Drug Response Prediction system Report - Shaheer Hussain

Introduction and Problem Statement

As a leading pharmaceutical company, PharmaTech is trying to improve diabetes treatment with the help of machine learning. The company aims to create a predictive model by analysing data from various clinical trials of a new medication. This model will support the development of personalised treatment plans leading to better patient outcomes.

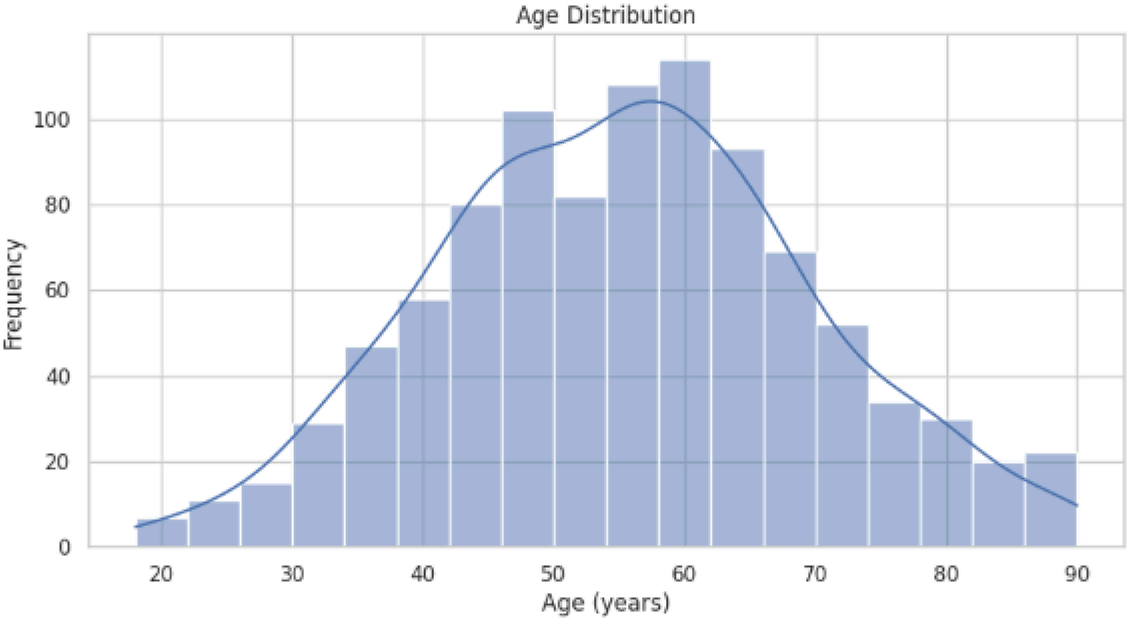
Current diabetes treatments produce different patient outcomes because of body variations. PharmaTech is trying to develop a machine learning model that can help predict the patient response to a new medication that it is developing. This will enable personalised treatment and therefore enhance the effectiveness of the treatment and care of the patient.

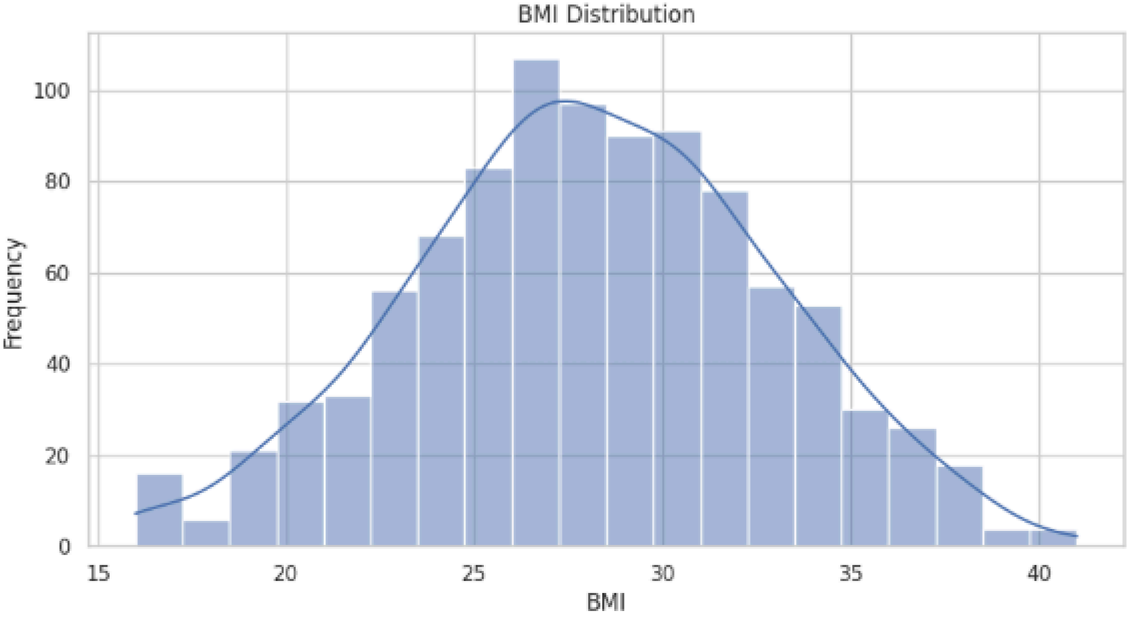
Data Analysis and Methodology

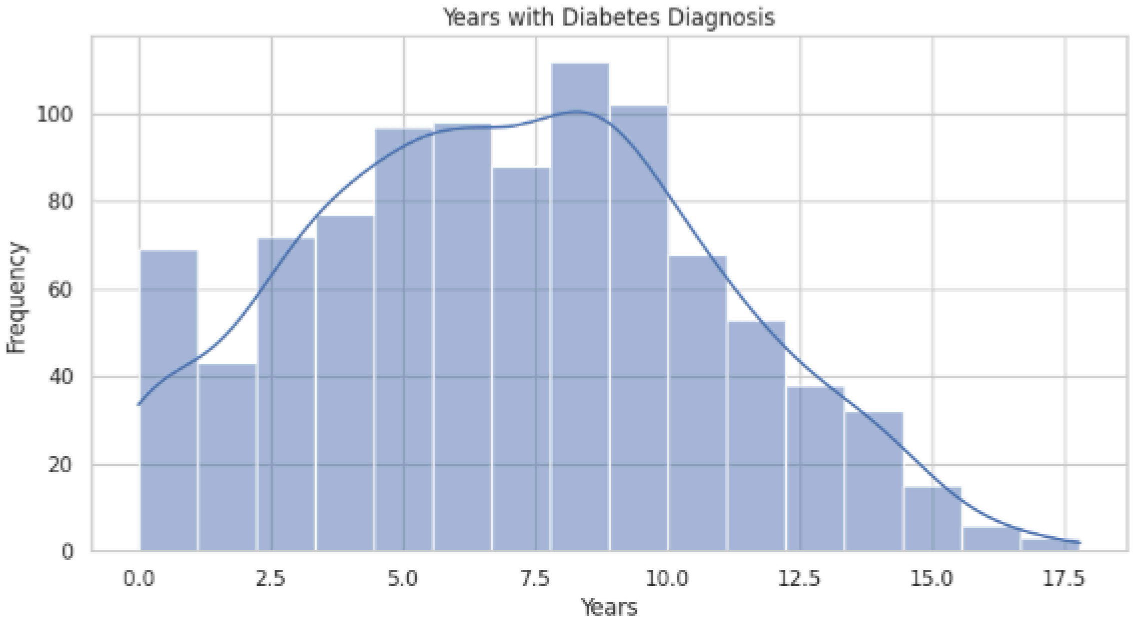
We investigated multiple dimensions of the dataset through analysis to understand its distribution patterns and how different machine learning models performed. Below is an overview of the key analyses conducted:

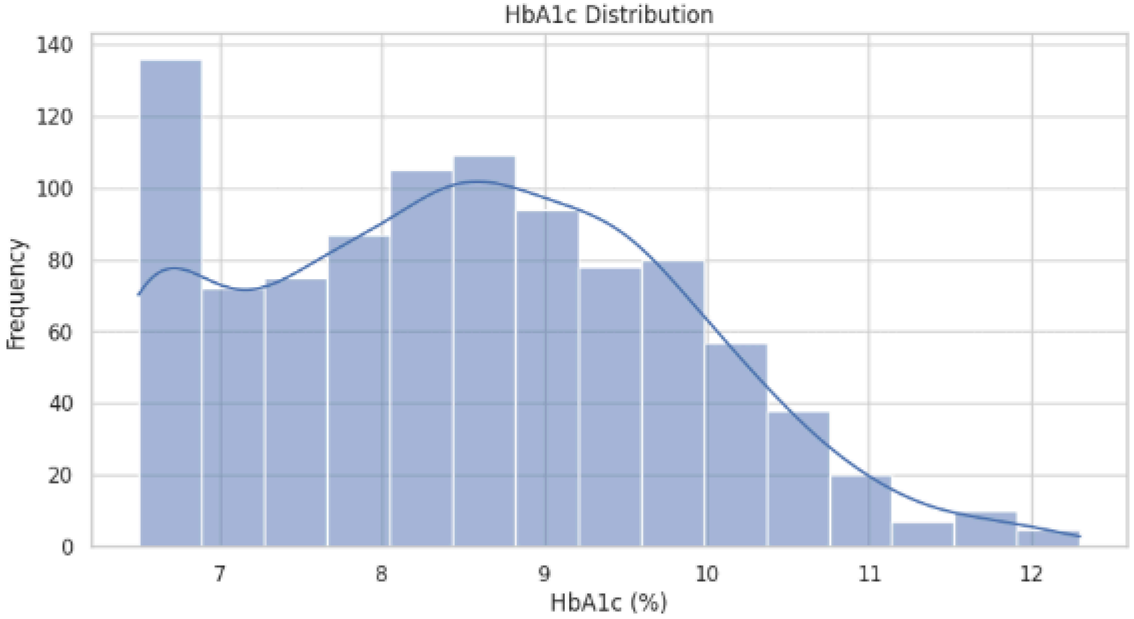
1. Feature Distributions

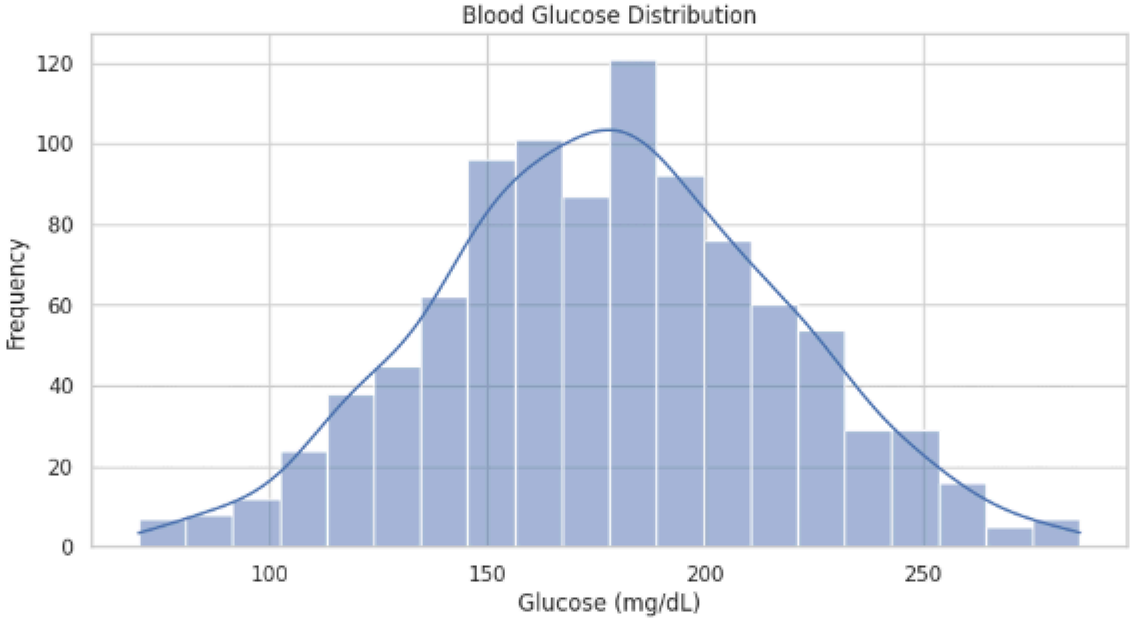
We reviewed the distribution patterns of essential attributes: age, BMI, years with diabetes diagnosis, HbA1c values, blood glucose measurements, cholesterol levels and treatment compliance. The histograms reveal that most of the features do exhibit normal distributions with some showing mild skewness however, HbA1c and blood glucose show a noticeable skewness, this tells me there could be outliers it that these two variables just don't have a perfect normal distribution

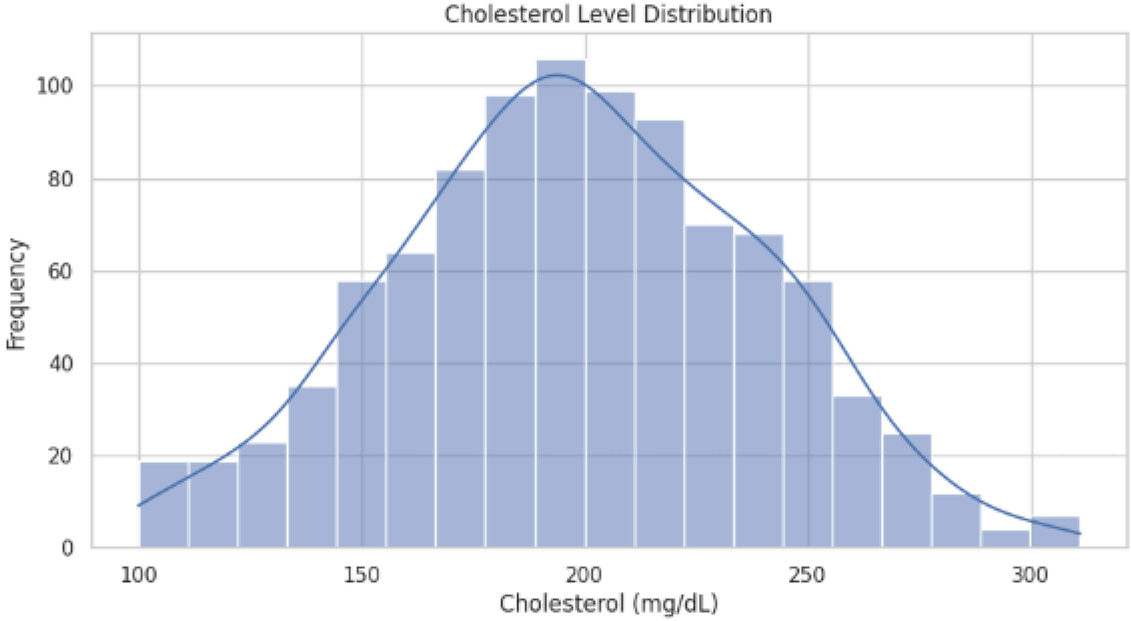

Age Distribution
Frequency
0
20
40
60
80
100
20
30
40
50
60
70
80
90
Age (years)


BMI Distribution
Frequency
0
20
40
60
80
100
15
20
25
30
35
40
BMI


Years with Diabetes Diagnosis
Frequency
0
20
40
60
80
100
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Years

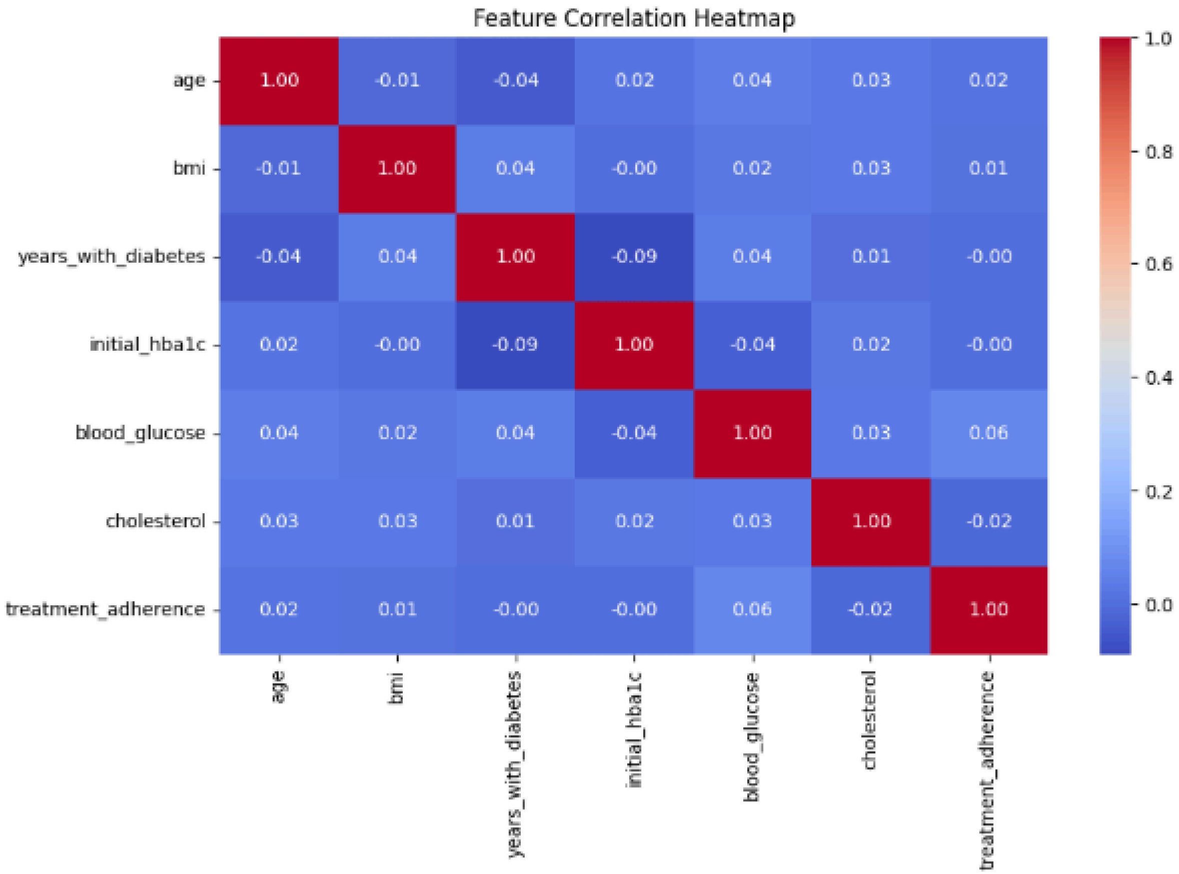

HbA1c Distribution
Frequency
0
20
40
60
80
100
120
140
7
8
9
10
11
12
HbA1c (%)





2. Feature Correlation Analysis

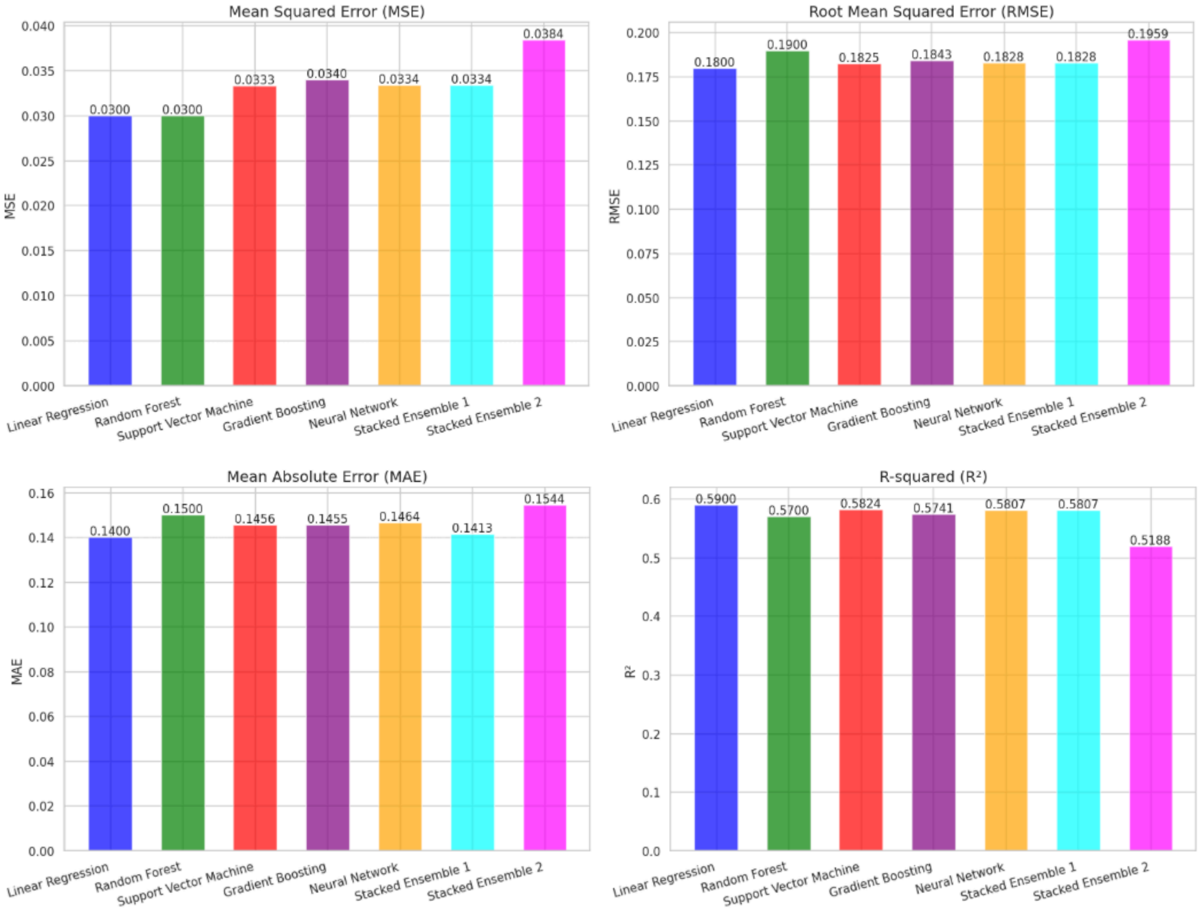
A correlation heatmap was also created to see how the variables in the dataset are related: Most of the features have a very low correlation with other features in the dataset. There was a weak negative correlation between the initial HbA1c and the years with diabetes, which may suggest that the length of the disease affects the glucose control.



3. Model Performance Comparison

Several machine learning models were evaluated using a set of error metrics:

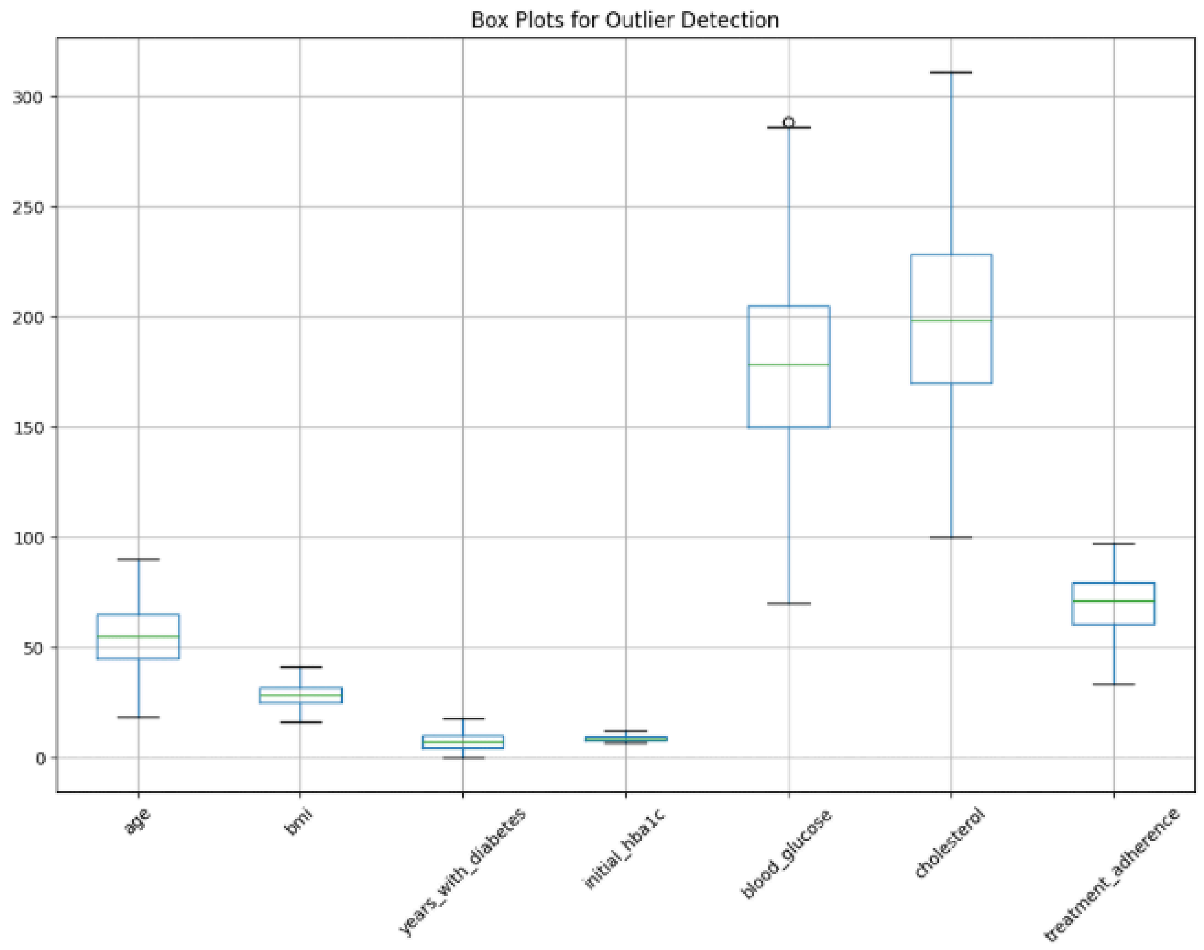
The performance of the models was evaluated using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) to assess the prediction accuracy. MAE gives information about the average prediction errors. R-squared (R^2) and Adjusted R^2 were used to evaluate model fit. From the results the gradient boosting and random forest produced the best results since they had the highest R^2 values.The support vector machine (SVM) had the lowest R^2 value of all the models, hence indicating a poor model fit. Neural Network and Linear Regression had fairly good accuracy and their performance was fairly similar.



4. Outlier Detection

Box plots were used to visualise the outliers in the dataset. There were high outliers in blood glucose and cholesterol which may compromise the models' performance.

The other variables had narrow ranges with no extreme variations.



I used an Agile approach during this project. I began by identifying project tasks as data preprocessing and model evaluation in addition to visualisations and results interpretation, this way I could concentrate on delivering better outcomes at each stage.Many data analysis projects require an agile methodology as it provides adaptability alongside continuous feedback and collaboration.

Model Development and Results

There are many different ways to tune the hyperparameters in a model, this is done to improve a models performance in terms of achieving higher accuracy, lower errors and better predictions.

Manual hyperparameter selection requires setting parameters through personal experience or a process of learning from results which is typically slow. Cross-Validation Grid Search systematically covers all hyperparameter settings through a predefined search within each validation subset which can become computationally costly. Randomised Search Cross-Validation (RandomisedSearchCV) is different from GridSearchCV in that it does not exhaust the hyperparameter space through random sampling of this space to find a balance between the time required and optimal model performance. Through probabilistic techniques such as Gaussian processes Bayesian Optimisation via Keras Tuner makes predictions about hyperparameter effectiveness through the application of past evaluation results to iteratively improve the search process and typically requires fewer trials to reach an optimal solution.

I have decided to have a go at doing random forest, support vector machine, neural network, linear regression and gradient boosting algorithms as well as two stacked ensembles so I will focus on only a few here:

Linear Regression (Ridge Regression)

My linear regression model I implemented was ridge regression which is L2 regularisation to prevent overfitting. To find the best value of alpha (that determines the amount of regularisation), I used grid search with 10-fold cross validation to assess the model. The final value of alpha was 21.54, which indicates a moderate amount of regularisation.

The feature importance was calculated from the model coefficients and it revealed that the initial HbA1c levels, severity of side effects and treatment compliance were the most important features. In order to determine bias and fairness, prediction errors were compared between demographics (gender, age group, and diet type). The fairness analysis showed rather small potential biases in errors for the older people and male patients which are the older people and male patients having slightly higher errors than the rest which needs to be looked at to increase equity of the predictions.

Support vector machine (SVM)

For my SVM model I performed grid search with 5-fold cross-validation (At times 10 fold can be really resource intensive and make my code run for a long time), to identify the best values of C (regularisation), epsilon (tolerance) and gamma (kernel coefficient). The optimal values of parameters (C = 1, epsilon = 0.1, gamma = 0.01) got R^2 of 0.5827 and adjusted R^2 of 0.5094 which indicates a moderate fit.

HbA1c values and treatment adherence were the most influential features according to permutation importance scores. The fairness analysis showed rather small differences in errors for the older people and males, that is the differences in performance between the groups are rather small and therefore predictions should be watched for equity.

Neural network

My neural network was optimised using Keras Tuner with random search over the number of neurons per layer, dropout rates and learning rate. To avoid overfitting, 5-fold cross validation was used which gave a cross validated R^2 of 0.4395 and test R^2 of 0.5369. Although neural networks do not provide the feature importance in the same way as linear models do, permutation importance or SHAP values can be used for feature importance. The fairness analysis showed similar errors for both genders but the errors were lower for the younger patients which could indicate a potential age bias in the model's performance which should be kept an eye on.

Business impact and Recommendations

The models incorporated in this work used patient demographics, medical history, lifestyle factors, and laboratory test results in predicting large HbA1c level reducers. Gradient boosting and support vector machine models were the best performing models and had a moderate level of accuracy in prediction that can be used in formulation of patient specific treatment plans. The feature importance evaluation can be used to identify factors that are likely to have an impact on treatment success such as the baseline HbA1c values and patient compliance which can be used to design patient education and adherence support strategies. The fairness analysis is used to determine that the model makes similar predictions for all patient demographics so that PharmaTech can determine where there are biases and correct them before implementing the model in clinical practice. This predictive system can be incorporated into the clinical workflows that PharmaTech develops to help doctors with model-based recommendations for medication decisions.

Conclusion and Future Work

My stacked ensemble model which has a neural network as the base model and a ridge regression as the meta model has the ability to improve the prediction of the patient's response to the diabetes treatment.The integration of the fairness analysis ensures that the model works uniformly for all the groups of people, avoiding the possibility of bias in recommending treatment options. Although the results are fairly good in terms of prediction accuracy, there is still much room for enhancement such as improving feature selection, collecting more clinical and lifestyle information and trying other ensemble methods including blending many base models. Future work should also involve hyperparameter tuning, more explainability tools for enhanced understandability, and practical validation using patient trial data to determine the clinical significance.