Diabetes Prediction Model

By: Jeffrey Ongicho

Overview

- Diabetes is a chronic disease that occurs when the pancreas does not produce adequate insulin or when the body is unable to effectively use the insulin produced.
- There are two types of diabetes: type 1 diabetes and type 2 diabetes. Type 1 diabetes is also known as insulin-dependent and is characterized by deficient insulin production by the body. Patients suffering from this require daily administration of insulin.
- Type 2 diabetes affects the body's use of insulin and leads to high levels of blood sugar

Business Understanding

- According to the World Health Organization, there was a 3% increase in age standardized mortality rates from diabetes between 2000 and 2019. In lower income countries it increased to 13%.
- Factors affecting type 2 diabetes include: overweight, genetics, lack of exercise. Keeping a healthy lifestyle is a key component in keeping diabetes away.

Business Problem

- There are certain health indicators that could assist someone in knowing how susceptible they are to suffering from diabetes.
- Genetics play a role as well as other factors such as lifestyle which includes eating habits, smoking, and exercise.
- Early intervention is key and would help give preventive care to individuals at high risk and ultimately reduce the number of deaths from diabetes.

Objective

- The primary goal is to create a diabetes prediction model that can be used to predict the individuals at risk of suffering from diabetes. By doing this, individuals are able to get early preventive care to reduce the impact of diabetes.

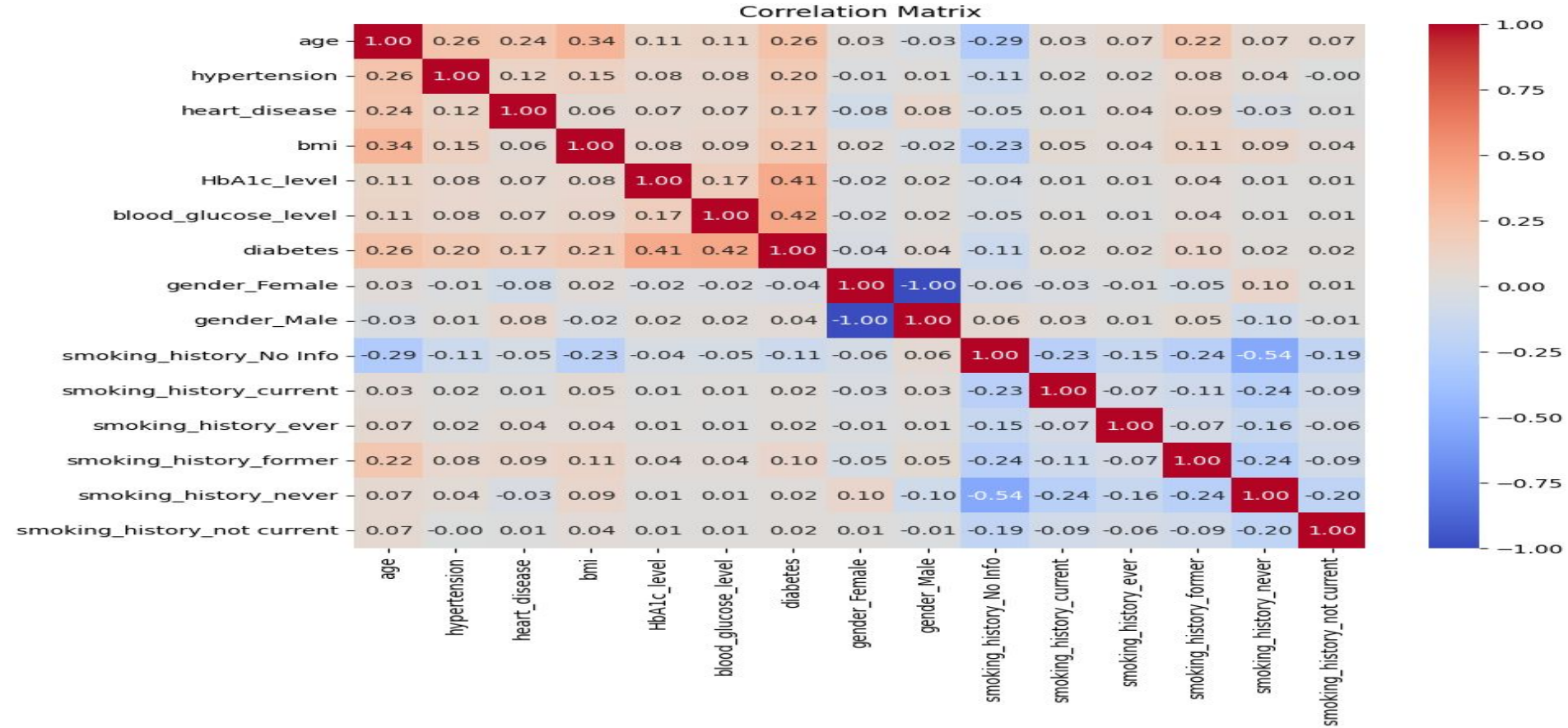
Data Understanding

- The dataset used in this project is sourced from here: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data.
- It contains 100,000 rows and 9 columns. The 9 columns are:
- - Gender
- - Age
- - Hypertension
- - Heart disease
- - Smoking history
- - BMI
- - HbA1c- level (average blood sugar level over the last 2-3 months)
- - Blood Glucose level
- - Diabetes
-

Data Preparation

- Before performing exploratory analysis and modelling on the data certain issues such as missing values and duplicates have to be handled first. The data was found not to have missing values which is good because missing values may tamper with the accuracy of the model.
- The were 3854 duplicates found in the dataset which were removed to ensure the integrity and consistency of data.
- In the gender column we remove 18 rows where the value is "Other" because the primary focus is on male and female individuals.
- There were two categorical variables in gender and smoking history. One hot encoding is used to incorporate them into the model and new columns are created.
- A correlation matrix is plotted to see how the features correlate to the target variable "diabetes"

Data Preparation (Cont..d)



Feature Selection

- The dataset is split into test and training sets. The test set is used for evaluation of the model based on the modelling done on the training set.
- For our baseline model we will be scaling the values to standardize the range of independent features in the dataset.

Modelling

- Under modelling three models will be created. The first one is a baseline model that is a simple logistic regression model, the second one is a decision tree without optimal hyper parameters and the third one is a decision tree using optimal hyper parameters after using Grid Search CV.
- The accuracy of 0.96 is high but maybe misleading because our target variable is imbalanced. Precision and recall are better metrics to use in this scenario and we can see that the precision is high while the recall is low.
- Our second model uses a decision tree which may be better because our dataset is quite large. The accuracy has reduced slightly while the precision has gone down considerably. The recall however, has improved.
- The third model uses Grid Search CV to find the optimal combination of parameters to use for our decision Tree classifier. The accuracy has improved to 0.97 the recall has largely remained the same and the precision has gone up to 0.99.

Model evaluation

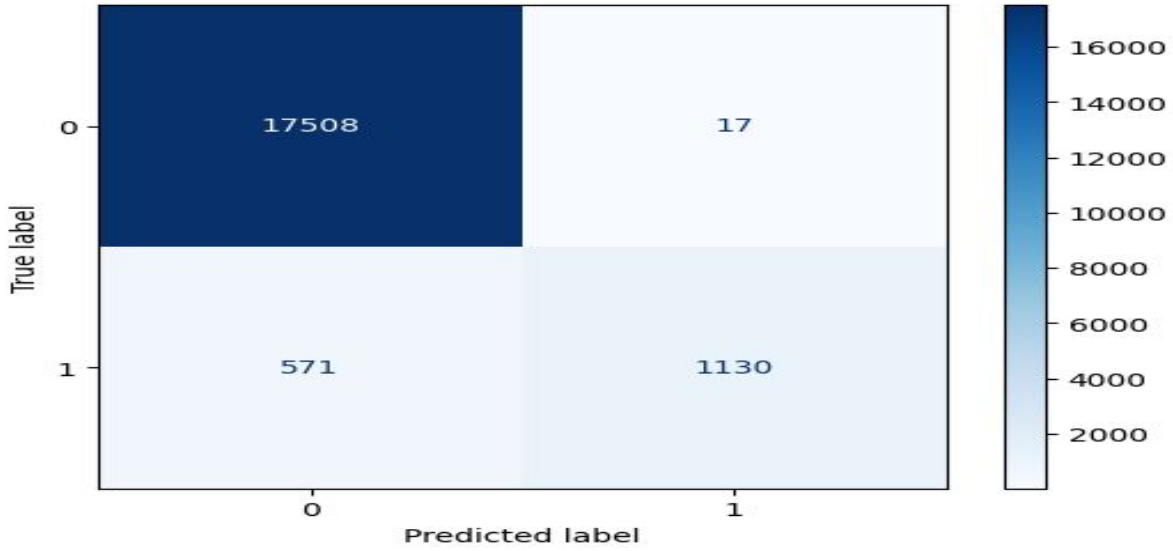
The metrics used to evaluate the model are:

- Accuracy
- F1-score
- Precision
- Recall

The final model will be the one used for evaluation.

Confusion matrix

- The confusion matrix is a table that summarizes the predictions against the actual outcomes. This will give us an indication of how well the model does at predicting who is actually suffering from diabetes.



Confusion Matrix (Cont..d)

Based on the confusion matrix:

- The TP (True Positives) are 1130. This means that 1130 individuals were correctly predicted to have diabetes.
- The TN (True Negatives) are 17508. This means that 17508 individuals were correctly predicted to not having diabetes.
- The FN (False Negatives) are 571. This means that 571 individuals actually had diabetes but were predicted not to.
- The FP (False Positives) are 17. This means that 17 individuals were wrongly predicted to having diabetes but they did not actually have it.

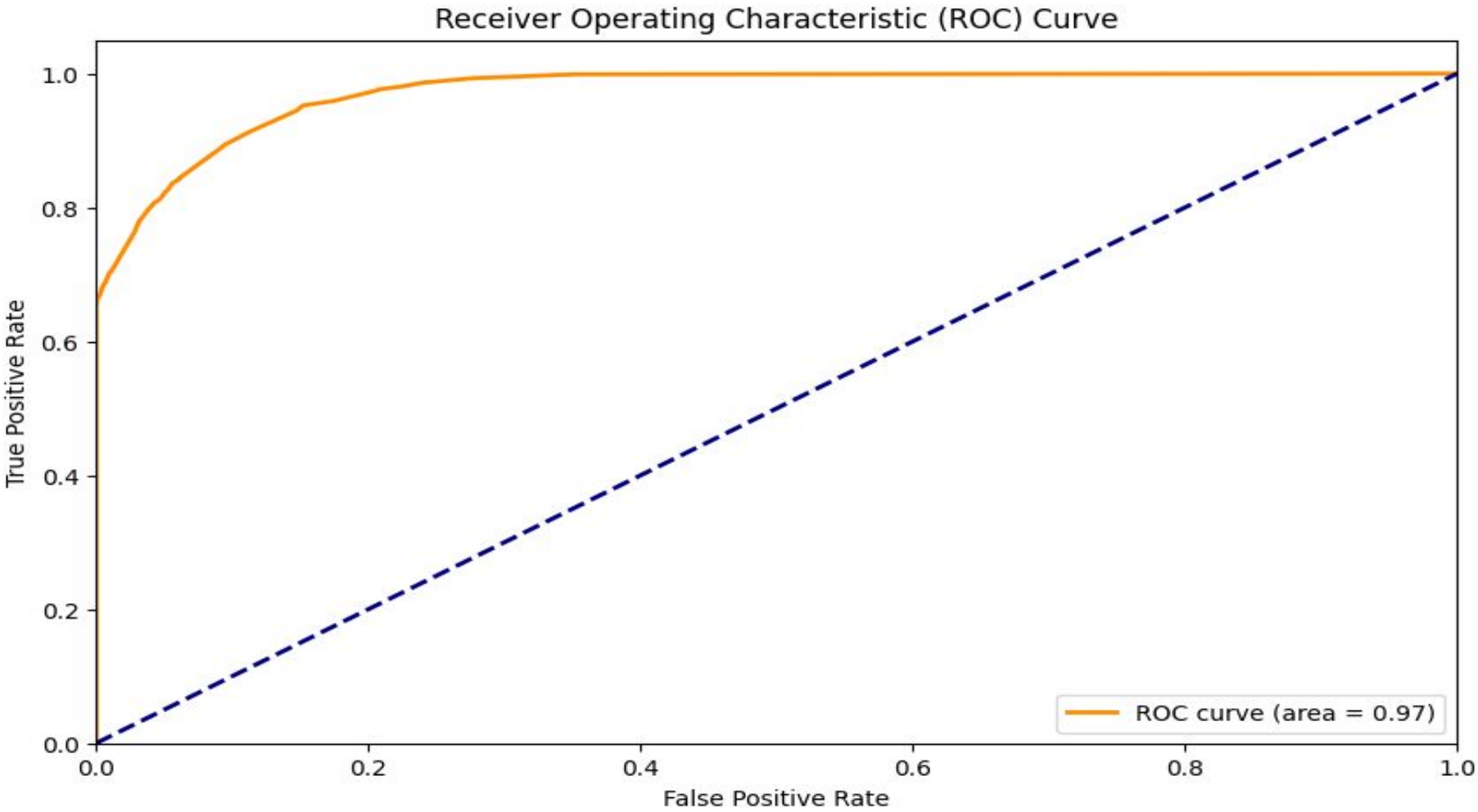
Metrics

The model achieved the following metrics:

- The accuracy is: 0.9694164152709872
- The Precision is: 0.985178727114211
- The Recall is: 0.6643151087595532
- The f1-score is: 0.7935393258426967

ROC curve and AUC

There is huge class imbalance in our target variable which means that the model gets a high accuracy because it is able to predict many instances of the majority class. An ROC curve would give us a better picture of the performance of the model. The ROC curve shows the true positive rate against the false positive rate of our model. Ideally the model with a ROC curve furthest top left has the best performance because the AUC is close to 1 which is a perfect model while 0.5 value for AUC means it is random guesses.


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.97)

Recommendations

From the results of the analysis we can see the following:

- The model shows a high level of performance with a 96% accuracy which means it correctly predicts individuals with diabetes.
- The AUC of 0.97 shows that the model has a great ability of distinguishing between the positive and negative cases.
- A f1-score of 0.79 shows that there is a decent balance between precision and recall from the model.
- 66% recall does indicate that there could be further improvement in fine tuning the model.
- 98% precision shows that the model can predict a lot of positive cases.

Health care providers can use the model for early detection and develop personalized treatment plans for individuals suffering from diabetes.

Health insurance companies can use the model to take note of high risk areas especially in lower income areas and adjust their health coverage accordingly.

Recommendations (Cont..d)

From the analysis the following recommendations can be drawn:

- Ensemble models can be used to improve the performance.
- Different data with new features i.e pregnancy can be used to further fine tune the model.