Health Care Diabetes Challenge

Every data science challenge begins with understanding the problem one is trying to solve. Taking a critical look at the challenge;

Diabetes Mellitus is an endocrine disease of importance, with multiple comorbidities and complications occurring due to the body's inability to manage its glucose level efficiently, It is characterized by varying levels of Insulin intolerance or deficit.

The types of Diabetes Mellitus (DM) occurrence include

1. Type 1 DM

2. Type 2 DM

3. Gestational DM

Gestational Diabetes Mellitus describes when a pregnant woman experiences any form of glucose intolerance with its first occurrence in pregnancy, this has notable risk to the mother and fetus and thus, the increase of Diabetes during pregnancy must be predicted and managed effectively to prevent mortality and other severe morbidities.

Objectives

- To explore factors that contribute to increased Diabetes incidence in pregnancy
- To explore health factors that can be used to most effectively predict Diabetes risk

Problem Statement

- To create a supervised learning, classification model that accurately predicts the risk for diabetes.

Now, that we have a full understanding of the problem, the approach is solving the problem needs to be well-defined for a successful project.

Here's the approach we used to tackle the problem;

1. Data Collection.

2. Data Cleaning and Validation.

3. Exploratory Data Analysis.

4. Feature Preprocessing.

5. Modelling

6. Feature Importance.

1. Data Collection

Every data science challenge needs the right data. Right data leads to a good solution. The dataset used for this project was sourced secondarily. Here are the columns available in the dataset;

1. Id: Unique identifier for each data entry.

2. Pregnancies: Number of times pregnant.

3. Glucose: Plasma glucose concentration over 2 hours in an oral glucose tolerance test.

4. BloodPressure: Diastolic blood pressure (mm Hg).

5. SkinThickness: Triceps skinfold thickness (mm).

6. Insulin: 2-Hour serum insulin (mu U/ml).

7. BMI: Body mass index (weight in kg / height in m^2).

8. DiabetesPedigreeFunction: Diabetes pedigree function, a genetic score of diabetes.

9. Age: Age in years.

10. Outcome: Binary classification indicating the presence (1) or absence (0) of diabetes.

The data was loaded into a Jupiter notebook file with the use of the Pandas library. The dataset was found to have 2768 observations with 10 columns.

2. Data Cleaning and Validation

The dataset was properly cleaned and validated to avoid working with dirty data. The dataset was found to be void of null values.

The Outcome column is the dependent variable with only two classes - 0 and 1. The BMI and DiabetesPedigreeFunction columns are of Float64 data type while the remaining seven columns are of Int64 data type.

After checking for duplicate values in the data, 1990 duplicates were found and dropped, leaving us with 776 observations.

3. Exploratory Data Analysis

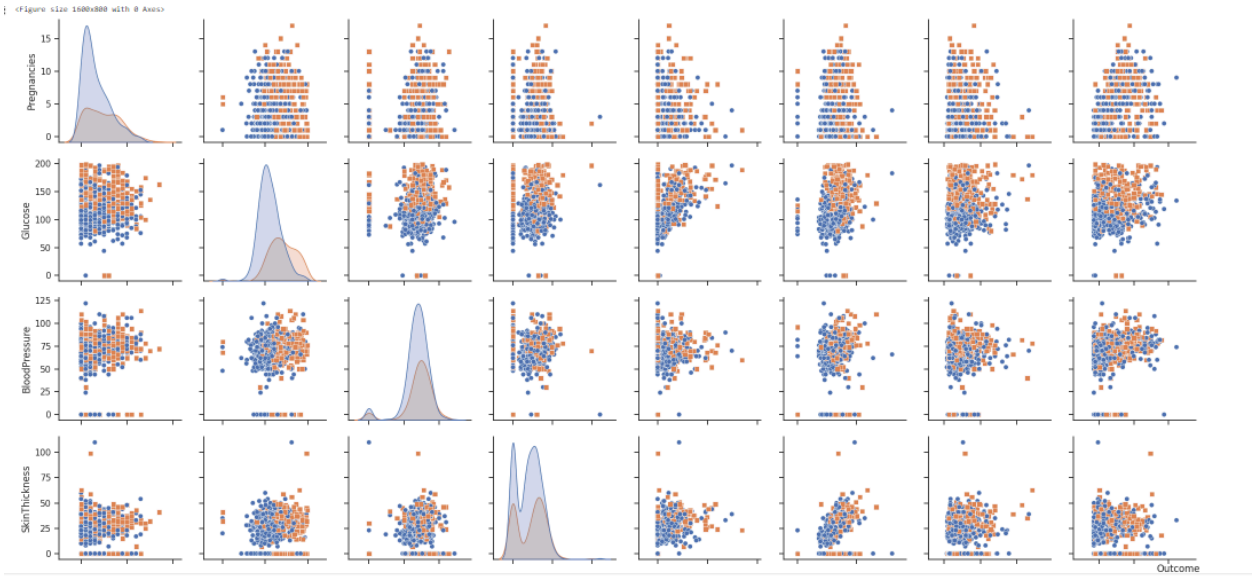
This aspect of data science helps in understanding dataset intricacies. Without EDA, there's no data science. The EDA was divided into two sections - Univariate and Bivariate.

Univariate Analysis

This analysis was done to explore the distribution of each independent variable with the aid of boxplots and histograms. Some of the features were skewed while some were normally distributed.

Bivariate Analysis

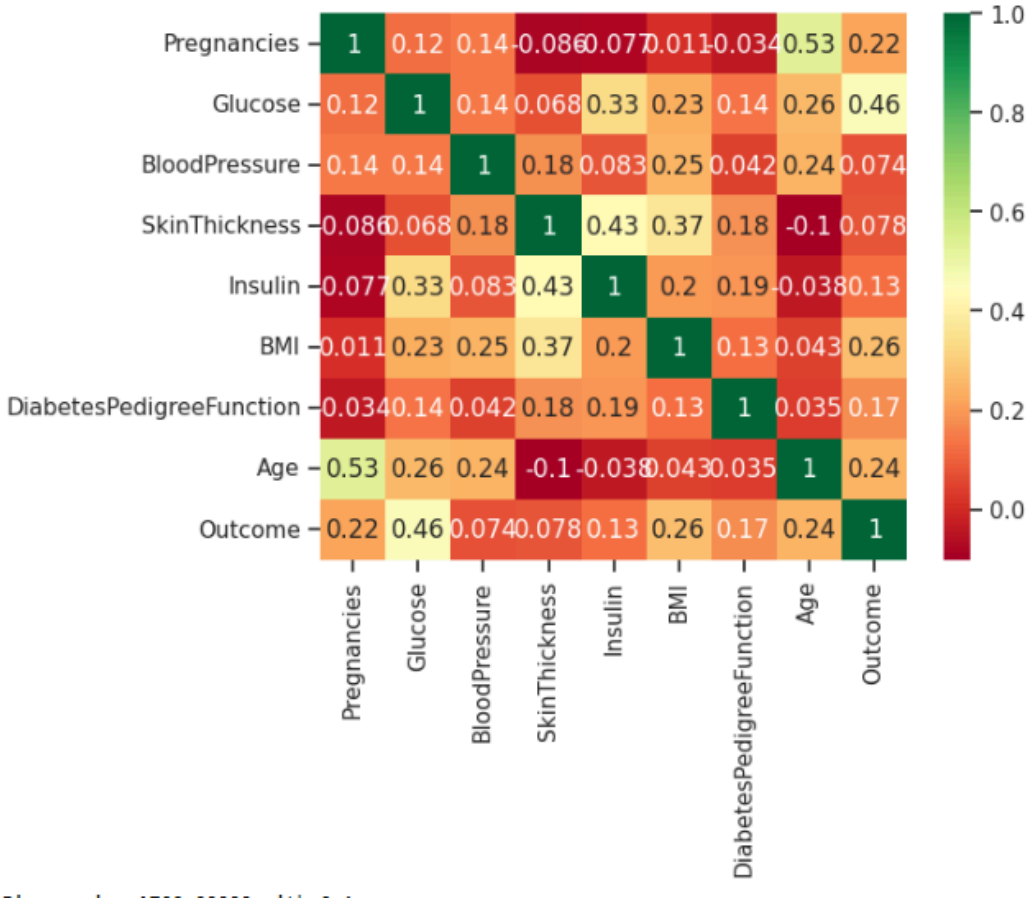
This analysis was used to explore the relationship between the dependent and independent variables with pair plots.



4. Feature Preprocessing

This aspect is essential to prepare the features well for modeling.

First, the correlation between the features was explored to ensure that each feature is independent of each other.



Also, the data was split into the train and test set in an 80:20 ratio using the train-test split validation library.

The train set features and the test set features were scaled using the Standard Scaler library for standardization.

5. Modeling

A baseline model - Logistic Regression was first trained on the data before introducing the Random Forest and XGBoost models.

The Random Forest and XGBoost models were tuned with the Randomized Search method for better optimization.

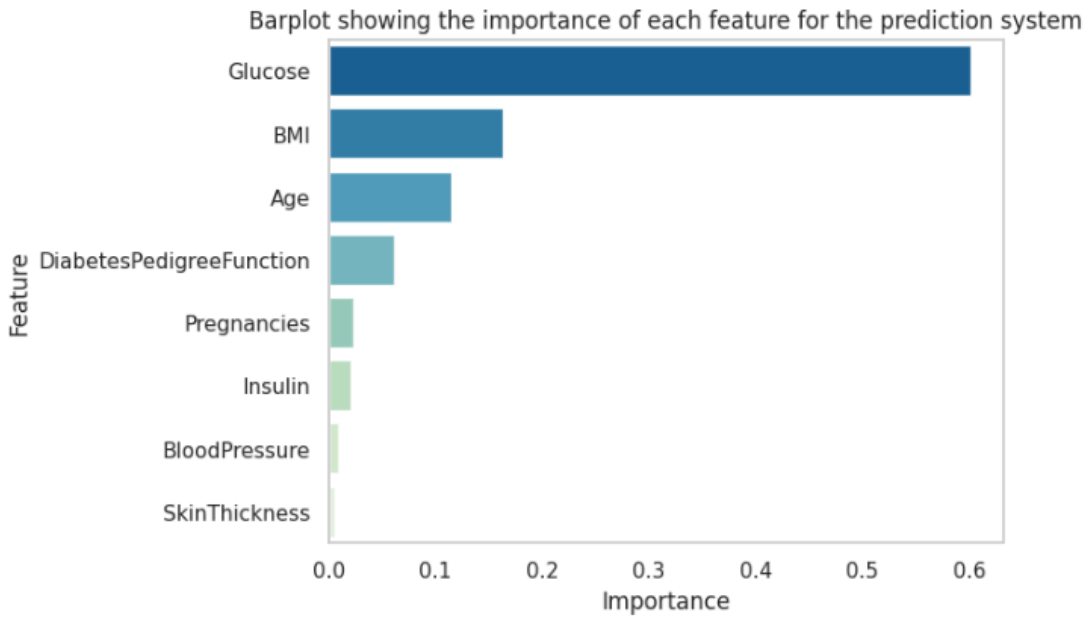
The models were evaluated based on the F1-score due to an imbalance of target classes.

Model	Training score	Validation score
Random Forest (without tuning)	1.0	0.640777
Xgboost (without tuning)	1.0	0.660377
Xgboost (tuned)	0.675192	0.666667
Random Forest (tuned	0.642105	0.653061

The tuned Random Forest model proved to be the best.

6. Feature Importance

The importance of each feature was derived to understand the factors that have a high influence on determining the diabetic status of a patient.



The glucose level of an individual has the highest influence on the diabetic status of an individual, followed closely by the BMI and Age.

Conclusion

Access to more data will greatly improve the performance of the model. Leveraging the power of data science, Team D has been able to develop an optimal model to predict the diabetic status of a patient.