Team3_D

HEALTH CARE CHALLENGE

PROJECT DESCRIPTION

Welcome to the Diabetes Prediction Challenge, a valuable resource for researchers, data scientists, and medical professionals interested in the field of diabetes risk assessment and prediction. The dataset provided contains a diverse range of health-related attributes, meticulously collected to aid in the development of predictive models for identifying individuals at risk of diabetes.

DATASET:

https://drive.google.com/drive/folders/1ubaqsATGLf0PvDxCgT7nQBl_4_2oehVG?usp=sharing

OBJECTIVES OF THE PROJECT

- To explore factors that contribute increased Diabetes incidence in pregnancy.

- To explore health factors that can be used to most effectively predict Diabetes risk.

WHAT IS DIABETES MELLITUS?

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.

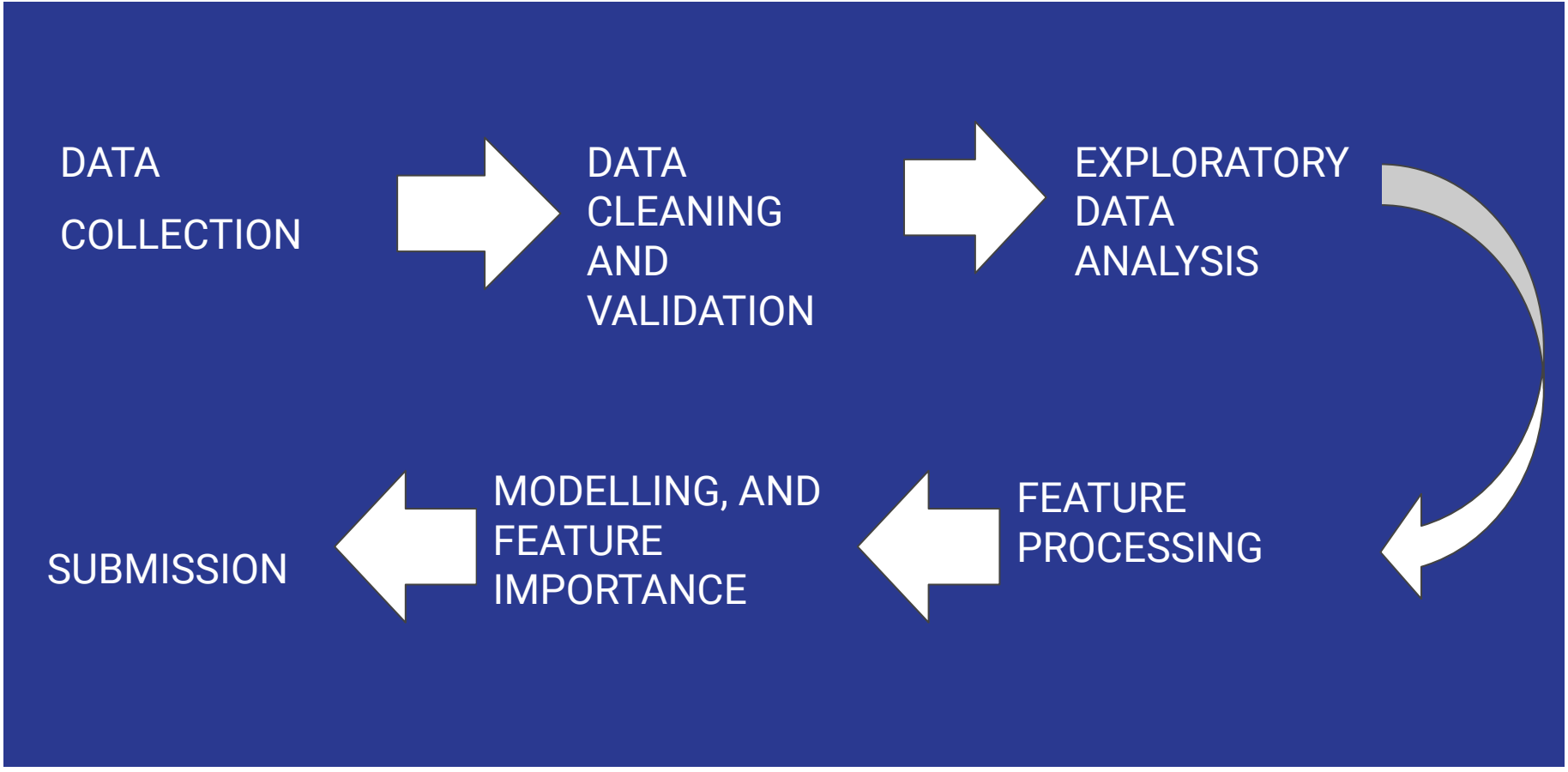
The types of Diabetes Mellitus (DM) occurrence include:

1. Type 1 DM
2. Type 2 DM
3. Gestational DM

PROBLEM STATEMENT

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

DEVELOPMENT WORKFLOW



IMPLEMENTATION

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.

IMPLEMENTATION

Exploratory Data Analysis

The EDA was divided into two sections - Univariate and Bivariate.

Univariate Analysis

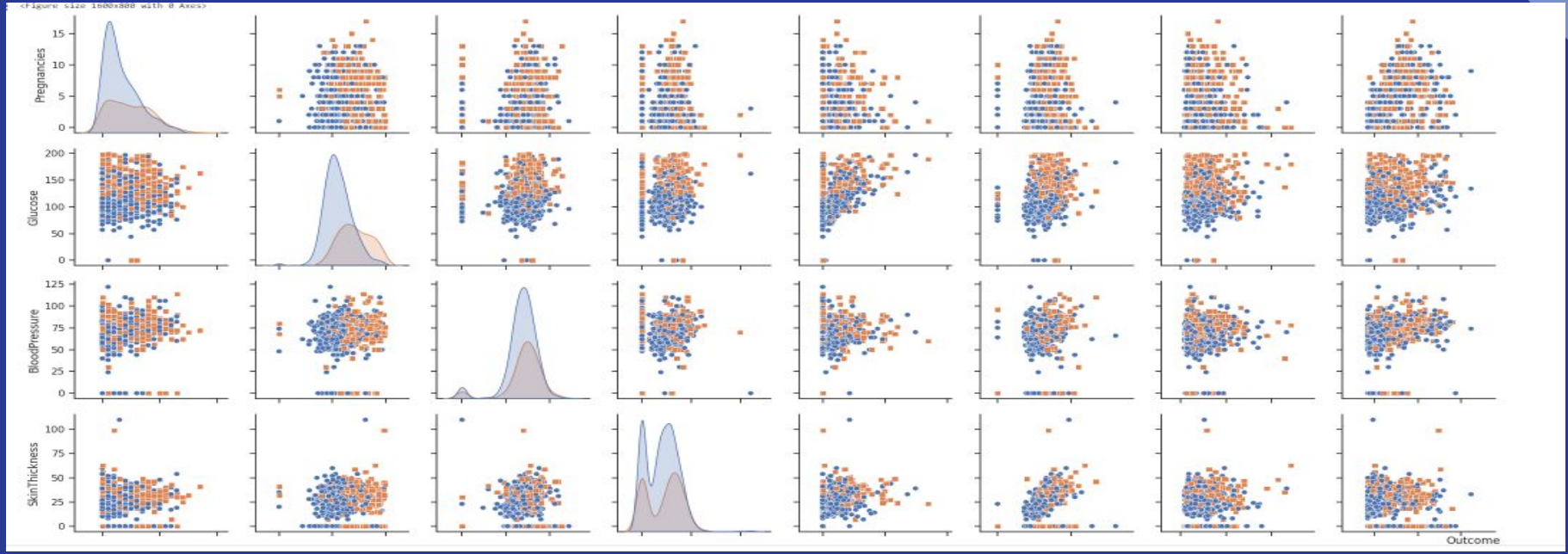
This analysis was done to explore the distribution of each independent variable with the aid of boxplots and histograms. Some of the features were skewed while some were normally distributed.

Bivariate Analysis

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

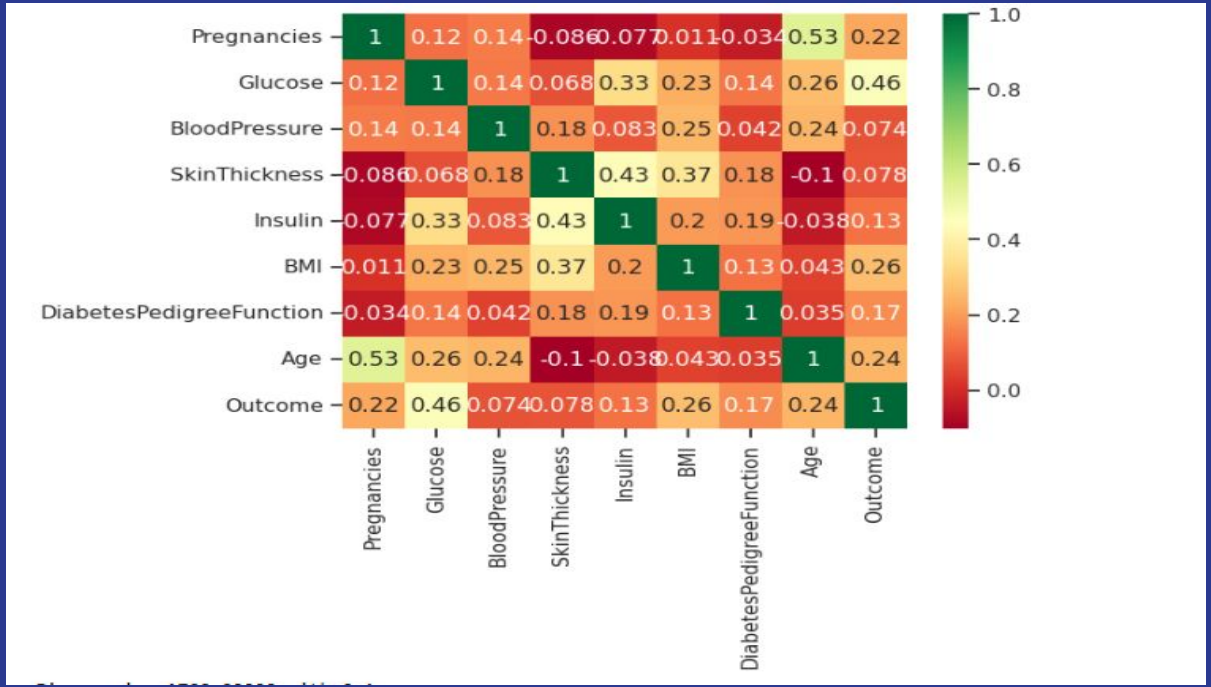
BIVARIATE ANALYSIS

Using Pair Plots



FEATURE ENGINEERING

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.



MODELLING

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.

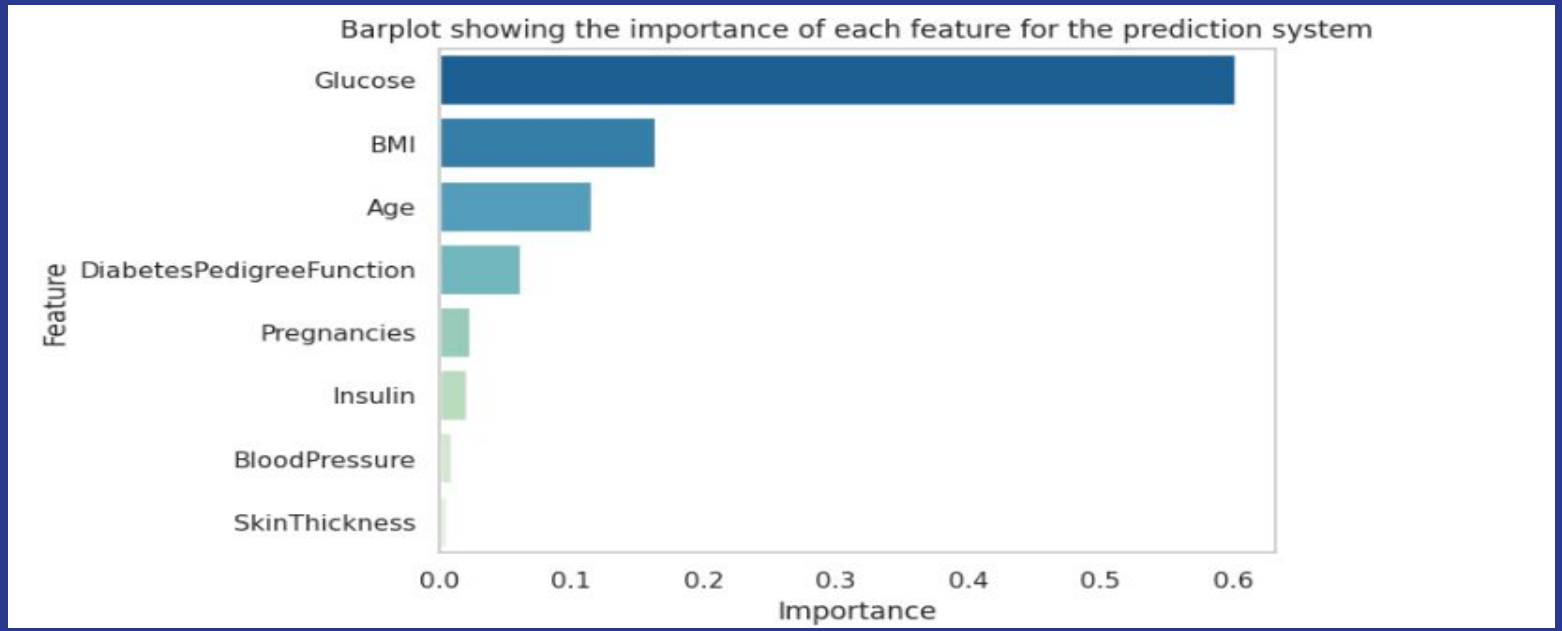
The models were evaluated based on the F1-score due to an imbalance of target classes.

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.

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.

The team

DATA SCIENTIST	DATA SCIENTIST	DATA SCIENTIST
NUNSI SHIAKI	GIFT UKPOWEH	Dolapo