Interpretable Modeling for Type 2 Diabetes Risk Factors

Ryan Soh

June 2025

Contents

Abstract

Diabetes mellitus is a chronic disease with significant health impacts worldwide. Early detection and risk prediction are essential for effective intervention and management. Recent advances in machine learning have led to the development of more accurate and interpretable models for diabetes prediction, leveraging publicly available datasets and sophisticated algorithms to improve predictions. In this project, traditional machine learning models are explored, and a select few are combined to improve the model's ability to generalise reduce the limitations of each individual model.

1 Introduction

Diabetes is a long-term illness characterised by a high sugar level in the blood[1]. A World Health Organisation study in 2022 found that 14% of adults aged 18 years and older were living with diabetes, a whopping increase from 7% in 1990[2]. Diabetes is also highly prevalent in Singapore, where one in three Singaporeans are at risk of developing diabetes, and approximately one in ten Singaporeans are diagnosed with diabetes[3]. Between Type 1 and Type 2 diabetes, Type 2 diabetes is the most common form of diabetes in the general population and often develops from pre-diabetes (where blood sugar level is higher than normal but not high enough to be diagnosed as Type 2 diabetes).

While there are no treatment plans available that will eliminate diabetes completely, managing lifestyle factors properly can help to reduce the risk of Type 2 diabetes. Some factors that increase the risk of pre-diabetes and Type 2 diabetes include[4] :

- Having close relatives with diabetes
- BMI of 23.0 kg/m^2 or higher
- Inactive lifestyle
- History of gestational diabetes (for women)
- Abnormal blood cholesterol or lipid levels
- High blood pressure
- Age 40 years and above
- Impaired glucose tolerance or fasting glucose

Diabetes can cause organ damage and higher risk of other diseases, such as coronary heart disease, heart failure, diabetic cardiomyopathy, eye disease, nerve damage and skin and mouth infections. As such, it is recommended to have periodic diabetes screening, taking blood glucose tests such as non-fasting blood test (HbA1c) that measures the average blood glucose levels over the past three months, and fasting plasma blood glucose (FPG) test, measuring blood glucose levels after 8 to 12 hours of fasting[1].

This project aims to:

1. Identify models which do well on the task of predicting whether an individual (aged 18 and above) has Type 2 diabetes using clinical and lifestyle features, classifying them as diabetic, non-diabetic or borderline diabetic
2. Identify important risk factors of Type 2 diabetes with explainable machine learning.
3. Assess whether an ensemble of machine learning models significantly improves metric scores compared to standalone models.

2 Related Work

A prominent trend in diabetes prediction research is the use of ensemble machine learning models, which combines the strengths of multiple models to enhance the predictive performance. Li et al. developed a stacking ensemble algorithm, using a genetic algorithm optimised XGBoost model with a LightGBM model as the base models, before passing the outputs from the base models to the random forest meta-classifier. The stacking model achieved an AUC value of 0.9890 and also showed superior performance in accuracy, precision, recall and F1-scores[5]. Hasan et al. introduced an automated classification pipeline, using a weighted ensemble of machine learning models optimised using a grid search hyperparameter approach for early prediction of diabetes. Different combinations of models were tested to observe that the AdaBoost and XGBoost model produced the best results amongst all other combination choices[6].

Li Et al. also highlighted the importance of addressing class imbalance, as the proportion of individuals with diabetes in the dataset was relatively low. A few sampling methods were tested, and SMOTEENN proved effective in improving model performance and ensuring that minority classes were adequately represented during training[5].

Recent research also focused on classifying Type 2 diabetes into their subtypes to enable a more personalised risk management plan. Hayato et al. introduced a random forest multiclass classification model to predict Ahlqvist's Type 2 diabetes subtypes using 15 clinical variables, achieving high accuracy of 0.94, an overall AUC value of 0.941 and high precision, recall and F1-scores above 0.9. Notably, the model maintained predictive performance despite missing insulin-related variables, proving its applicability to external datasets[7].

3 Data Pre-processing

The data used in this project has been sourced from the Centers for Disease Control and Prevention. The National Health and Nutrition Examination Survey August 2021-August 2023 Cycle[8] has a vast amount of data that is suitable for analysis and model training.

The relevant datasets were merged and cleaned, before exploratory data analysis and further pre-processing steps were performed.

3.1 Dataset

As clinical and survey data collected were segmented into more granular categories, a few datasets of interest are downloaded for analysis and model training.

Demographic Data

- DEMO_L.xpt : Demographic Variables and Sample Weights

Examination Data

- BMX_L.xpt : Body Measures

Laboratory Data

- HDL_L.xpt : Cholesterol - High-Density Lipoprotein
- GHB_L.xpt : Glycohemoglobin
- GLU_L.xpt : Fasting Plasma Blood Glucose

Questionnaire Data

- BPQ_L.xpt : Blood Pressure and Cholesterol
- DIQ_L.xpt : Diabetes
- SMQ_L.xpt : Smoking - Cigarette Use

Each of these datasets have their set of data, not all of which are relevant to the study. As such, the merged dataset only consists of the following data for analysis and model training:

- Body Mass Index (kg/m^2)
- History of high blood pressure (0=No, 1=Yes)
- Gender (0=Female, 1=Male)
- Age at screening (years)
- Diabetic status (0=Yes, 1=No, 2=Borderline)

- Glycohemoglobin level (%)
- Fasting Plasma Blood Glucose (mmol/L)
- Direct HDL-Cholesterol (mmol/L)
- Smoked at least 100 cigarettes in lifetime (1=Yes, 2=No)
- Current smoking frequency (1=Every day, 2=Some days, 3=Not at all)

Additional details on how the data was collected and response types are available in the documentation provided by the CDC NHANES website.

While there are other variables such as alcohol consumption habits and physical activity level that can influence the prediction of Type 2 diabetes, the health survey response is highly subjective (e.g. different interpretation of high-intensity activity from person to person), and also provide a range of values, which leads to approximation of results at best. Hence, these data have been explicitly excluded and not used for analysis or model training. This will be further discussed in the later section "Lessons Learnt".

In order to make EDA reliable, some data cleaning was done prior to EDA. Specifically, rows with missing values or values encoded as "Refused [to answer]" or "Don't know" were removed from the dataset. Participants whose age is below 18 are also excluded from the dataset.

3.2 Exploratory Data Analysis

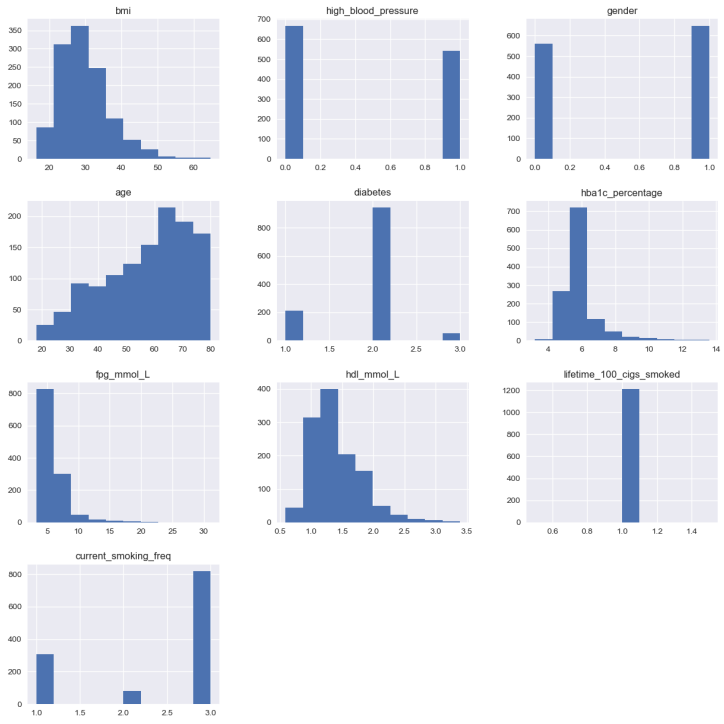

Figure 1: Histogram plot of all features

Features representing glycohemoglobin levels, fasting plasma blood glucose level and direct high-density lipoprotein are highly skewed to the right, suggesting the need for transformation later in pre-processing. While BMI and age are also skewed features, BMI is designed as a standardised way to categorise individuals based on weight relative to height and age maintains high interpretability in its original form.

It is also worth noting that all participants within this dataset after merging and cleaning have smoked at least 100 cigarettes in their lifetime, suggesting that this will not provide any useful information for the models to train on and should be removed during pre-processing.

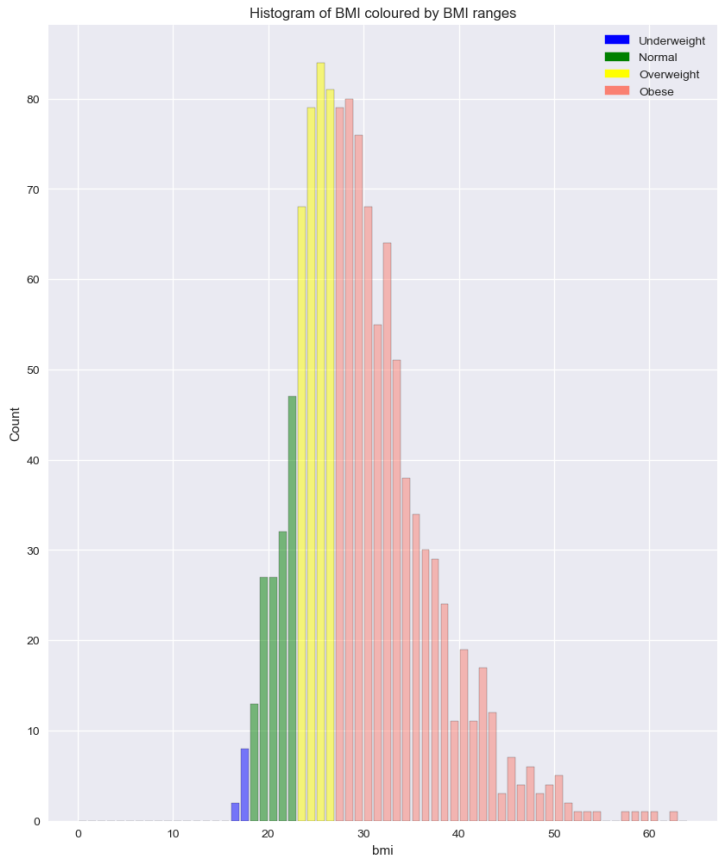

Figure 2: Histogram of BMI coloured by BMI range

Most of the participants do not belong in the "Normal" BMI range, with a significant proportion of them being at least "Overweight".

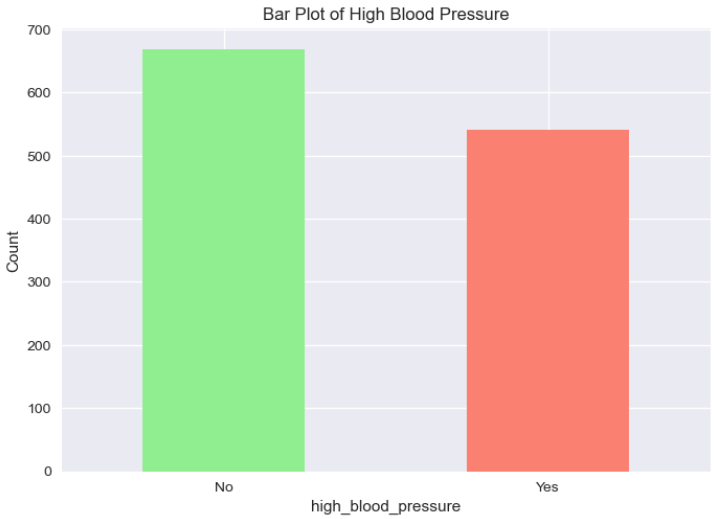

Figure 3: Bar plot of high blood pressure

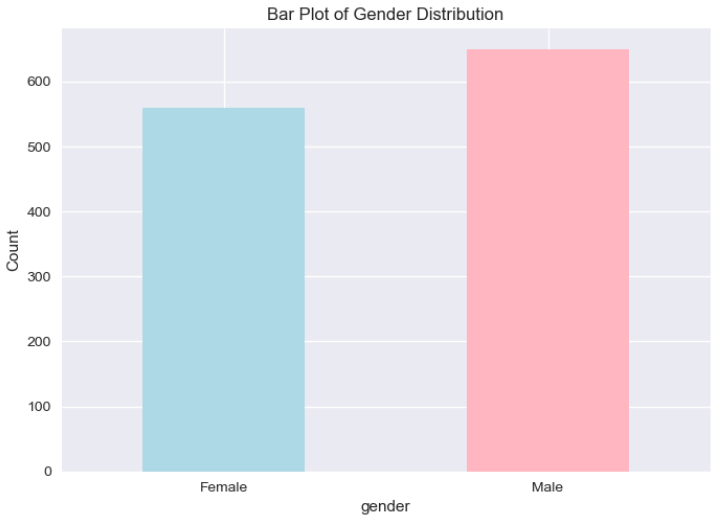

Figure 4: Bar plot of gender distribution

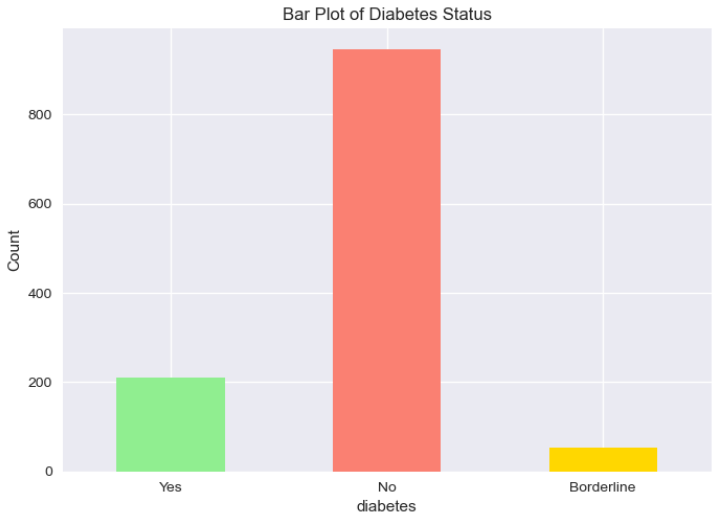

Figure 5: Bar plot of diabetes status

This bar plot shows that the diabetes status data is highly disproportionate, where there are at least four times more non-diabetic participants than diabetic participants, and at least ten times more non-diabetic participants than borderline diabetic participants. As this will be the target variable, it is important to handle class imbalance during model training.

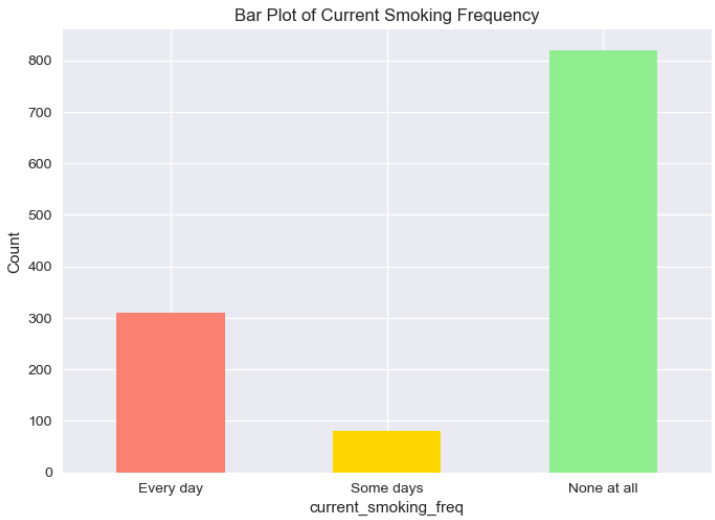

Figure 6: Bar plot of current smoking frequency

While all participants have smoked at least 100 cigarettes in their lifetime, more than half of them have quit smoking, with slightly more than 300 of them still smoking every day and less than 100 of them smoking on some days.

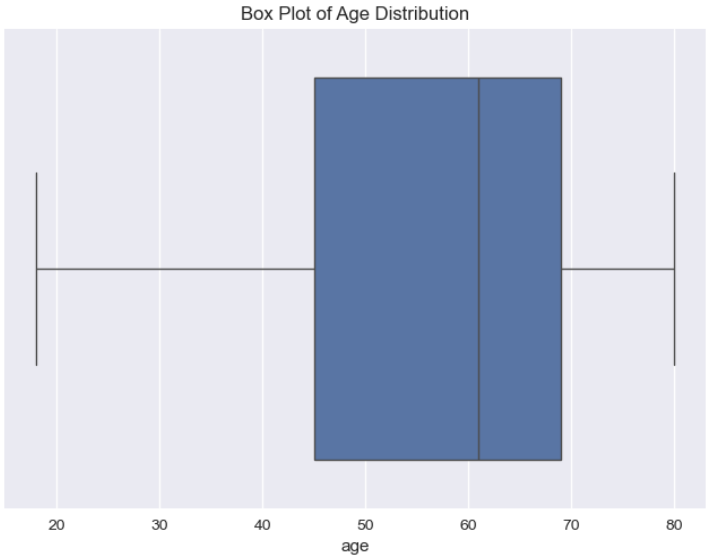

Figure 7: Box plot of age distribution

The box plot suggests a left-skewed distribution, suggesting that most participants are of the higher age ranges, with the interquartile range being between the age of 45 and 69.

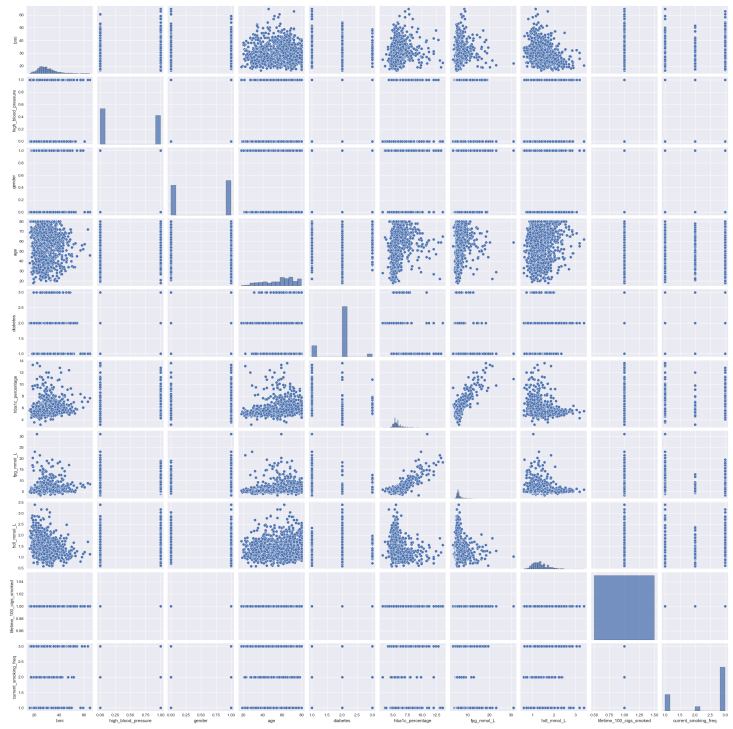

Figure 8: Pairplot of features

This pairplot shows a correlation between glycohemoglobin and fasting plasma blood glucose levels, both moving in the same direction as the value increases. Most of the other pairs of variables do not show any significant relationship.

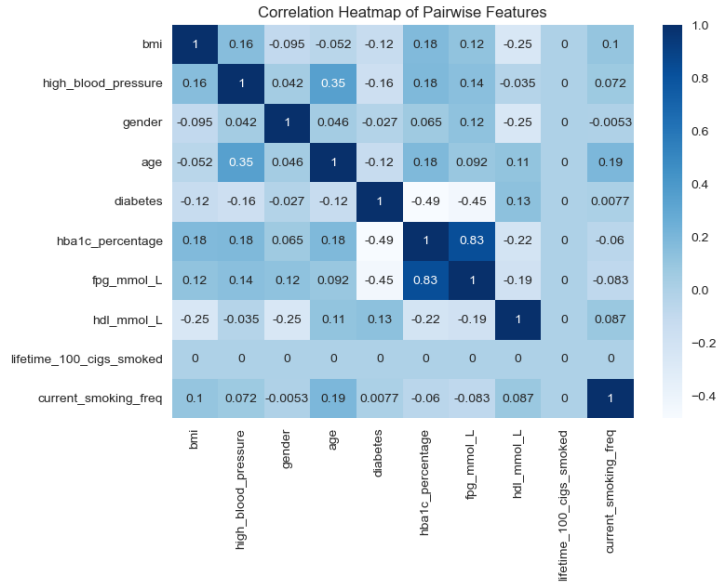

Figure 9: Correlation heatmap of pairwise features

Similar to the plot in Figure 8, the correlation heatmap shows a numerical representation of the pairwise relationship of the features. As identified earlier, glycohemoglobin and fasting plasma blood glucose levels have a stronger relationship in the positive direction as indicated by the darker shade of blue.

3.3 Pre-processing

Log transformation was performed on the features representing glycohemoglobin levels, fasting plasma blood glucose level and direct high-density lipoprotein identified as right-skewed during exploratory data analysis.

Outliers from non-categorical columns and features that provide no predictive information identified during EDA were removed from the dataset.

Categorical features were converted to integer values, and the target variable "diabetes" was label encoded for model training.

4 Model Construction

For this multiclass classification task, the following classification models were explored:

- Decision Tree
- Random Forest
- XGBoost
- AdaBoost
- Logistic Regression
- K-Nearest Neighbours
- Support Vector Machine
- Neural Network

GridSearchCV is employed before building each model to identify optimal hyperparameter values from a list of values provided. This allows for a more accurate comparison of each model's performance, as opposed to manually testing hyperparameters individually with different values, then choosing the hyperparameter values which return the highest average cross validation scores. While it may be plausible to build a strong model using those values, it is also prone to missing out on the strongest combination of values, therefore building sub-optimal models.

To demonstrate what manually evaluating hyperparameter values using metric scores look like, two examples are demonstrated in building the decision tree and random forest models. However, GridSearchCV was performed later and the optimised search results were used to build the final decision tree and random forest models.

4.1 Class Imbalance

The class imbalance in the dataset presents the issue where the models will mostly see "no diabetes" examples, learn those patterns well and subsequently, predict correctly examples that are "no diabetes" in the test set. However, the models will struggle to learn the patterns representing "diabetic" and "borderline diabetic" classes. This will result in a higher number of misclassified examples of the minority classes, and hence lower metric scores.

As such, oversampling of the minority classes, undersampling of the majority class or a hybrid approach that employs both can be considered to handle the class imbalance issue. This will give the model more opportunity to learn the patterns of the minority classes and possibly lead to better metric scores. However, these strategies do come with their own caveats.

In this project, four strategies - RandomOverSampler, SMOTE, ADASYN and SMOTEENN are experimented. The metric results post-sampling are compared against the base model with no manipulation to the dataset. The base model is built using an XGBoost classifier, a gradient boosting model with built-in regularisation that prevents overfitting, and is computationally efficient, optimising the speed and performance of the algorithm.

To find out which class balancing strategy is the most effective, the data processed from each sampling method serves as input for the XGBoost model before GridSearchCV is performed. The hyperparameters that produced the best macro F1-score during cross validation is used to build the optimised model and make predictions on the test set.

RandomOverSampler randomly duplicates existing samples from minority classes until classes are balanced, but this may increase the risk of overfitting. SMOTE generates synthetic samples by interpolating linearly between existing minority samples, but this may introduce samples that cross class boundaries. ADASYN also generates synthetic samples similar to SMOTE, but focuses on regions where there is low neighbourhood support for the minority classes. However, this may add noise in overlapping regions. SMOTEENN employs the same oversampling strategy as SMOTE, and also undersamples the majority class by removing majority class instances from regions with high neighbourhood support from other classes. Hence, SMOTEENN is often referred to as a more aggressive approach.

Strategy	Accuracy	Macro Precision	Macro Recall	Macro F1	ROC AUC
NoResampling	**0.8678**	0.5109	0.5322	0.5214	**0.8033**
RandomOverSampler	0.8326	0.4971	0.5269	0.5108	0.7965
SMOTE	0.8326	**0.5596**	**0.6079**	**0.5794**	0.7762
ADASYN	0.8150	0.5186	0.5513	0.5321	0.7679
SMOTEENN	0.6476	0.5063	0.6226	0.5009	0.7860

Table 1: Metric score comparison of the class balancing strategies

Among the four strategies tested, SMOTE shows the greatest improvements in macro precision, macro recall and macro F1-score, while achieving a slightly lower accuracy and ROC AUC score than RandomOverSampler. The dip in accuracy compared to NoResampling is expected as SMOTE balances out the classes such that the majority class no longer dominates. The model will start to predict the previously underrepresented classes more, which leads to higher recall and F1-score, but can also introduce more false positives in the minority class, which explains the reduction in overall accuracy and ROC AUC scores.

SMOTEENN showed the largest improvement in macro recall, but all other metric scores were lower, especially in overall accuracy. Its aggressive approach may have removed too many borderline samples which are informative but harder-to-classify, making the dataset overly clean and hence, the model loses its ability to generalise well.

SMOTE is a good resampling strategy that will be used to train and evaluate the models.

4.2 Model Training and Evaluation

To ensure consistency in training and evaluating the different models, the same train-test split data was used across all models. To ensure the minority class is present within the test set, the train test split was stratified. SMOTE was then applied onto the training set to balance out the classes for learning. StratifiedKFold was used to evaluate the model before testing[9]. Stratification ensures that the original distribution of classes is maintained in the training and test set and across each fold, which allows the model to learn to classify the minority classes as well, resulting in better generalisation capabilities.

5 Results Interpretation, Comparison, and Recommendation for Fusion Model

To understand how each model makes predictions on new data, permutation feature importance and SHapley Additive exPlanations (SHAP) are useful in explaining how each feature affects the models' predictions.

5.1 Permutation Feature Importance

Permutation feature importance is a global model-agnostic post-hoc method, measuring the increase in the prediction error of the model after permuting the values of a feature[10]. A larger increase in prediction error indicates higher feature importance.

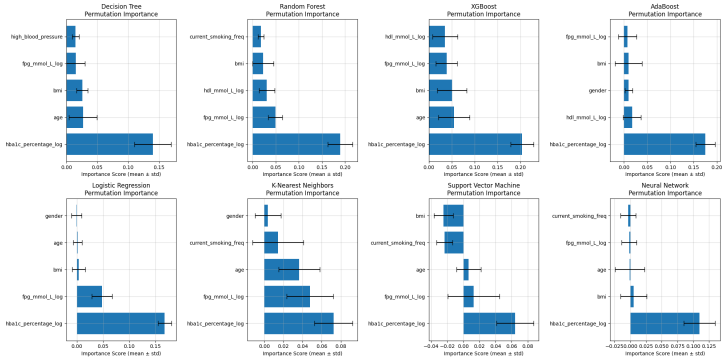

Figure 10: Permutation feature importance for each trained model

As observed in Figure 10, glycohemoglobin level is consistently the most important feature across all models, with importance scores ranging from 0.0645 to 0.2038.

Feature	DT	RF	XGB	ADB	LGR	KNN	SVM	NN	Avg
hba1c_log	0.1399	0.1896	0.2038	0.1754	0.1677	0.0724	0.0645	0.1100	0.1404
fpg_log	0.0151	0.0490	0.0389	0.0078	0.0481	0.0479	0.0130	-0.0018	0.0272
age	0.0269	0.0124	0.0548	0.0071	0.0017	0.0366	0.0067	-0.0008	0.0182
bmi	0.0256	0.0224	0.0506	0.0100	0.0037	-0.0031	-0.0246	0.0054	0.0113
hdl_log	-0.0001	0.0305	0.0354	0.0174	-0.0079	-0.0104	-0.0334	-0.0064	0.0031
smoking_freq	0.0005	0.0180	0.0103	-0.0031	-0.0086	0.0145	-0.0232	-0.0033	0.0006
gender	0.0128	-0.0072	0.0025	0.0102	-0.0003	0.0039	-0.0263	-0.0180	-0.0028
high_bp	0.0150	0.0151	0.0081	-0.0132	0.0092	0.0037	-0.0338	-0.0215	0.0045

Table 2: Permutation importance scores for each model

Averaging out across all models with equal weighting, fasting plasma blood glucose level and age are the next most important features with an average importance score of 0.0272 and 0.0182 respectively . While BMI, direct high-density lipoprotein level and current smoking frequency have positive permutation feature importance scores, their smaller magnitudes suggest that they do not contribute much to the models' abilities to make predictions.

Gender and history of high blood pressure attained a negative importance score on average. This suggests that these features may be irrelevant and could even be detrimental to the model's predictive performance, possibly introducing noise.

However, feature importance does not inform how the feature influences the prediction as it is measured by prediction error. Feature importance is just an internal ranking of features.

5.2 SHAP Analysis

SHapley Additive exPlanations (SHAP) is a local model-agnostic post-hoc method, explaining the prediction of a model by quantifying the contribution of each feature to the prediction[10]. Unlike permutation feature importance, the SHAP summary plot combines feature importance with feature effects. The summary plot orders the features by importance in the y-axis, and shows how that feature increases or decreases the prediction relative to the base value, given a high or low feature value.

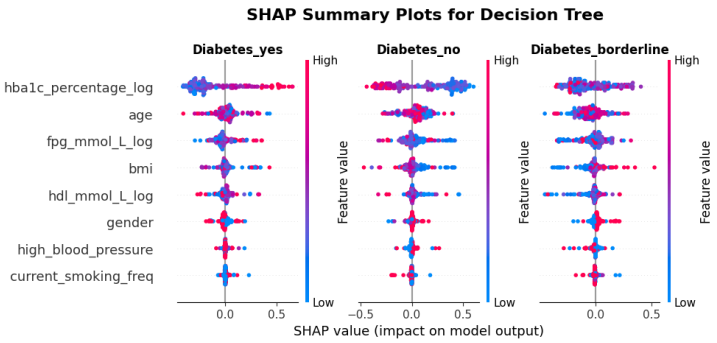

Figure 11: SHAP beeswarm plot for decision tree

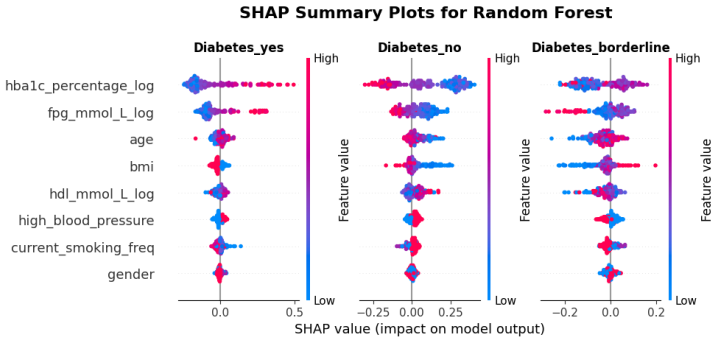

Figure 12: SHAP beeswarm plot for random forest

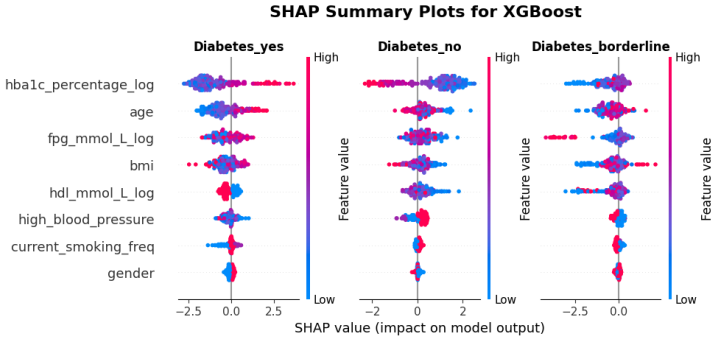

Figure 13: SHAP beeswarm plot for XGBoost

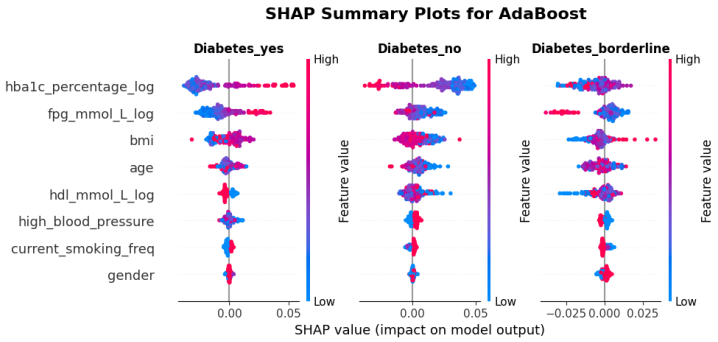

Figure 14: SHAP beeswarm plot for AdaBoost

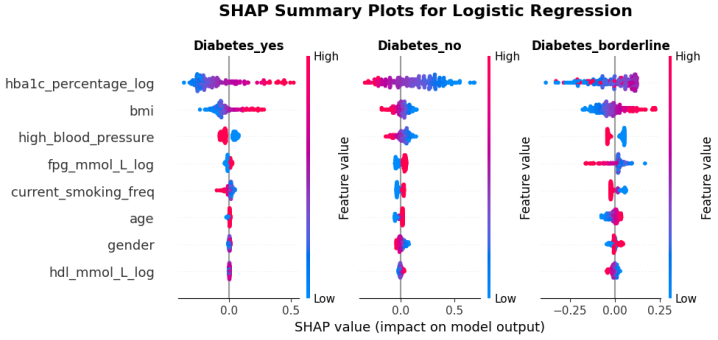

Figure 15: SHAP beeswarm plot for logistic regression

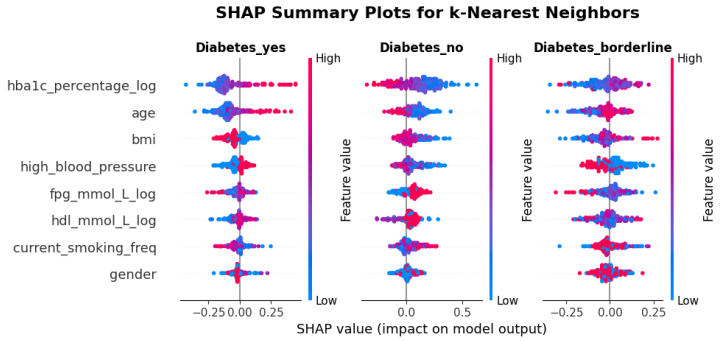

Figure 16: SHAP beeswarm plot for k-nearest neighbours

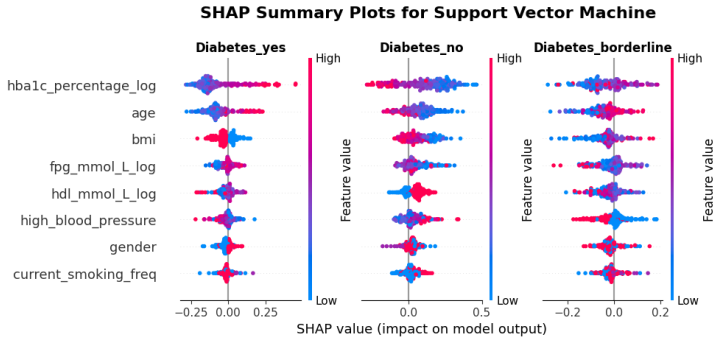

Figure 17: SHAP beeswarm plot for support vector machine

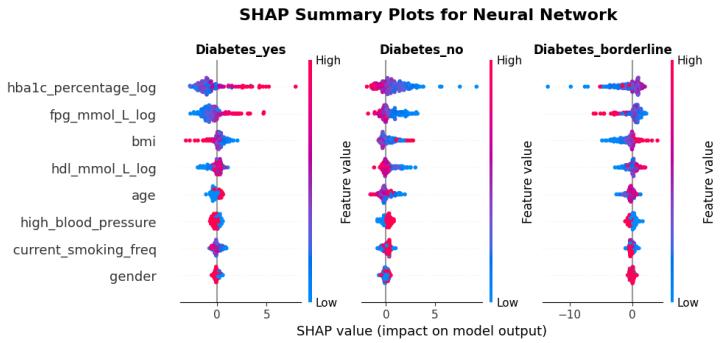

Figure 18: SHAP beeswarm plot for neural network

Using the beeswarm plot, the feature effects are displayed, demonstrating how the magnitude of a feature value pushes the model's prediction towards or away from a certain class. In general, high glycohemoglobin levels strongly push predictions towards diabetic while low glycohemoglobin levels push predictions towards non-diabetic. In most cases, moderate glycohemoglobin levels push predictions towards borderline diabetic.

Most of the diabetic and non-diabetic plots are approximately mirror images of one another in terms of the distribution of high and low feature values and the corresponding SHAP value direction, while borderline diabetic plots do not show any clear trend with the other two classes. In fact, some features show opposing trends between diabetic and

borderline diabetic classes, which questions the intuition that features pushing predictions to diabetic should also push the prediction to borderline diabetic. For instance, in the random forest plot, higher values of fasting plasma blood glucose level pushes the predictions to diabetes_yes, while moderate values of fasting plasma blood glucose levels push predictions to borderline diabetes.

From model to model, the distribution of the SHAP plots remains generally the same - a high glycohemoglobin level pushes predictions to diabetic in every model. There are little contradictions observed.

5.3 ROC AUC and Classification Report Scores

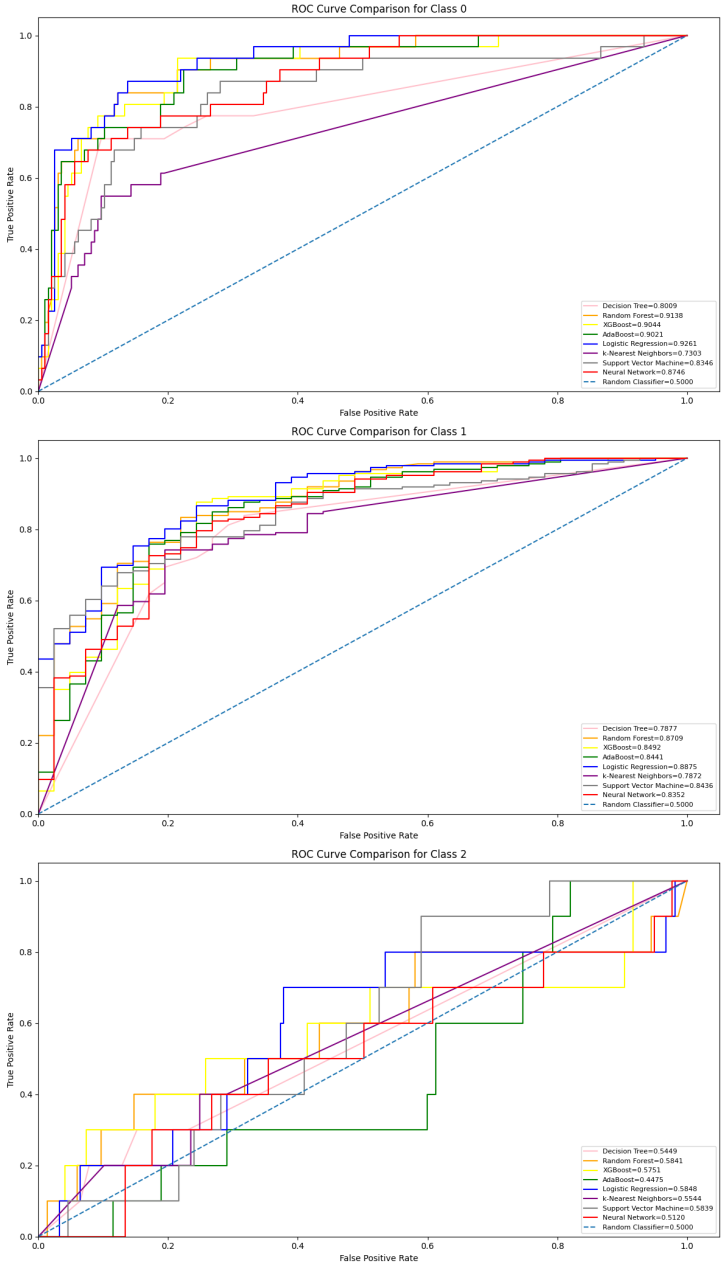

Figure 19: ROC curve comparison for all models and classes

Model	Macro Precision	Macro Recall	Macro F1	ROC AUC
Logistic Regression	0.5569	0.5642	0.5227	**0.7995**
Random Forest	0.5468	**0.6305**	0.5700	0.7896
XGBoost	**0.5596**	0.6079	**0.5794**	0.7762
Support Vector Machine	0.4566	0.4688	0.4546	0.7540
Neural Network	0.4891	0.5172	0.4904	0.7406
AdaBoost	0.5288	0.5692	0.5450	0.7312
Decision Tree	0.4732	0.5534	0.4671	0.7112
k-Nearest Neighbours	0.4816	0.4968	0.4681	0.6906

Table 3: Macro score and ROC AUC comparison between models

From Table 3, XGBoost achieved the best F1-score of 0.5794 with a well-balanced precision and recall score. Random forest has the second highest F1-score of 0.5700 and the highest macro recall score of 0.6305. Logistic regression has the best ROC AUC score of 0.7995, but a weaker F1-score.

A simple model like logistic regression did surprisingly well, suggesting the dataset may not contain complex patterns that require more advanced models to capture. The decision boundary that separates the classes might be approximately linear or simple enough to be captured by a linear model, making them easily distinguishable.

Despite neural networks being a strong model that can capture complex, non-linear relationships, it severely underperformed compared to simpler ones like logistic regression and AdaBoost. This may suggest that the dataset is not too complex or that the hyperparameters chosen were sub-optimal, and other values should be tested to improve the results.

5.4 Fusion Model

Instead of using individual models to make predictions for the Type 2 diabetes class, a stack ensemble can be created to combine the advantages of each model and reduce the limitations. A stack ensemble combines the predictions of a few base models, then passes the predictions through a meta-model, which learns to make the final predictions[11]. It is found that stacking has potential to improve predictive performance, capturing patterns that individual models may not be able to capture alone. Model diversity in the base models is also important to make the stacked ensemble reduce overfitting and generalise well[12].

Based on the macro scores and ROC AUC of the different models, the base models can include XGBoost, random forest and logistic regression given the high macro F1-score, macro recall score and ROC AUC respectively. Logistic regression is a simple meta-model suitable for classification tasks[13]. The goal of this stacking ensemble is to identify if it does better than base models do. If base models perform better or equally well compared to the stacking ensemble, the stacking ensemble should not be used to ensure low complexity.

5.5 Overall Results Comparison

Depending on the prediction results for Type 2 diabetes, follow-up plans are proposed accordingly. As such, it is important to distinguish when precision or recall is a more important metric based on the classification.

For prediction classes diabetic and borderline diabetic, a high recall score is important to ensure fewer false negatives, and timely intervention and early treatment for Type 2 diabetes.

Model	Macro Precision	Macro Recall	Macro F1	ROC AUC
Logistic Regression	0.5569	0.5642	0.5227	**0.7995**
Random Forest	0.5468	**0.6305**	0.5700	0.7896
XGBoost	**0.5596**	0.6079	**0.5794**	0.7762
Support Vector Machine	0.4566	0.4688	0.4546	0.7540
Neural Network	0.4891	0.5172	0.4904	0.7406
Stacked Ensemble	0.5583	0.6061	0.5780	0.7405
AdaBoost	0.5288	0.5692	0.5450	0.7312
Decision Tree	0.4732	0.5534	0.4671	0.7112
k-Nearest Neighbours	0.4816	0.4968	0.4681	0.6906

Table 4: Macro score and ROC AUC comparison with Stacked Ensemble

Model	Recall
Random Forest	0.7742
AdaBoost	0.7419
XGBoost	0.7419
Stacked Ensemble	0.7419
Decision Tree	0.7097
Logistic Regression	0.7097
Neural Network	0.6774
k-Nearest Neighbours	0.6774
Support Vector Machine	0.5161

Table 5: Recall score comparison for class "Diabetes_yes"

For the identification of diabetic cases, random forest performed the best with a recall score of 0.7742, while AdaBoost, XGBoost and the stacked ensemble performed slightly worse with a similar score of 0.7419. The neural network, k-nearest neighbours and support vector machine models struggled to classify the positive class correctly, with the lowest recall score being 0.5161.

Model	Precision	F1-score
XGBoost	0.0.9371	0.9086
Stacked Ensemble	0.9314	0.9030
AdaBoost	0.9253	0.8944
Random Forest	0.9441	0.8761
Support Vector Machine	0.9130	0.8473
Neural Network	0.9351	0.8471
k-Nearest Neighbours	0.9388	0.8288
Logistic Regression	0.9549	0.7962
Decision Tree	0.9380	0.7683

Table 6: Precision and F1 score comparison for class "Diabetes_no"

A high precision score for the non-diabetic class is ideal to reduce the number of misclassified cases of diabetes so less patients are put through unnecessary treatment and worry. However, it is easy to achieve a high precision score due to the class imbalance in this dataset, with the majority of them being non-diabetic. Hence, F1 score, being the harmonic mean of precision and recall, is a good metric that shows if the models are achieving high precision at the expense of recall.

Although the logistic regression model has achieved the highest precision score, it scores one of the worst in F1-score relative to the other models. The same pattern can be observed for k-nearest neighbours and decision tree models. XGBoost is strong in identifying true negatives and avoiding false positives, achieving the highest F1-score of 0.9086 with a reasonably high precision score of 0.9371. The stacked ensemble achieved very similar results with an F1-score of 0.9086 and recall score of 0.9371.

Model	Recall
Decision Tree	0.3
Random Forest	0.3
Logistic Regression	0.3
XGBoost	0.2
k-Nearest Neighbours	0.2
Stacked Ensemble	0.2
AdaBoost	0.1
Support Vector Machine	0.1
Neural Network	0.1

Table 7: Recall score comparison for class "Diabetes_borderline"

The low recall scores for the borderline diabetic class across all models suggest difficulty in identifying the rarest class in this imbalance dataset. The decision tree, random forest and logistic regression models performed the best amongst all models with a recall score of 0.3, while the XGBoost, k-nearest neighbours and stacked ensemble models achieved a recall score of 0.2. Even though SMOTE was applied to train the models to identify the minority classes, the models still struggle to distinguish these examples from others. This could imply that the borderline classes may not be well-defined given the

features used, and hence the synthetic samples interpolated lack discriminative features and could not help improve the model's ability to generalise much.

- For diabetic and borderline diabetic classes, high recall is important to minimize false negatives.
- Random forest had the best recall for diabetic cases (0.7742).
- XGBoost and the stacked ensemble had the highest F1-score for non-diabetic cases (0.9086).
- All models struggled to identify borderline diabetic cases, even after SMOTE.

6 Further Work

There are a few limitations faced while building this project. Starting from the initial brainstorming of this project, the goal was to build a predictive model that can be further explored by the Singapore healthcare system, to improve efficiency in the healthcare sector while still maintaining accurate medical diagnosis of Type 2 diabetes. However, special request is needed in order to access Singapore healthcare data, and thus the secondary option was to find a dataset within the Southeast Asian region, where lifestyle factors may be less dissimilar. There was a lack of accessible datasets that were highly informative, thus the final decision was made to use the NHANES dataset from CDC, even though there was a lack of cases where diabetes and borderline diabetes were present. Nonetheless, the goal remains unchanged, to be able to build a predictive model that can generalise well on any test dataset and to further improve metric scores while capturing complex patterns.

While GridSearchCV was employed to find the optimal hyperparameter values to build each model, there are many other values that can be tested with more time and computation power. It is hence not guaranteed that the hyperparameters used to train the models here are the most optimal.

More features can also be explored to build a more robust model. For instance, Hayato et al. utilised different features, such as HOMA-2B and HOMA-2IR, were used to build predictive models and even classified Type 2 diabetes based on their subtypes [7]. They also achieved higher accuracy and AUC scores. Additionally, it is notable that the NHANES dataset managed to capture useful data such as alcohol consumption, duration of physical activity and intensity of physical activity which are correlated with the risk of developing Type 2 diabetes, these features were intentionally excluded from model training due to its highly subjective nature or lack of specificity in the data collected. For instance, excessive drinking of alcohol will increase the risk of developing diabetes, but it is also dependent on the amount of alcohol and level of alcohol content consumed. Within the NHANES dataset, participants were only asked on their frequency of alcohol consumption, while the amount of alcohol and level of alcohol content consumed were not identified as it is hard to quantify. Similarly, while physical activity can help with the prevention of Type 2 diabetes, it is also highly dependent on the amount of exercise and

intensity of exercise. As the data ranges by the intensity of exercise, duration of exercise and frequency of exercise, feature engineering of physical activity features would cause overly complex features to be introduced into the model.

All models did poorly in correctly identifying borderline diabetic cases. Other data rebalancing techniques and feature engineering should be explored to discover if the models' performances will improve. The models should also be evaluated against other datasets to observe the consistency in its predictive performance, but it should be noted that publicly available datasets may not fully capture the heterogeneity of real-world populations.

Other models and hyperparameter search optimisers can be further explored to evaluate their performance in detecting Type 2 diabetes. Different combinations of models and meta-classifiers should also be tested for the stacking ensemble. StackingCVClassifier is a great tool to integrate neural networks into the stacking ensemble and the performance of each combination can be observed. It is still important to ensure diversity in the base models rather than choosing the combination that offers the highest validation score.

7 Lessons Learnt

Initially, there were considerations to engineer some features such as combining BMI and age to reduce the number of dimensions. However, after more research was done, while higher BMI and age leads to increased risk of Type 2 diabetes, younger participants are less prone to Type 2 diabetes[14]. Combining the features will thus cause us to lose interpretability, not knowing if the prediction was influenced more by a high BMI or older age.

While SMOTE managed to handle the class imbalance problem and improved model performance, the overall generalisation ability did not improve significantly. This suggests that the synthetic samples added may not be representative of the true minority classes, possibly due to overlaps between classes in the feature space, which makes it difficult to separate the minority class from the rest.

To understand the importance of each feature, permutation-based feature importance was chosen over impurity-based feature importance. Impurity-based feature importance works well on tree-based models, measuring the importance of features based on how much the feature helps to reduce impurity. However, it is biased towards high cardinality features and computed on training set, hence it does not reflect its usefulness in helping the model generalise on unseen data[15]. On the other hand, permutation importance is model-agnostic, which works consistently across all model types. When permuting a feature, it also eliminates the interaction effects between features, giving a more accurate representation of the feature's effects on the prediction. It also offers the flexibility of permuting on the test set, which allows for an unbiased evaluation of how each feature contributes to the model's ability to generalise on unseen data[16].

Model results can be saved using pickle as it preserves the dictionary structure of classification reports, handles mixed data types, and quickly saves and loads. Trained models can be saved using joblib which is optimised for scikit-learn models, and it has

built-in compression that reduces file size further, making it more memory efficient.

8 Conclusion

Compared to the stacking ensemble, traditional machine learning models showed poor performance in diabetes risk prediction. The stacking ensemble showed promising results in macro scores compared to many of the individual models, including significant improvements in identifying diabetic and non-diabetic cases, and performed as well as the XGBoost model in identifying borderline diabetic cases. However, it still fell short as compared to one of its base models, XGBoost. XGBoost achieved slightly higher macro scores and performed better in the multiclass classification problem. Given the complexity of the models, XGBoost should be chosen over the stacking ensemble to achieve lower computational complexity.

There is still room for improvement in improving the effectiveness of the models in predicting the risk of diabetes. Such improvements include hyperparameter search optimisation, feature selection, feature engineering, building the stacking ensemble on different base models and meta-classifiers, and obtaining datasets with more data representative of the minority classes.

Through these efforts, there is a good possibility in constructing a more accurate model that is able to provide stronger clinical support for our healthcare professionals in their continuous work of diabetes detection and management.

References

[1] Singapore Heart Foundation. *Diabetes Risk Factors*. Accessed: 2025-07-06. 2022. URL: https://www.myheart.org.sg/health/risk-factors/diabetes/.

[2] World Health Organization. *Diabetes*. Accessed: 2025-07-06. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/diabetes.

[3] HealthHub Singapore. *What is Diabetes?* Accessed: 2025-07-06. URL: https://www.healthhub.sg/programmes/diabetes-hub/what-is-diabetes.

[4] HealthHub Singapore. *Diabetes – Are You at Risk?* Accessed: 2025-07-06. 2022. URL: https://www.healthhub.sg/live-healthy/diabetes-are-you-at-risk.

[5] W. Li, Y. Peng, and K. Peng. "Diabetes prediction model based on GA-XGBoost and stacking ensemble algorithm". In: *PLoS ONE* 19.9 (2024), e0311222. DOI: 10.1371/journal.pone.0311222. URL: https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0311222.

[6] Md. Kamrul Hasan et al. "Automated Classification Pipeline Using Weighted Ensemble for Early Prediction of Diabetes". In: *2020 IEEE 6th World Forum on Internet of Things (WF-IoT)*. 2020, pp. 1–6. DOI: 10.1109/WF-IoT48112.2020.9221212. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9076634.

[7] N. Hayato et al. "Random forest multiclass classification model to predict Ahlqvist's Type 2 diabetes subtypes using 15 clinical variables". In: *Diabetologia* (2024). DOI: 10.1007/s00125-024-06248-8. URL: https://link.springer.com/article/10.1007/s00125-024-06248-8.

[8] Centers for Disease Control and Prevention. *National Health and Nutrition Examination Survey, 2021-2023*. Accessed: 2025-07-06. 2023. URL: https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2021-2023.

[9] GeeksforGeeks. *Stratified K Fold Cross Validation*. Last updated: April 15, 2025. Accessed: 2025-07-06. 2025. URL: https://www.geeksforgeeks.org/machine-learning/stratified-k-fold-cross-validation/.

[10] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 3rd ed. Creative Commons, 2025. ISBN: 978-3-911578-03-5. URL: https://christophm.github.io/interpretable-ml-book.

[11] Brijesh Soni. *Stacking to Improve Model Performance: A Comprehensive Guide on Ensemble Learning in Python*. https://medium.com/@brijesh_soni/stacking-to-improve-model-performance-a-comprehensive-guide-on-ensemble-learning-in-python-9ed53c93ce28. Accessed: 2025-07-06. 2023.

[12] H2O.ai. *What is Stack Ensemble?* https://h2o.ai/wiki/stack-ensemble/. Accessed: 2025-07-06. 2025.

[13] Jason Brownlee. *Stacking Ensemble Machine Learning With Python*. https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/. Accessed: 2025-07-06. 2021.

[14] Ying Chen et al. "Association of body mass index and age with incident diabetes in Chinese adults: a population-based cohort study". In: *BMJ Open* 8.9 (2018). PMCID: PMC6169758. Accessed: 2025-07-06, e021768. DOI: 10.1136/bmjopen-2018-021768. URL: https://bmjopen.bmj.com/content/8/9/e021768.

[15] scikit-learn developers. *Permutation Importance vs Random Forest Feature Importance (MDI)*. https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance.html. Accessed: 2025-07-06. 2024.

[16] scikit-learn developers. *5.2. Permutation feature importance*. https://scikit-learn.org/stable/modules/permutation_importance.html. Accessed: 2025-07-06. 2024.