreviated as Sklearn, this library was used in the project codes to support data preprocessing and models selection and evaluation tools.

**Imblearn:** This library was used in the project codes to provide the sampling method to handle class imbalance in order to improve the models performance.

**Matplotlib:** This library was used in the project codes to provide plots needed for data visualisation and analysis.

**Seaborn:** This library was used in the project codes to provide a variety of built-in chart types for adanced data visualization.

**Joblib:** This library was used in the project codes for parallel computing purpose. It was used for saving and loading models, as well as handling data processing quickly, making it convenient to reuse the models and data processors later without retraining, thereby improving efficiency.

**Shap:** This library was used in the project codes to explain the prediction results of the proposed model by assigning a specific contribution value to each feature, helping users understand how the model makes predictions based on the input features.

## 3.2 Data Collection and Preprocessing

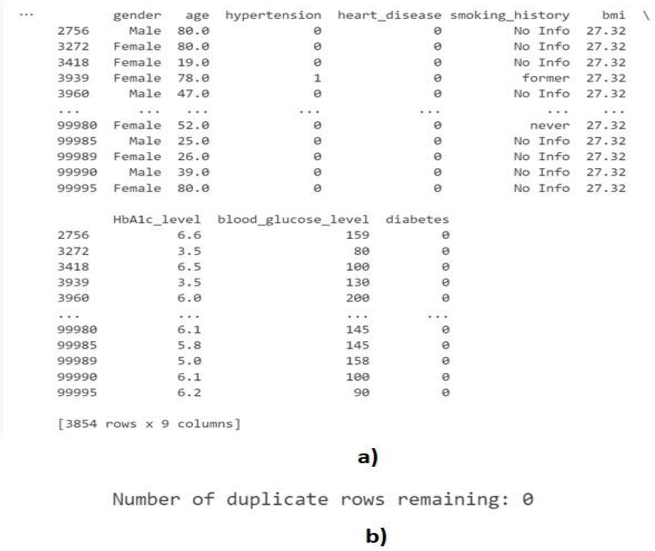
### 3.2.1 Data Collection

A publicly available dataset from Kaggle was used for the project [9]. The dataset includes a total of one hundred thousands (10,0000) observations and eight (8) features. The features are described as follow:

- Gender: representing both women and men
- Age: continuous numerical value, indicating the patient's age
- Hypertension: binary variable, where “1” represents having hypertension, and “0” represents not having hypertension
- Heart disease: two categorical variables, where “1” represents having heart disease, and “0” remains unchanged without heart disease
- Smoking history: categorical variables, where “Never” represents never smoking, “No Info” represents no information, “Current” represents current smoking, “Former” represents past smoking but has quit, and “Ever” represents past smoking and still present
- BMI: representing the body mass index in kg/m$^2$
- HbA1c level: continuous numerical value, representing the level of glycated hemoglobin
- Blood glucose level: Integer value representing the patient's blood pressure level

### 3.2.2 Data Preprocessing

As first step in this project data preprocessing, missing values and duplicate values were checked to understand the data quality and prevent redundancy, erroneous and biases in analysis and modeling. The checking showed that there was no missing value in the dataset, but there were 3854 duplicate values. Therefore, the “drop_duplicates()” syntax was used to remove all the duplicate values. The detection and removal of the duplicate values are shown in figure 5.



```
...       gender   age  hypertension  heart_disease smoking_history    bmi  \
2756        Male  80.0             0              0         No Info  27.32
3272      Female  80.0             0              0         No Info  27.32
3418      Female  19.0             0              0         No Info  27.32
3939      Female  78.0             1              0          former  27.32
3960        Male  47.0             0              0         No Info  27.32
...          ...   ...           ...            ...             ...    ...
99980     Female  52.0             0              0           never  27.32
99985       Male  25.0             0              0         No Info  27.32
99989     Female  26.0             0              0         No Info  27.32
99990       Male  39.0             0              0         No Info  27.32
99995     Female  80.0             0              0         No Info  27.32

       HbA1c_level  blood_glucose_level  diabetes
2756           6.6                  159         0
3272           3.5                   80         0
3418           6.5                  100         0
3939           3.5                  130         0
3960           6.0                  200         0
...            ...                  ...       ...
99980          6.1                  145         0
99985          5.8                  145         0
99989          5.0                  158         0
99990          6.1                  100         0
99995          6.2                   90         0

[3854 rows x 9 columns]
```

a)

```
Number of duplicate rows remaining: 0
```

b)

Figure 5. Duplicate Values Handling: a) Detection, b) Removal

Outlier detection was the next step of the data preprocessing. The process consists of detecting and removing or adjusting outliers, and it was used to improve data quality and prevent erroneous analysis, and to make the proposed model more robust and to enhance its generalisation ability [23]. Figure 6, which is a box plot was used to check for outliers, and figure 7 shows the exact percentages. It was found  that the percentage of outlier in blood glucose level is 2.11% with 2031 as the number of extreme values, the percentage of outliers in the HbA1c level is 1.36% with 1312 as the number of mechanics, the percentage of extremums in BMI is 5.57% with 5354 as the number of mechanics.

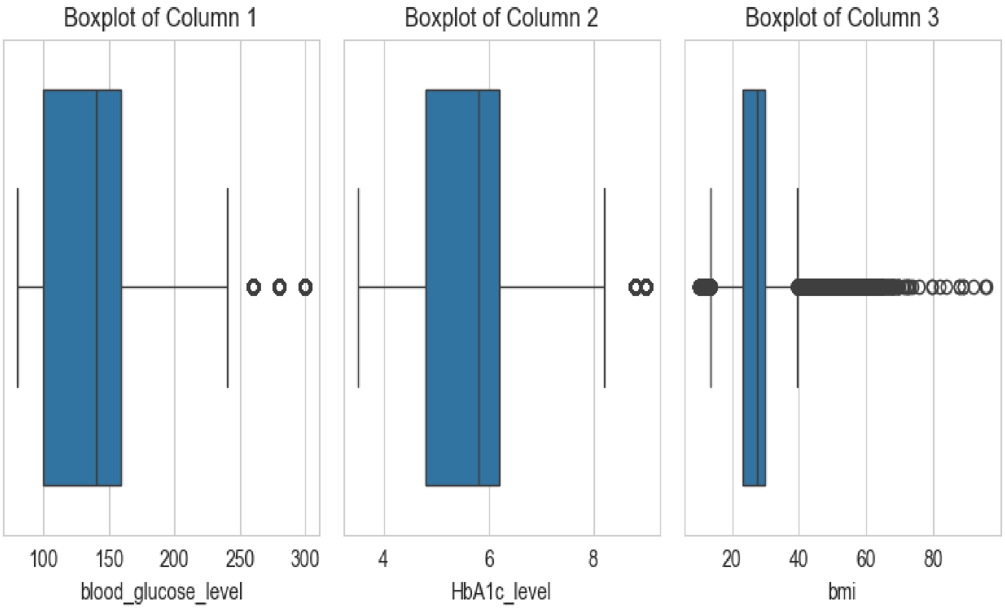
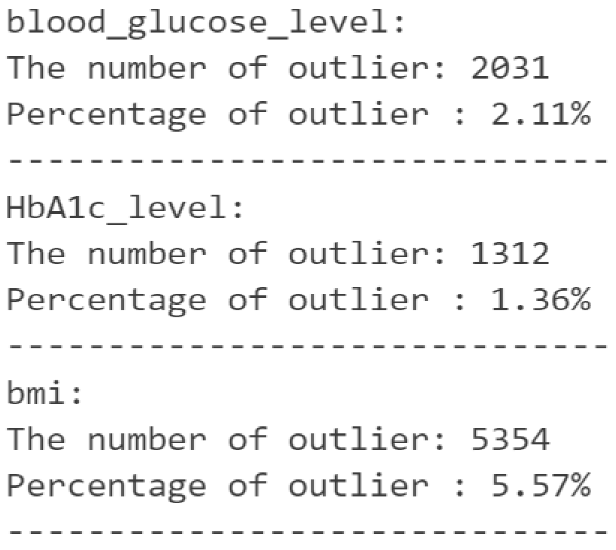


**Figure 6. Outliers Detected for Three Features**



```
blood_glucose_level:
The number of outlier: 2031
Percentage of outlier : 2.11%
------------------------------
HbA1c_level:
The number of outlier: 1312
Percentage of outlier : 1.36%
------------------------------
bmi:
The number of outlier: 5354
Percentage of outlier : 5.57%
------------------------------
```

**Figure 7. Outliers Percentages**

After these detections, the extreme values were removed using the median method [24], and the result is shown in figure 8.

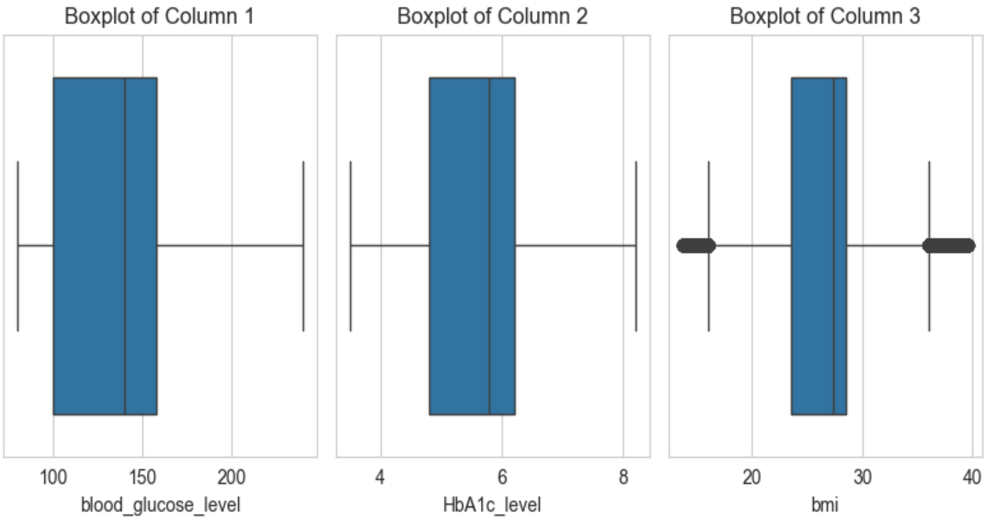


**Figure 8. Result from Outliers Removal**

After dealing with outliers, one-hot encoding was performed to convert the two features, namely "gender" and "smoking history", which were categorical variables in the dataset into binary variables that can be processed by the models.

Given that using imbalanced dataset to train models to predict diabetes is one of the limitations from existing studies that this project aimed to address, it was necessary to check the dataset class distribution after the one-hot encoding. This was to determine whether the data was balanced or not. As depicted by figure 9, the result from checking showed an imbalanced data distribution. Specifically, the result showed that the dataset contained 84405 negative cases (labeled as 0) and 8156 positive cases (labeled as 1). This indicated a noticeable imbalance in class distribution, which may impact the models performance. For this project for example, the imbalance distribution could lead the models to favor predicting the negative cases which constitute the larger class, and ignoring the positive cases which constitute the minority class, and this would affect the overall prediction performance.

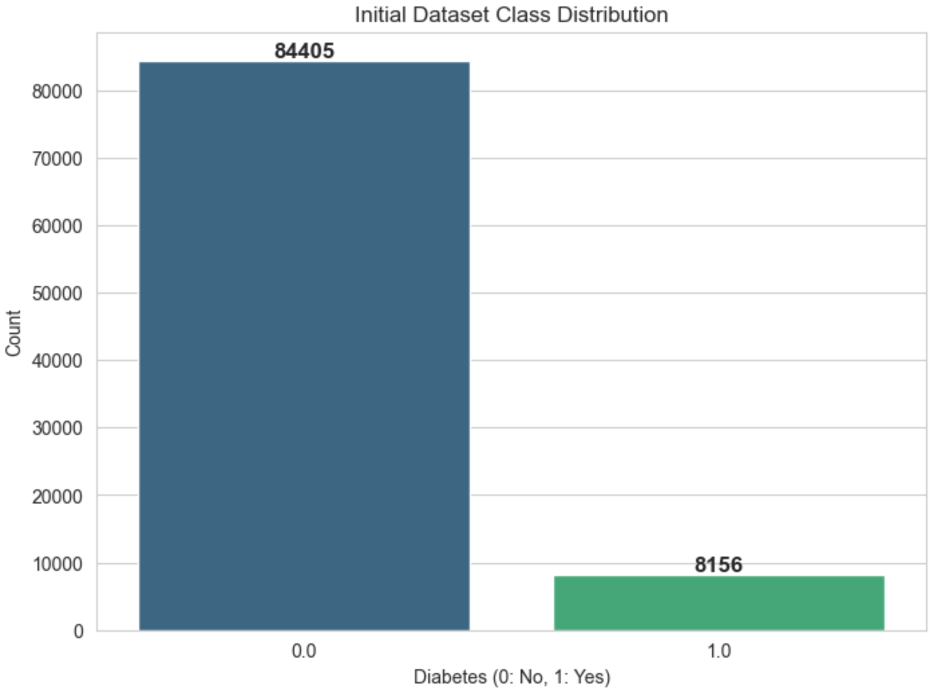


**Figure 9. Initial Dataset Class Distribution**

It was then necessary to handle the class imbalance by adjusting the distribution to prevent overfitting to the majority class and underfitting to the minority class. Class imbalance handling generally reduces bias toward majority class, makes accuracy become a more reliable metric, and improves precision, recall, and F1-Score [25]. There are several techniques to handle class imbalance. Common techniques include oversampling and undersampling which adjust the class distribution to prevent overfitting to the majority class and underfitting to the minority one, and synthetic minority oversampling technique (SMOTE) which can generate some new samples for the minority class to enhance the model's ability to learn from it [26]. In this work, only downsampling technique helped to handle the class imbalance, and the result is depicted by figure 10. Even though a perfect 50:50 balance was not achieved, the result showed that the dataset was almost balanced with 8127 number of negative cases against 8080 number of positive cases, and this has provided a more comprehensive assessment of the models performance.

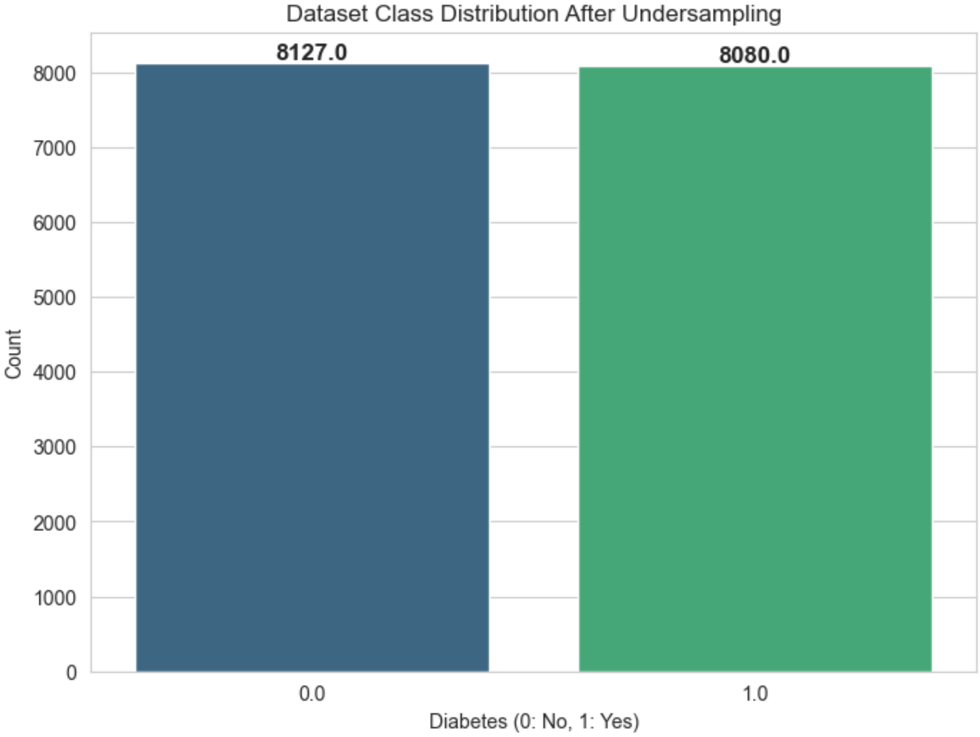


**Figure 10. Dataset Class Distribution After Undersampling**

After handling the class imbalance it was necessary to study the feature correlation of the project dataset as it helps identify redundant features and reduce dimensionality [27]. Highly correlated features may contain similar information, leading to multicollinearity, which can affect the coefficients stability in some models [27]. By computing the correlation matrix, highly correlated features can be identified, and redundant features can be removed or dimensionality reduction techniques can be applied to improve the model's generalisation ability. Additionally, analysing feature correlation helps in selecting an appropriate modeling approach [27], [28]. In this study, a heatmap was then used to visualise the correlation between different features, and the result is shown in figure 11. As it can be observed, HbA1c level exhibits the highest correlation with diabetes with 0.30, followed by age with a correlation of 0.26, blood glucose level with a correlation of 0.22, hypertension with a correlation of 0.20, BMI and heart disease with correlation of 0.17 each. So, overall, the result showed no redundant features detected, and the correlated ones are not high.

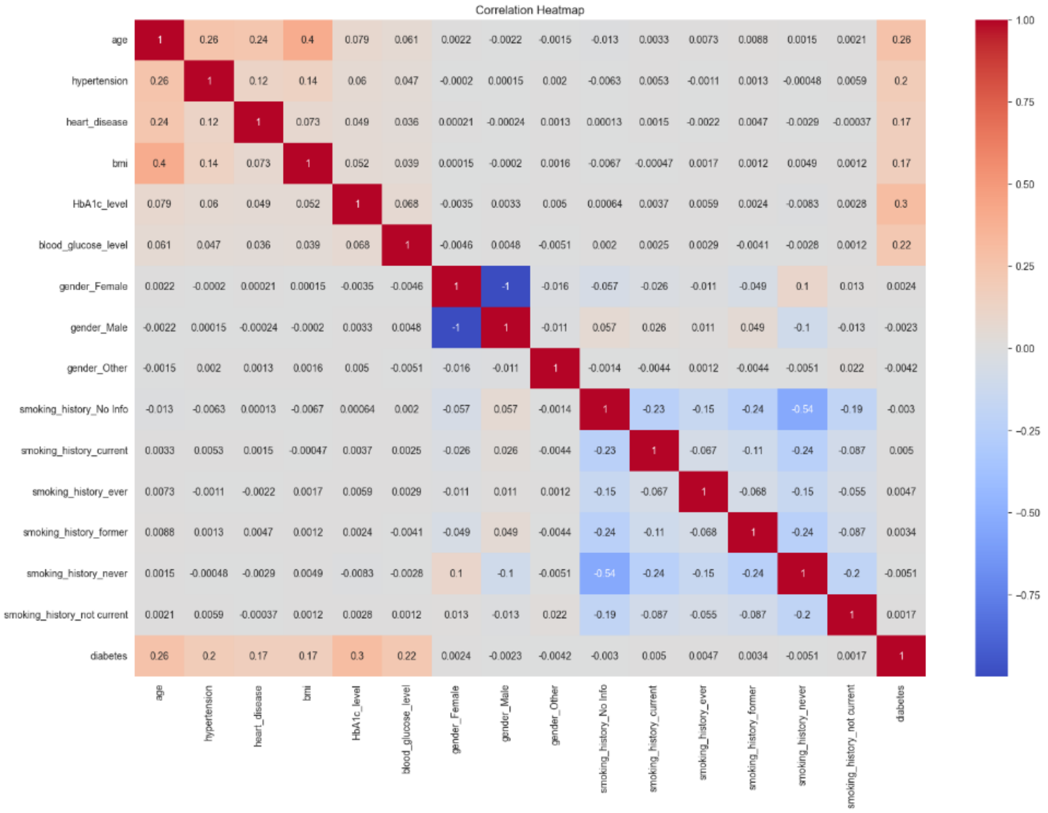


**Figure 11. Heatmap Showing Correlation Between Different Features**

Based on the feature correlation result, it was a wise decision to prevent certain features from having an excessive impact on the models performance, even though their correlations are not high. Data standardisation was then performed to ensure some features do not overshadow others due to scale differences.

Lastly, the dataset was splitted into training and test sets to ensure proper model training. This helped prevent overfitting by separating training data from test data and also helped evaluate the models generalisation ability. The 8:2 ratio was used to split the data.

## 3.3 Base Models Training

### 3.3.1 Models Selection

To accomplish the project' goals objectives, the following base models were selected. The selection was based on the models unique abilities in handling classification tasks and their complementary characteristics in an ensemble model [6], [7].

- Random Forest Classifier: It was selected because of its suitability for processing high-dimensional data and automatically handling missing values and outliers. By integrating multiple decision trees, the excessive risk of a single decision tree is reduced and the generalisation ability of the model is improved. With its ability to easily capture non-linear patterns, it can help identify the features that are most critical for the prediction of diabetes [6].
- AdaBoost: It was selected because of its capability to focus on errors made by other classifiers to improve the prediction ability of the overall ensemble model. It can give extra attention to some some diabetes patients profiles that are difficult to classify [6].
- XGBoost Classifier: It was selected because it is one of the most computationally efficient, powerful and suitable classifiers for large-scale datasets. It has built-in

regularisation terms helpful for overfiting control and generalisation enhancement and can then provide more consistent performance in diabetes prediction tasks [6], [29].

- CatBoost Calssifier: It was selected because of its unique advantage with dealing well with structured medical data. Its training process is fast, and it can achieve better performance under default parameters, reducing the cost of tuning parameters. It provides robust generalisation and is known for its ability to make prediction more reliable when dealing with unseen patient data, and to perform well in diabetes prediction scenarios [6].

### 3.3.2 Models Training

The training procedure of each base model is described in this subsection. Each model was separately and carefully trained to learn relevant patterns from the dataset, ensuring proper parameter initialisation, and the training process reproducibility.

The Random Forest Classifier was trained first by bootstrapping multiple subsets of training data. Each tree was trained independently on a different subset, and the final prediction is the set of all the trees. The following hyperparameters were used for the training: n_estimators = 50 for 50 decision trees in the forest; random_state = 5 to ensure repeatability; control over max_depth = 2 to limit the depth of each tree to 2 to control overfitting; min_samples_split = 2 as the minimum number of samples required to split an internal node; min_samples_leaf = 4 to make sure there are at least 4 samples per leaf node.

AdaBoost Classifier was trained next. Thirty (30) weak classifiers (typically decision stumps) were used by the model. Each subsequent classifier adjusted the weights to focus on the samples that were misclassified in the previous iteration, and the final prediction was a weighted combination of all classifiers. The following hyperparameters are were used: n_estimators =30 for 30 decision trees; random_state = 5 to ensure repeatability.

XGBoost Classifier was trained next with the following hyperparameters: n_estimators = 50 for 50 gradient boosting trees constructed; random_state = 42 to ensure repeatability. Each tree was destined to attempt to correct the residual error of the previous tree, with the loss function guiding how to build a new tree.

CatBoost Classifier was trained last with the following hyperparameters: iterations = 50 representing the number of promotion rounds (tree); learning_rate = 0.1 to control the step size of each update; depth = 6 for the maximum depth of each tree; random_state = 42 to ensure repeatability; verbose = 0 to disable verbose output.

### 3.3.3 Models Evaluation on Test Set

Model evaluation on test set serves to assess its performance on unseen data, thus measuring its generalisation ability. For each trained model of this project, the test set was then used to evaluate some performance metrics and compare them with the ones from the training set to ensure there is no overfitting on the training data, and to ensure that the model can make accurate predictions on new, unseen data in real-world applications. The performance metrics include [30]:

- Accuracy which represents the proportion of samples that are correctly predicted out of the total number.
- Precision which indicates the proportion of the actual positive samples among all the ones predicted as positive by the model.
- Recall which evaluates models ability to detect true positive cases.
- F1 Score which is the recall and precision harmonic mean that provides a balanced measure of models' correctness and their ability to identify positive cases.

The values of these performance metrics from the test set were used not only to judge whether there was overfitting, but also to compare the base models in terms of performance and complexity which served as a foundation for model selection for the ensemble framework described in the next section.

## 3.4 Ensemble Model Building

The results from the base model evaluation based on the test set showed best prediction performance from CatBoost classifier, followed by XGBoost classifier, AdaBoost classifier, and Random Forest classifier, in that order. Given that one of the main goals of this project is to maximise predictive performance and enhance generalisation, CatBoost and XGBoost were selected as the base learners to propose the ensemble model. The rationale for this selection includes:

- Empirical Performance: Both models outperformed the other models across all the evaluation metrics, and this indicates their strong learning capability on the dataset.
- Complementarity: Their prediction performance analysis suggests different types of errors from them, which could be beneficial in making better final predictions when combining their strengths.
- Model Diversity: Even though both models are gradient boosting based, they differ from each other in key aspects such as optimisation techniques and handling categorical features [6], [29]. This diversity could help in correlated errors risk minimisation and enhance the ensemble's robustness.
- Reduced Complexity: Using only the top two performing models to build the ensemble model limits the framework computational complexity.

### 3.4.1 Approach Used: "Stacking"

The stacking approach was used to build the ensemble model. It is a meta-learning technique that integrates the predictions of several models through a meta-classifier in order to leverage their complementary strengths to form a more robust final prediction [31]. The predictions from CatBoost and XGBoost classifiers were then combined using stacking, with linear regression as the meta classifier in the ensemble method. The stacking ensemble model architecture is presented in figure 12. It includes the two base models and the meta classifier.

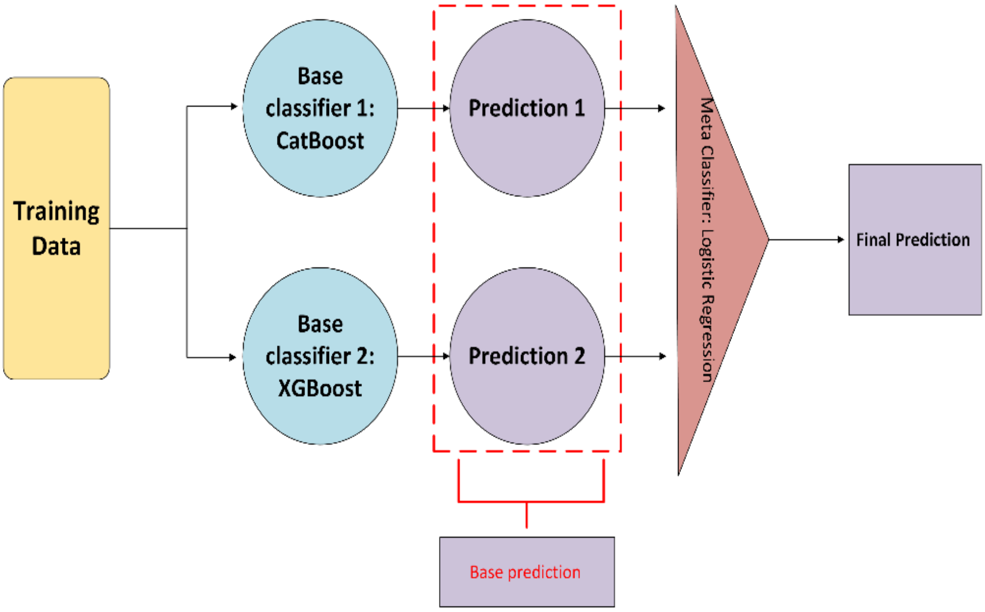

Training Data
Base classifier 1: CatBoost
Base classifier 2: XGBoost
Prediction 1
Prediction 2
Meta Classifier: Logistic Regression
Final Prediction
Base prediction

**Figure 12. Stacking Esemble Model Architecture**

Both models use the training dataset to independently learn patterns from the data and generate their respective predictions which serve as the input features for the meta classifier to make the final decision. CatBoost adopts a symmetric tree structure along with the Ordered Boosting technique, and dynamically adjusts the weights of features during training according to their importance, ensuring that more relevant features have a greater influence on the model's prediction [31]. XGBoost automatically performs feature selection during the training process by evaluating feature importance based on criteria such as information gain and Gini index, effectively filtering out less important features and suppressing noisy ones [31].

In order to instantiate the stacking claasifier, the following parameters were used inside the meta classifier:

- Saga solver was used for large-scale data to support L1/L2 regularisation, with the advantages of fast speed and high efficiency. L1 and L2 stand for Lasso regularisation and Rigid regularization, respectively [32].
- Penalty was used for overfitting prevention, with L2 and C = 0.1 indicating appropriate enhancement of regularisation [33].

### 3.4.2 Ensemble Model Evaluation on Test Set

The built ensemble model was evaluated on the test set to assess its prediction performance and generalisation. In addition to the metrics used to evaluate the base models, confusion matrix was also used in the evaluation to provide a representation to summarise the model performance.

The confusion matrix contains four elements such as true positives (TP) which represent the number of patients who have diabetes and are predicted to have it, false positives (FP) representing the number of patients who don’t have the disease but predicted to have it, true negatives (TN) representing the number of patients who don’t have the disease and predicted as not having it, and false negative (FN) representing the number of patients who have the disease but predicted as not having it [30].

## 3.5 Ensemble Model Explainability Using SHAP

After evaluating the ensemble model, SHAP was applied as the explainable AI technique for its prediction interpretability. SHAP values were computed for each feature in the test set, where each of them represents the marginal contribution of the corresponding feature to the final prediction [8]. A positive value of SHAP indicates a positive impact of the associated feature on predicting the positive class, while a negative value suggests a negative contribution to the positive class prediction [8]. The process is shown in the following steps:

- Step 1: The SHAP values were computed on the ensemble model to find out which base model has more impact on its building, and the result showed that it was XGBoost.
- Step 2: The SHAP values were computed on XGBoost to find out the features that have more impact on it, and it was concluded that those features are the ones that impact the ensemble model the most.

After computing the SHAP values, a summary plot was used to illustrate the overall impact of each feature on the ensemble model. In addition, a feature importance bar chart was generated, where features are ranked according to their mean absolute SHAP values, clearly showing the degree of influence each feature has on the model’s prediction results.

# 4 Results and Discussions

This chapter provides the results from the performance evaluations of the base models. The evaluation results of the proposed ensemble model is also provided in the chapter, followed by the SHAP performance analysis.

## 4.1 Base Models Performance on Training Set

The training results of the base models are provided in table 1 and show that XGboost outperformed the other three across all the metrics, with 0.9288, 0.9302, 0.9288, and 0.9288 for accuracy, precision, recall, and F1-Score, respectively. This is followed by CatBoost, AdaBoost, and Random Forest in that order. These results indicate that XGBoost has a lower misclassification rate by effectively learning from the large dataset used in this project, confirming its choice justified in chapter 3.

**Table 1 Training Results of the Base Models**

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Random Forest | 0.8611 | 0.8640 | 0.8611 | 0.8608 |
| AdaBoost | 0.8923 | 0.8937 | 0.8923 | 0.8922 |
| XGBoost | 0.9288 | 0.9302 | 0.9288 | 0.9288 |
| CatBoost | 0.8963 | 0.8994 | 0.8963 | 0.8961 |

## 4.2 Base Models Performance on Test Set

The performance of the models were evaluated on test set to ensure there is no overfitting and for generalisation purpose, and the results are presented in table 2. It can be observed from the table that the performance values have slightly dropped for each metric and for each model, but the overall results show that there was no overfitting. Moreover, the results show the superiror performance of Catboost over all the other three models across all the metrics on the test set, followed by XGBoost, AdaBoost, and Random Forest in that order. These results also confirmed the choice of CatBoost for this project as justified in chapter 3 that it could provide robust generalisation, make prediction more reliable when dealing with unseen patient data, and perform well in diabetes prediction scenarios.

A clearer performance comparison of all the models is provided in figure 13. The figure compares only the models accuracies from training set and test set. In summary, CatBoost and XGBoost emerged as the best two performing models not only in terms of prediction accuracy, but also in terms of complexity and generalisation balance. This, then confirms their choice to build the ensemble model as justified in section 3.4 of chapter 3.

**Table 2 Test Results of the Base Models**

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Random Forest | 0.8510 | 0.8538 | 0.8510 | 0.8507 |
| AdaBoost | 0.8846 | 0.8865 | 0.8846 | 0.8845 |
| XGBoost | 0.8868 | 0.8888 | 0.8868 | 0.8866 |
| CatBoost | 0.8871 | 0.8912 | 0.8871 | 0.8866 |

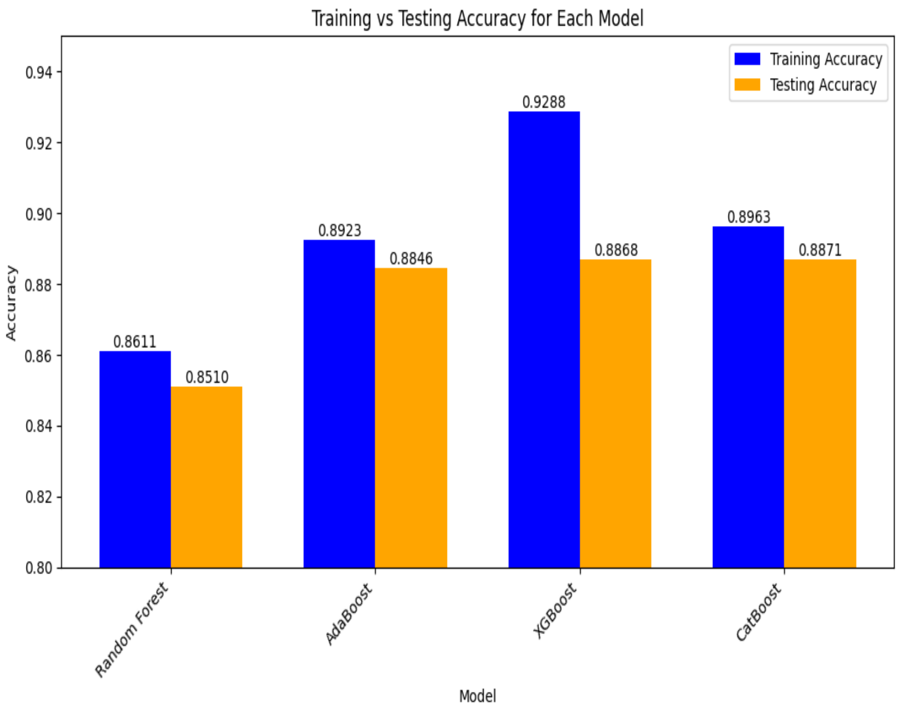


**Figure 13. Comparison of the Base Models Accuracy on Both Sets**

## 4.3 Ensemble Model Performance Analysis

The ensemble model performance on both training and test sets is provided in figure 14. The results indicate that the metrics are highly consistent for both sets, indicating that the model's predictions across all categories are balanced with no significant bias.

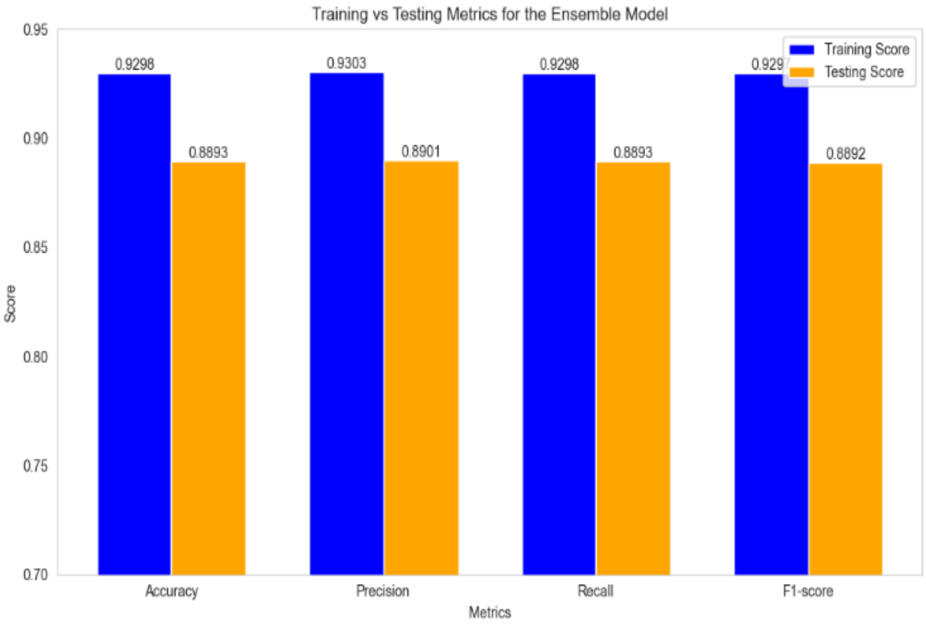


**Figure 14. Ensemble Model Performance Results**

The performance of the ensemble model is next, compared with that of the best base model, with focus on the prediction accuracy from the test set, and the results are depicted by figure 15. With an accuracy of 0.8893 against 0.8871 for the best base model, this demonstrates that the ensemble model outperforms the best base model. Based on the results, it can be concluded that the performance of feature-based ensemble models is feasible and effective. The result also demonstrates that one of the objectives of this project is achieved.

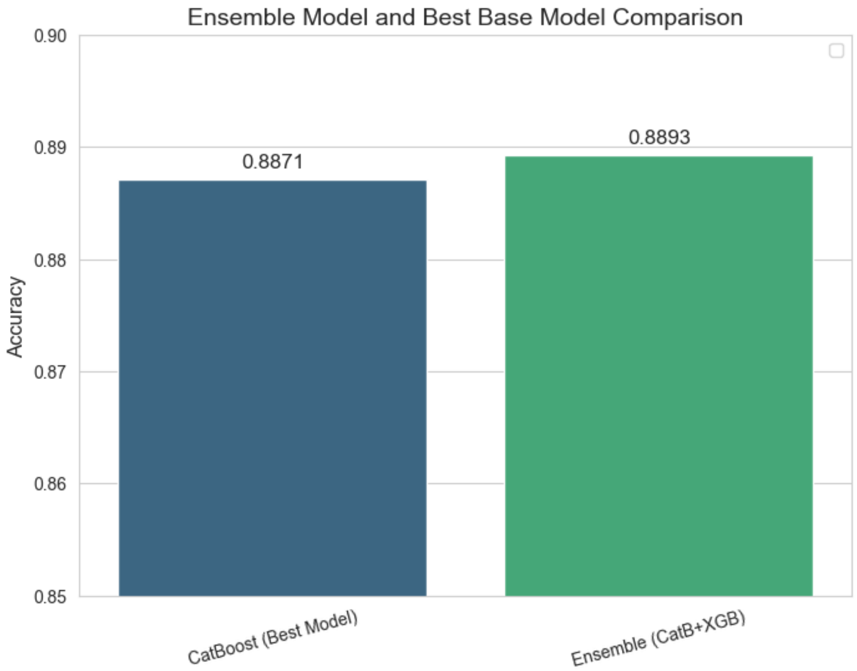


**Figure 15. Ensemble Model and Best Base Model Performance Comparison**

In order to illustrate the ensemble model's predictive performance across different classes and analyse its strengths and weakness, a confusion matrix was plot and the result is shown in figure 16. According to the result, the model demonstrates high precision, as 1,480 out of 1,622 samples predicted as "Positive" are true positives. Additionally, among the 1,697 samples predicted as "Positive," 1,480 are actually positives, indicating a reasonable recall rate. The model also shows a good handling of class distribution (TN = 1,403, TP = 1,480) with no significant bias.

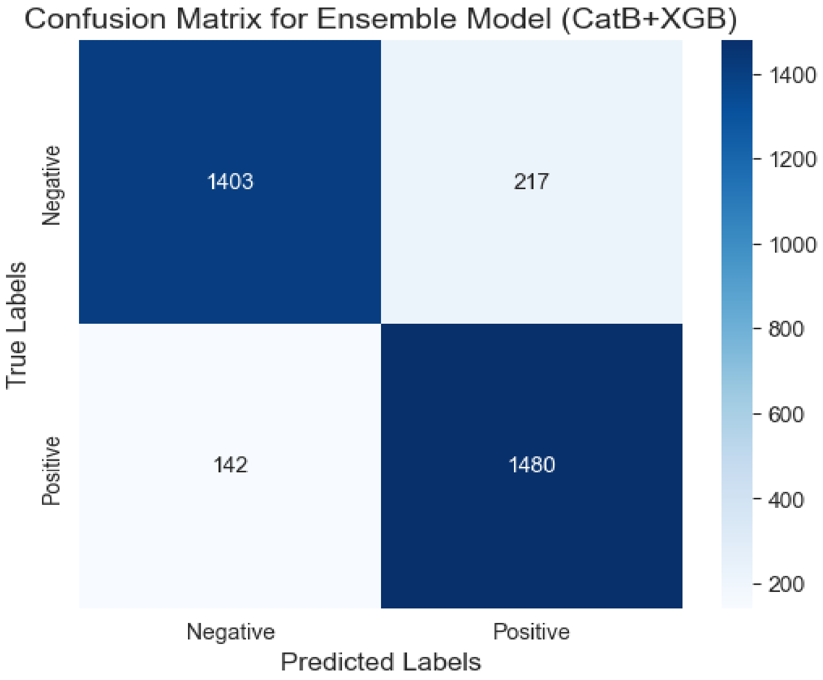


**Figure 16. Confusion Matrix of the Ensemble Model**

## 4.4 SHAP Performance Analysis

In order to illustrate the overall impact of each feature on the ensemble model, a summary plot of the result from the computed SHAP values on the model is first, provided. The objective is to find out which base model has more impact on the ensemble model prediction output. The result is depicted by figure 17 and shows that the SHAP values of XGBoost are higher than the ones of CatBoost. Moreover, the higher value has positive impact for Class 1 in XGBoost, while it has negative impact on its Class 0, which means that this feature has significant discriminatory power within the XGBoost model.

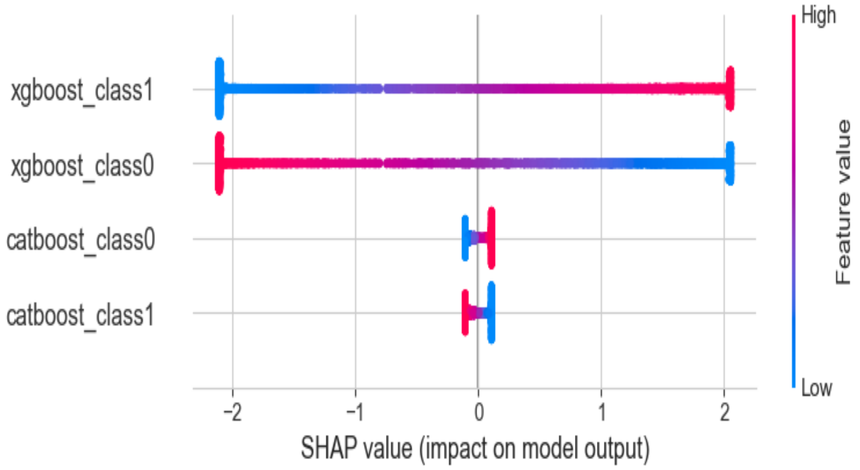


**Figure 17. SHAP Summary Plot Showing the Impact of the Base Models Features on the Ensemble Model Prediction Output**

Next, the SHAP values are ranked in a bar chart according to their absolute mean to clearly show their degree of impact on the model prediction output, and the result is shown in figure 18.

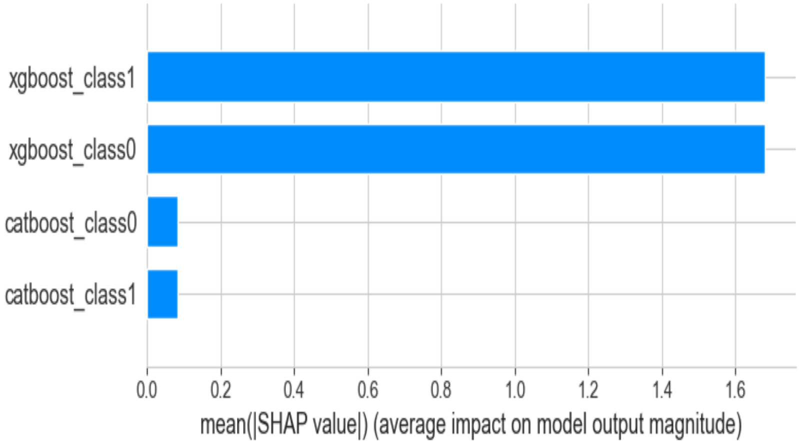


**Figure 18. SHAP Plot Showing the Degree of Impact of the Base Models Features on the Ensemble Model Prediction Output**

The result shows that the magnitudes of XGBoost Class 1 and XGBoost Class 0 are higher than the ones of Catboost Class 1 and CatBoost Class 0, which illustrates that the precision of XGBoost depends on the contribution of those features. Moreover, the magnitudes of XGBoost Class 1 and XGBoost Class 0 are equal, meaning that those features have similar impact, and do not particularly favor one class over the other. The magnitudes of CatBoost

Class 1 and CatBoost Class 0 are also equal, which also means that those features have similar impact, and do not particularly favor one class over the other.

The overall results from both plots show that the SHAP values for CatBoost are insignificant and then very negligeable compared to the ones of XGBoost. It can then be concluded that the features learned from XGBoost have more significant impact on the ensemble model prediction output. Therefore, to determine the features names, the SHAP values were computed on XGBoost, and the results are shown in figure 19 and figure 20.

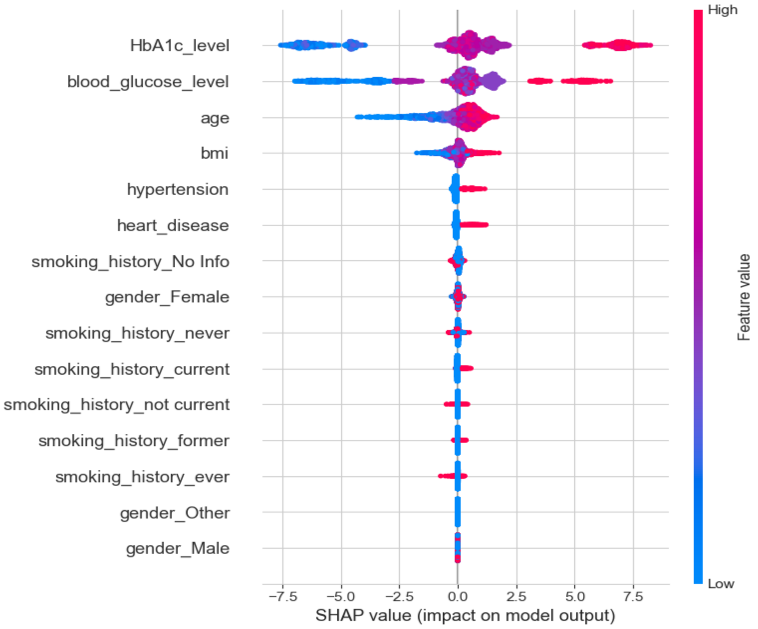


**Figure 19. SHAP Summary Plot Showing Feature Impact on the Ensemble Model Prediction Output, Using XGBoost**

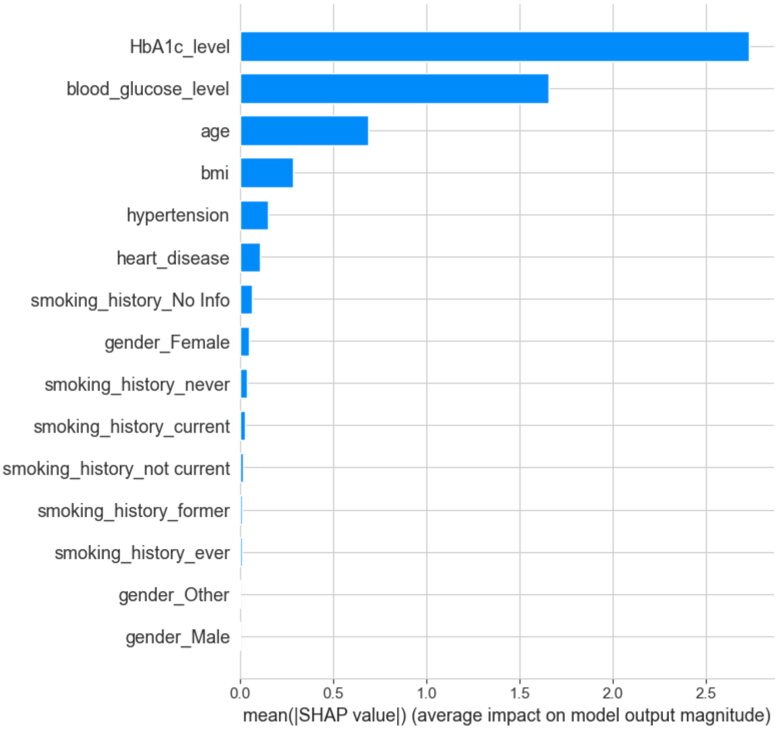


**Figure 20. SHAP Plot Showing the Degree of Impact of the Features on the Ensemble Model Prediction Output, Using XGBoost**

Figure 19 shows the summary plot of the features impact on the ensemble model prediction output while figure 20 shows the degree of impact of the features. The analysis of the results from both figures shows that HbA1c level and blood glucose level exhibit the highest SHAP values, indicating that they are the most critical features in the model prediction outcome. This reflects the medical understanding that elevated HbA1c and blood glucose levels directly correlate with diabetes risk. Age and BMI show the next highest SHAP values, and the values demonstrate that both features also have a significant impact on the model prediction outcome. This aligns with the association of aging with declining metabolic function and the potential for elevated BMI to increase chronic disease risks. Hypertension and heart disease show low SHAP values, implying their limited direct influence on the model prediction output. For moking history, the SHAP values are very low and vary widely across categories (e.g., current, former). Some categories (e.g., current) may slightly increase risk, while others (e.g., never) might reduce it. This indicates a minimal impact of thes features on the model prediction outcome. Finally, gender features (e.g., male, female) have SHAP values near 0, also indicating no impact or a minimal impact of gender on the model prediction outcome.

# 5 Conclusion

## 5.1 Summary

This study proposed a machine learning framework to predict diabetes. The framework consists of an ensemble model and an explainable artificial intelligence. The ensemble model was built by combining two base models, namely CatBoost Classifier and XGBoost Classifier using stacking method, with the objective of improving prediction performance (especially the accuracy), robustness, and generalisation. The main objective of using XAI is to make the prediction outcomes transparent and understandable, and SHapley Additive exPlanation (SHAP) was used for this purpose. A very large dataset consisting of 100,000 observations and 8 features was used, and the framework achieved a test accuracy of 88.93%, outperforming the base models, and demonstrates its effectiveness of handling complex medical datasets. With its ability to handle class distribution with no significant bias, the model also shows its stability potential in imbalanced data scenarios, ensuring its applicability to diverse clinical dataset. Moreover, the results from SHAP reveal that features such as HbA1c level and blood glucose are the most critical features in the model prediction outcome, and then have a substantial impact on the risk of developing diabetes. This is followed by age and BMI which also have significant impact on the model prediction outcome. Hypertension and heart disease have limited direct impact on the model prediction output.

## 5.2 Critical Evaluation

The project successfully achieved its predefined objectives of enhancing the prediction performance and robustness of diabetes, and understanding of the predictions output by healthcare professionals and patients. By integrating two machine learning base models to form an ensemble model, the proposed prediction method has demonstrated its effectiveness with complex medical datasets, a good generalisation ability, and its capability to handle imbalance datasets with no significance bias. The predictive performance of the two base models used to form the ensemble approach complemented each other, leading to prediction results that outperform all the four individual base models evaluated in the study. Additionally, with the application of XAI, the proposed model not only offers possibilities for physicians and patients to know the features associated with the risk of developing the disease, but also provides transparency and trust of AI in medical sector.

Compared to some existing related works, this study offers a balanced trade-off between prediction accuracy and computational complexity, as well as a competitive advantage in terms of model interpretability. For example, while studies [16], [21], and [22] reported higher prediction accuracy, their models lack explainability. They also lack generalisation due to the limited size and nature of their datasets used, and they introduced some bias. Additionally, the model from [21] is more computationally complex with three base models used to form the hybrid one. Even though XAI is applied to the model in [20], the model lacks generalisation with very limited dataset. The comparisons are summarised in table 3.

**Table 3 Results Comparisons with Related Works**

| Reference | Proposed Approach | Number of Models Involved | Accuracy | Model Explainability | Limitations |
|---|---|---|---|---|---|
| [16] | Ensemble Model | 3 | 95% | No | Limited generalisation; bias from imbalance dataset |
| [20] | Hybrid Model with XAI | 2 | 92.85% | Yes | Very limited dataset; no generalisation |
| [21] | Hybrid Model | 3 | 90.10% | No | No generalisation; bias from dataset; higher computational complexity |
| [22] | Ensemble Model | 2 | 90.4% | No | No generalisation; bias from dataset |
| This work | Ensemble Model with XAI | 2 | 88.93% | Yes | Couldn't be optimised; Potential stability issue due to limited generalisation of XGBoost |

## 5.3 Limitations and Future Work

Despite its achievements, this study has some limitations. The first limitation is that the base models and the ensemble models did not go under any optimisation process. The reason for this is that all attempts to improve the models performance using hyperparametres tuning either decreased some models accuracies or lead to overfitting for others in case their accuracies increase. Even the use of k-fold cross-validation during the hyperparameter search process could not help achieving the objective. Another limitation is the generalisation gaps of XGBoost base model and the ensemble model which are 0.42% and 0.41%, respectively. Even though the gaps are acceptable (less than 0.5%) and did not lead to overfitting, it could be an issue for the models stability.

In future work, advanced techniques such as Bayesian optimisation could be explored for hyperparamter tuning to avoid overfitting while improving models performance. Further fine-tuning of regularisation parameters could also be explored to maintain the models generalisation while benefiting from hyperparameter tuning. Moreover, "early stopping", which is a technique that avoids over-training by helping models to keep generalisable patterns instead of memorising the training data [34], could be applied to enhance the models stability.

## 5.4 Critical Reflection

Reflecting back on the process of doing this project, several aspect went well. A well-defined plan was established early on, helping in deciding the research direction and topic which were promptly discussed and structured with the supervisor. An extensive literature review was also done to understand the state of the art of the research direction, and this helped in detecting and evaluating potential models to be used in the project. Moreover, a strategic model selection was done where multiple base models were initially trained and evaluated but the selection was later streamlined in collaboration with my supervisor to focus on four models based on their potential contributions in forming an ensemble model to predict diabetes.

However, some aspects of the project did not go well and need improvement. For example, due to my limited code management skills, the code scripts were initially not well-organised, and this led to loss of experimental results and forced me to start afresh. Also, the application of SHAP took long time to be successful, and this is due to my lack of experience with the technique and the limited related resource available in the context of diabetes prediction to help for its quick understanding and implementation.

# References

[1] E. Cousin *et al.* Diabetes Mortality and Trends Before 25 Years of Age: An Analysis of the Global Burden of Disease Study 2019, *Lancet Diabetes Endocrinol*, 10(3):177–192, 2022. doi: 10.1016/S2213-8587(21)00349-1.

[2] A. Z. Woldaregay *et al.* Data-driven Modeling and Prediction of Blood Glucose Dynamics: Machine Learning Applications in Type 1 Diabetes, *Artif Intell Med*, 98(2019):109-134, August 2018. doi: 10.1016/j.artmed.2019.07.007.

[3] V. Jaiswal, A. Negi, and T. Pal. A Review on Current Advances in Machine Learning Based Diabetes Prediction, *Prim Care Diabetes*, 15(3):435-443, 2021. doi: 10.1016/j.pcd.2021.02.005.

[4] E. Afsaneh, A. Sharifdini, H. Ghazzaghi, and M. Z. Ghobadi. Recent Applications of Machine Learning and Deep Learning Models in the Prediction, Diagnosis, and Management of Diabetes: A Comprehensive Review, *Diabetol Metab Syndr*, 14(1):1-39, 2022. doi: 10.1186/s13098-022-00969-9.

[5] O. Virgolici et al. Diabetes Prediction Using Machine Learning Techniques: A Brief Overview Diabetes & its Complications, 8(1): 1-9, 2024. Available:https://www.scivisionpub.com/pdfs/diabetes-prediction-using-machine-learning-techniques-a-brief-overview-3189.pdf.

[6] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda. Machine Learning and Data Mining Methods in Diabetes Research, *Comput Struct Biotechnol J*, 15(2017):104–116, 2017. doi: 10.1016/j.csbj.2016.12.005.

[7] R. L. Naidu Boddeda, R. Prasad, S. S. Amiripalli, and M. S. N. V. Jitendra. Prediction of Diabetes Using Hybridization based Machine learning algorithm, *International Conference on Smart Computing and Application (ICSCA 2023)*, pp. 1–5, 2023. doi: 10.1109/ICSCA57840.2023.10087491.

[8] E. Tjoa and C. Guan. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI, *IEEE Trans Neural Netw Learn Syst*, 32(11):4793-4813, 2021. doi: 10.1109/TNNLS.2020.3027314.

[9] M. Mustafa, Diabetes prediction dataset, available: : www.kaggle.com/datasets/iammustafatz/diabetes-prediction.

[10] S. F. Ahmed *et al.* Deep Learning Modelling Techniques: Current Progress, Applications, Advantages, and Challenges, Springer Netherlands, 56(11):1-98, 2023. doi: 10.1007/s10462-023-10466-8.

[11] M. Iman, H. R. Arabnia, and K. Rasheed. A Review of Deep Transfer Learning and Recent Advancements, *Technologies (Basel)*, 11(2):1-14, 2023. doi: 10.3390/technologies11020040.

[12] S. Shridevi and S. Elias. Explainable AI Based Neck Direction Prediction and Analysis During Head Impacts, *IEEE Access*, 12(2024):31399–31408, 2024. doi: 10.1109/ACCESS.2024.3367602.

[13] A. Dutta *et al.* Early Prediction of Diabetes Using an Ensemble of Machine Learning Models, *Int J Environ Res Public Health*, 19(19):1–25, 2022. doi: 10.3390/ijerph191912378.

[14] G. R. Ashisha, X. A. Mary, H. M. Ashif, I. Karthikeyan, and J. Roshan. Early Diabetes Prediction with Optimal Feature Selection Using ML Based Prediction Framework, *4th International Conference on Signal Processing and Communication (ICSPC)*, pp. 391–395, 2023. doi: 10.1109/ICSPC57692.2023.10125956.

[15] H. El Massari, N. Gherabi, F. Qanouni, and S. Mhammedi. Diabetes Prediction Using Machine Learning with Feature Engineering and Hyperparameter Tuning, *International Journal of Advanced Computer Science and Applications*, 15(8):171–179, 2024. doi: 10.14569/IJACSA.2024.0150818

[16] Y. Zheng. Diabetes Prediction and Analysis based on Ensemble Learning Method, *3rd International Conference on Electronic Information Engineering and Computer Science, (EIECS 2023)*, pp. 1353–1358, 2023. doi: 10.1109/EIECS59936.2023.10435397.

[17] A. A. Alzubaidi, S. M. Halawani, and M. Jarrah. Towards a Stacking Ensemble Model for Predicting Diabetes Mellitus using Combination of Machine Learning Techniques, *International Journal of Advanced Computer Science and Applications*, 14(12):348–358, 2023. doi: 10.14569/IJACSA.2023.0141236

[18] K. Suga Priya, A. Saranya, S. Abinesh, and S. Chidambaram. A Performance Analysis of Predicting Diabetics Disease Using Hybrid Machine Learning Approach, *9th International Conference on Advanced Computing and Communication Systems (ICACCS 2023)*, pp. 1105–1108, 2023. doi: 10.1109/ICACCS57279.2023.10113086.

[19] R. Goudar and N. Aftab. Diabetes Prediction using Hybrid Model, *IEEE 9th International Conference for Convergence in Technology (I2CT 2024)*, pp. 1–10, 2024. doi: 10.1109/I2CT61223.2024.10544246.

[20] D. K. Vishwakarma, A. Bilal, and A. Zahoor. Hybrid XGBoost and Extreme Machine Learning Algorithm on Diabetes Disease Prediction, *14th International Conference on Computing Communication and Networking Technologies (ICCCNT 2023)*, pp. 1–7, 2023. doi: 10.1109/ICCCNT56998.2023.10307445.

[21] R. F. Albadri, S. M. Awad, A. S. Hameed, T. H. Mandeel, and R. A. Jabbar, A Diabetes Prediction Model Using Hybrid Machine Learning Algorithm, Mathematical Modelling of Engineering Problems, 11(8): 2119–2126, Aug. 2024, doi: 10.18280/mmep.110813.

[22] P. Sampath et al., Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique, Sci Rep, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-78519-8.

[23] V. J. Hodge and J. Austin, "A Survey of Outlier Detection Methodologies," Artificial Intelligence review, 22(2): 85-126, 2024, http: https://www-users.york.ac.uk/~vjh5/myPapers/Hodge+Austin_OutlierDetection_AIRE381.pdf

[24] V. J. Hodge and J. Austin, "A Survey of Outlier Detection Methodologies," Artificial Intelligence review, 22(2): 85-126, 2024, http: https://www-users.york.ac.uk/~vjh5/myPapers/Hodge+Austin_OutlierDetection_AIRE381.pdf

[25] Y. Cui, M. Jia, T. Y. Lin, Y. Song, and S. Belongie, Class-balanced loss based on effective number of samples, *in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* pp. 9260–9269, Jun. 2019. doi: 10.1109/CVPR.2019.00949.

[26] M. Abdelhamid and A. Desai, Balancing the Scales: A Comprehensive Study on Tackling Class Imbalance in Binary Classification, Sep. 2024, Available: http://arxiv.org/abs/2409.19751

[27] R. Rodríguez-Pérez and J. Bajorath, Feature importance correlation from machine learning indicates functional relationships between proteins and similar compound binding characteristics, Sci Rep, 11(1), Dec. 2021, doi: 10.1038/s41598-021-93771-y.

[28] H. Peng, F. Long, and C. Ding, Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, Aug 2005, doi: https://doi.org/10.1109/TPAMI.2005.159

[29] T. Chen and C. Guestrin, XGBoost: A scalable tree boosting system, in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, pp. 785–794, Aug. 2016,. doi: 10.1145/2939672.2939785.

[30] P. Singh, N. Singh, K. K. Singh, and A. Singh. Diagnosing of Disease Using Machine Learning, INC, 2021. doi: 10.1016/B978-0-12-821229-5.00003-3.

[31] L. Rokach, Ensemble-based classifiers, Artif Intell Rev, 33(1-2), pp. 1–39, Feb. 2010, doi: 10.1007/s10462-009-9124-7.

[32] A. Defazio Ambiata, F. Bach, and S. Lacoste-Julien, SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives, Dec. 2014, doi: 1407.0202v3

[33] J. Friedman, T. Hastie, and R. Tibshirani, Regularization Paths for Generalized Linear Models via Coordinate Descent, Journal of Statistical Software, 33(1),Jan 2010. doi: https://doi.org/10.18637/jss.v033.i01

[34] R. Caruana and S. Lawrence, Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping, Advances in Neural Information Processing Systems 13 (NIPS 2000), 2000, available: https://papers.nips.cc/paper_files/paper/2000/hash/059fdcd96baeb75112f09fa1dcc740cc-Abstract.html