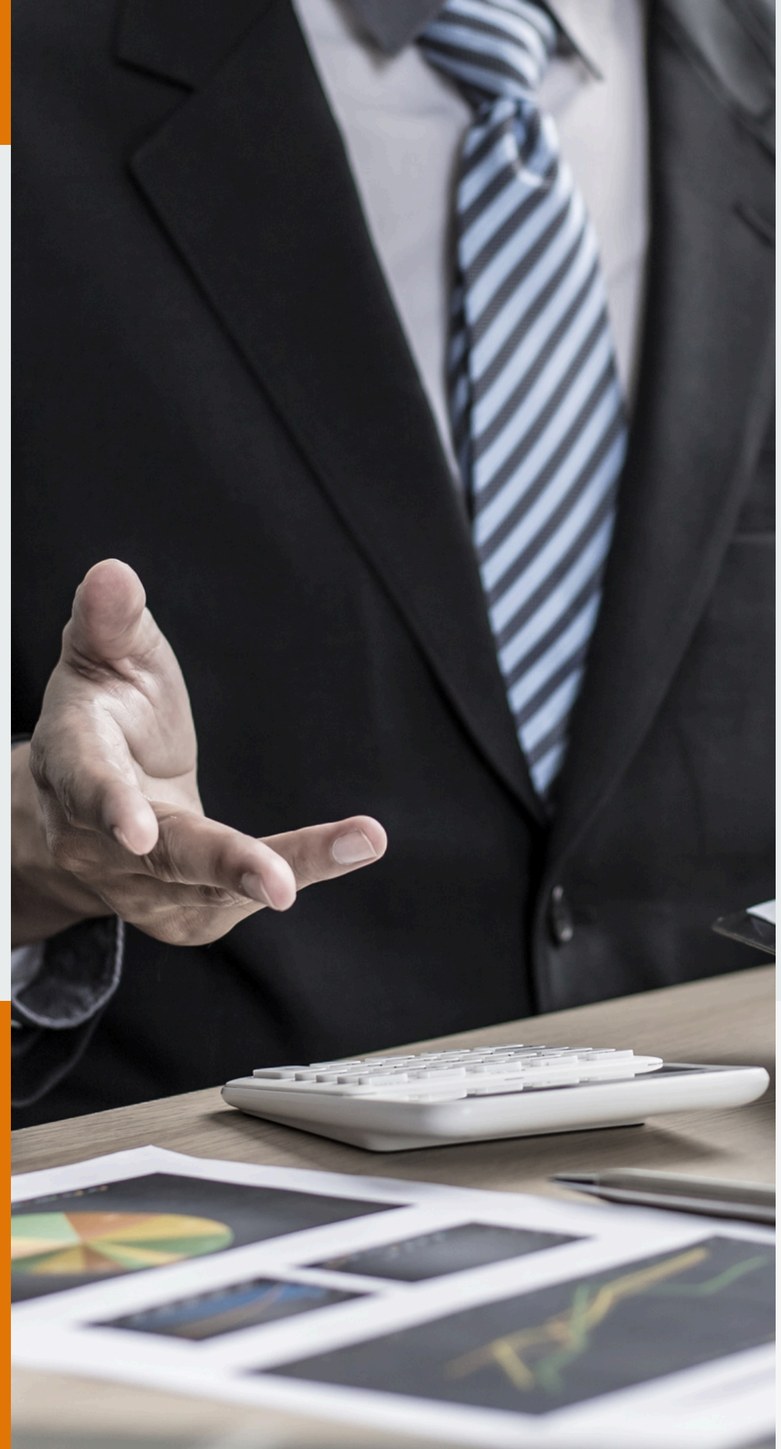
STARK HEALTH CLINIC

DIABETES PREDICTION PROJECT

Helen Ojo



CONTENT

01 Introduction

02 Problem Statement

03 Objectives

04 Workflow

05 Data Visualisation

06 Modeling

07 Evalutaion Metrics

PROBLEM STATEMENT

Diabetes is a health hazard for the patients of Stark Health and a high cost as well. Even though diagnostic tools are at the disposal of this health facility, early detection is less precise. Early interventions are missed due to a lack of proper early detection methodologies, which negatively affects the patient's health and increases costs associated with advanced stages of the disease. The bottom line is an urgent need for better, more accurate, and reliable predictive tools to enable early identification and treatment of at-risk individuals.

OBJECTIVES

Predictive Accuracy

Develop a machine learning model to accurately predict diabetes onset with high sensitivity and specificity, minimizing false negatives and positives.

Operational Impact

Enable Stark Health Clinic to implement targeted measures, improving outcomes through timely interventions and reducing diabetes progression.

Strategic Value

Contribute to reducing patients' financial burden by lowering the costs of advanced diabetes management and supporting proactive, data-driven care.

WORKFLOW

Data Collection and Preparation

- Libraries: pandas, numpy, scikit-learn
- Data Handling: Use pandas to load and manage datasets.
- Data Cleaning: Handle missing values with custom imputation using numpy.
- Outlier Detection: Use statistical methods (interquartile range) to identify and address anomalies in the data.
- Normalization: Apply MinMaxScaler from scikit-learn for scaling numerical features.

Feature Engineering

- Libraries: pandas, numpy, scikit-learn, matplotlib, seaborn
- Feature Selection: Perform correlation analysis using seaborn heatmaps
- Derived Features:
- Create a new feature age_group to categorize patients into bins ("child," "adult," "elderly").
- Visualization for EDA:
- Pairplot with hue="Diabetes" to explore relationships between features.
- Use histplot to visualize feature distributions
- Encoding: Apply one-hot encoding for categorical variables as needed.

WORKFLOW

Model Development

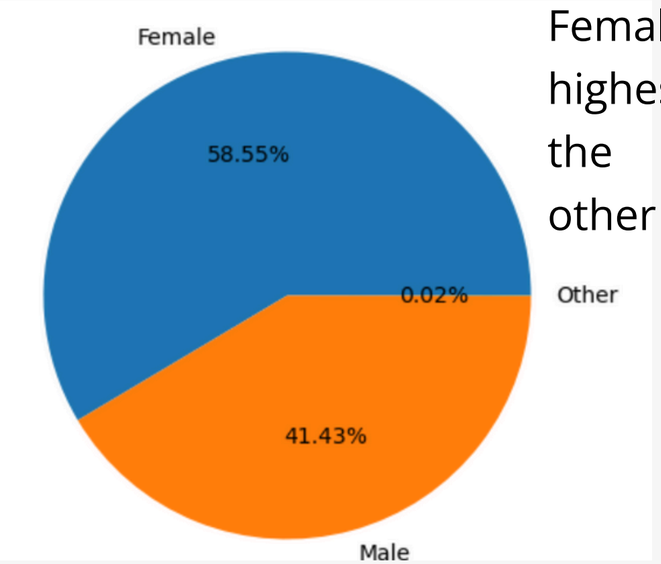
- Libraries: scikit-learn
- Algorithm Selection: Use RandomForestclassifier.
- Training and Validation: Split data using train_test_split and perform cross-validation with GridSearchCV.
- Hyperparameter Optimization: Use GridSearchCV for more advanced tuning.

Model Evaluation and Validation

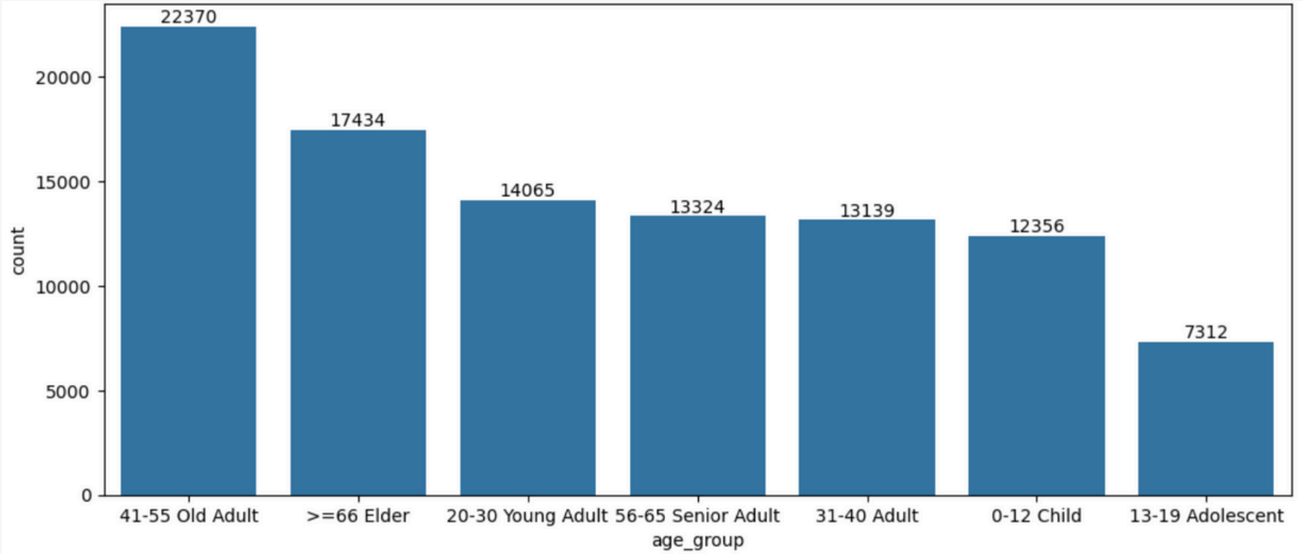
- Libraries: scikit-learn, matplotlib, seaborn
- Evaluation Metrics: Use classification_report and confusion_matrix from scikit-learn.
- Visualization: Plot confusion matrices using matplotlib for performance analysis.
- Error Analysis: Analyze misclassified cases by comparing predictions to actual labels in pandas.

DATA VISUALISATION

EXPLORATORY CHART

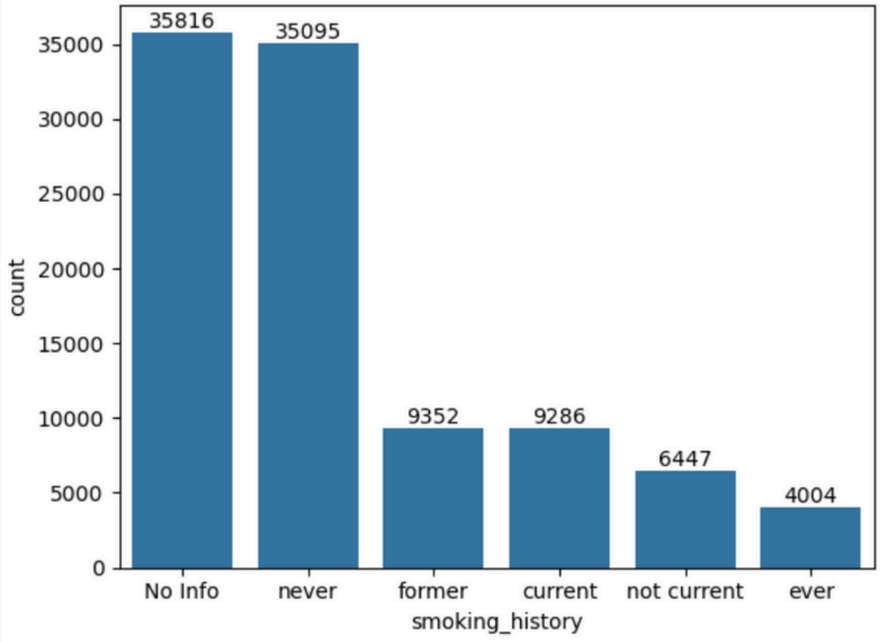


Female Patient highest while the lower is other

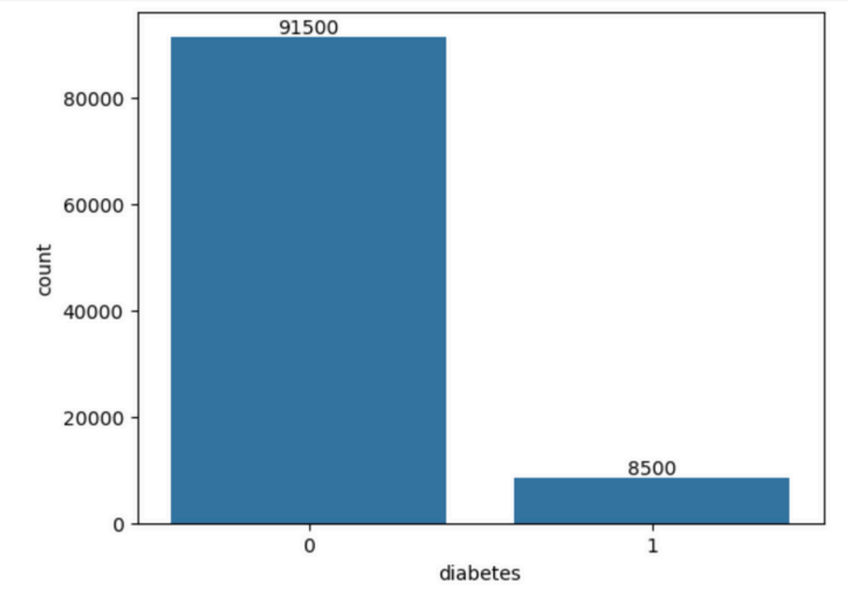


The 41-55 (Old Adult) highest while the lowest is the 12-19 (Adolescent)

Patient with no info record has highest while the lower is ever

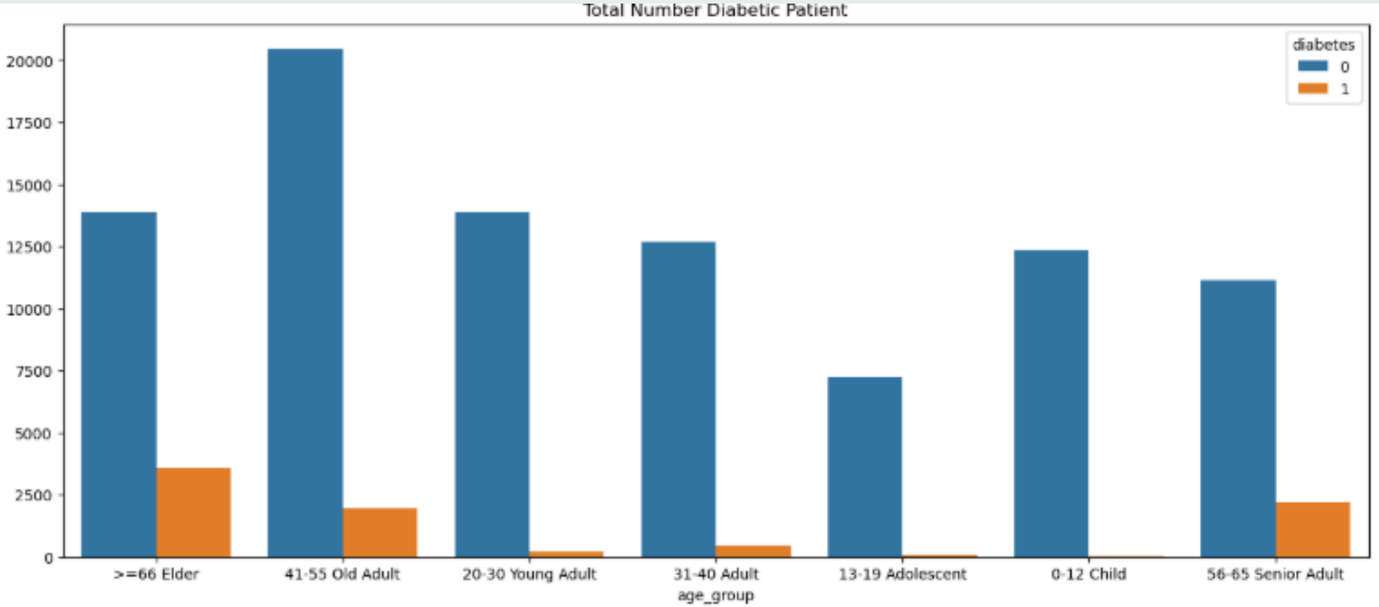


Record of 0 has higher than while 1

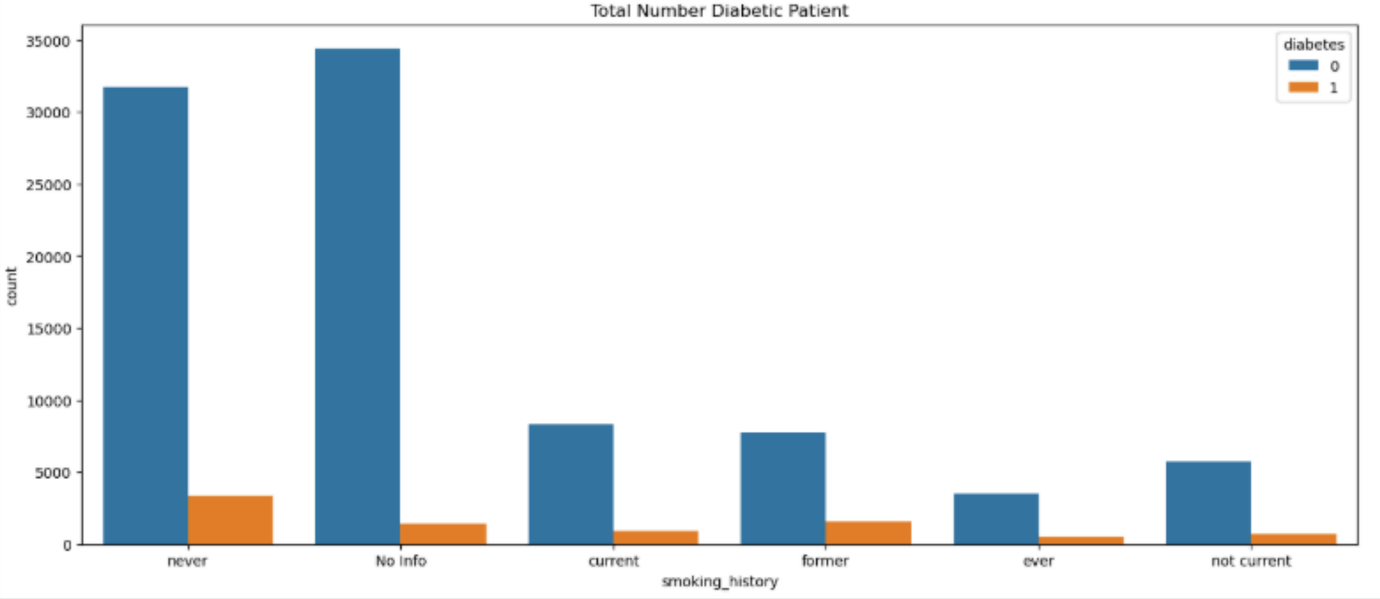


DATA VISUALISATION

EXPLORATORY CHART



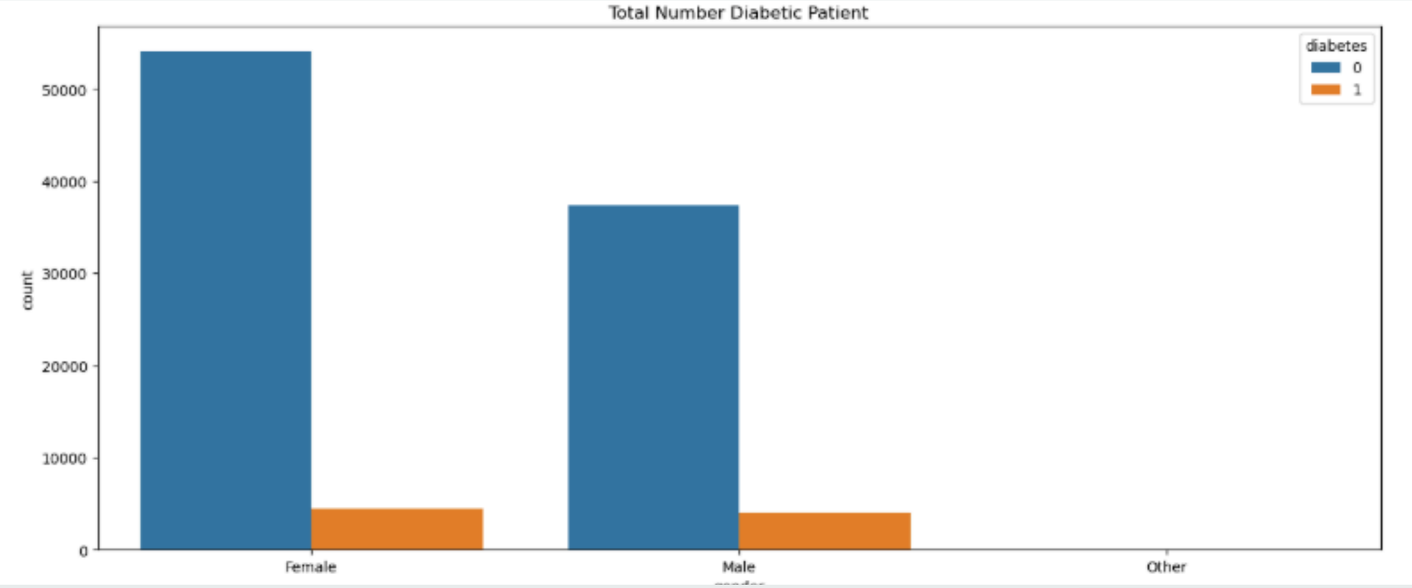
Diabetes is prominent in the age group above 66 less for the age group 0-12.



Diabetes is prominent in the never smokes and lesser for the ever.

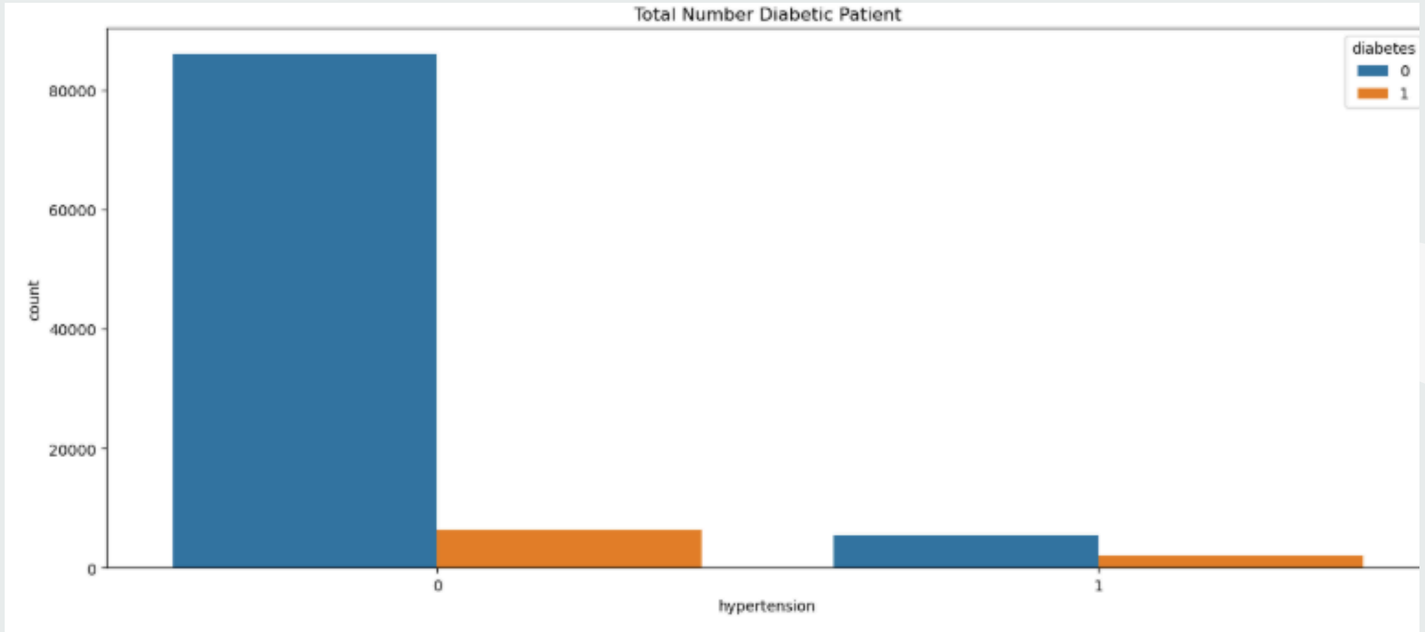
DATA VISUALISATION

EXPLORATORY CHART



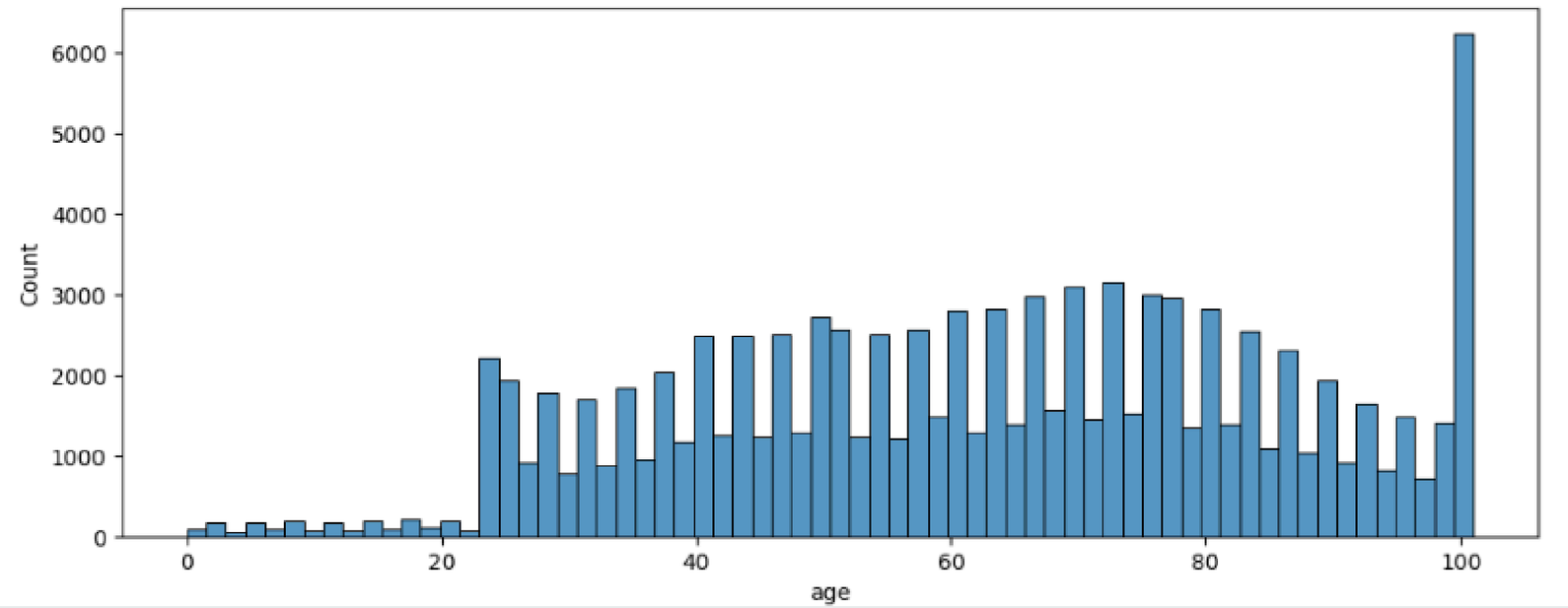
Diabetes is higher in females and less in males.

Diabetes is higher in type 0 hypertension and lesser in type 1 hypertension.

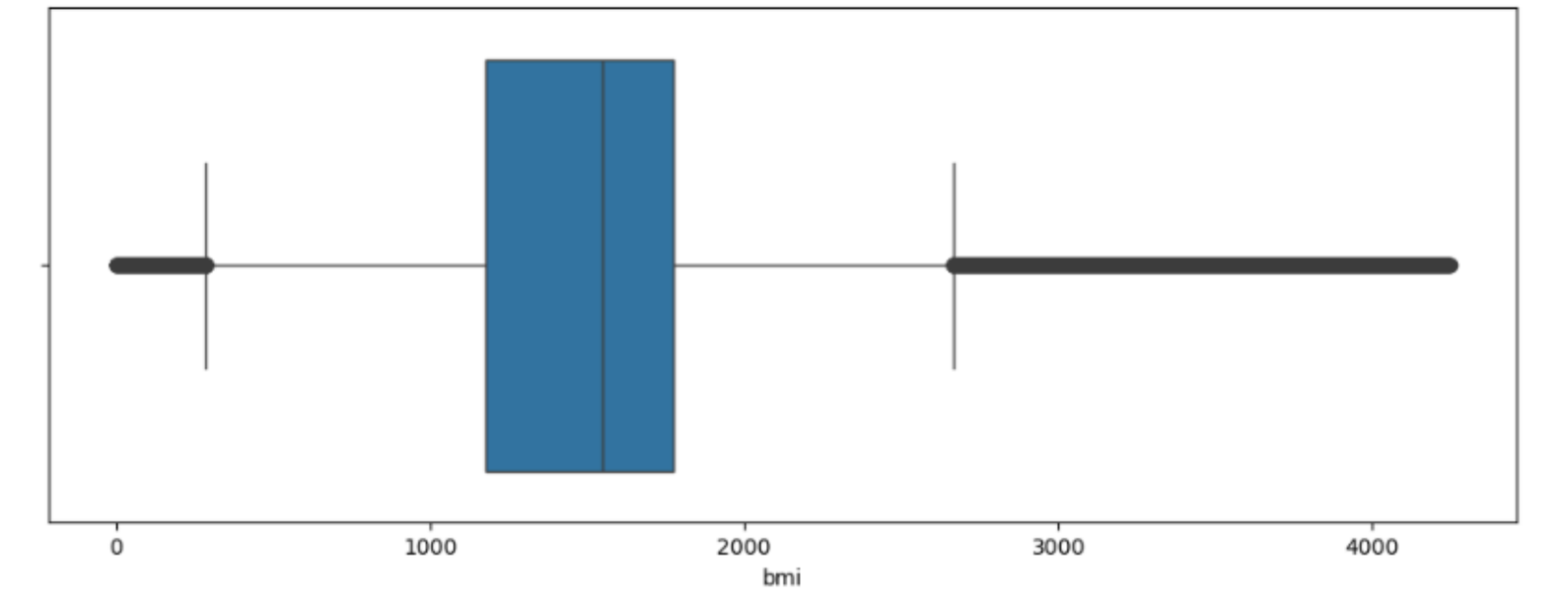


DATA VISUALISATION

DATA DISTRIBUTION



An impressive client satisfaction rate underscores our unwavering commitment to delivering exceptional service and exceeding expectations.



An impressive client satisfaction rate underscores our unwavering commitment to delivering exceptional service and exceeding expectations.

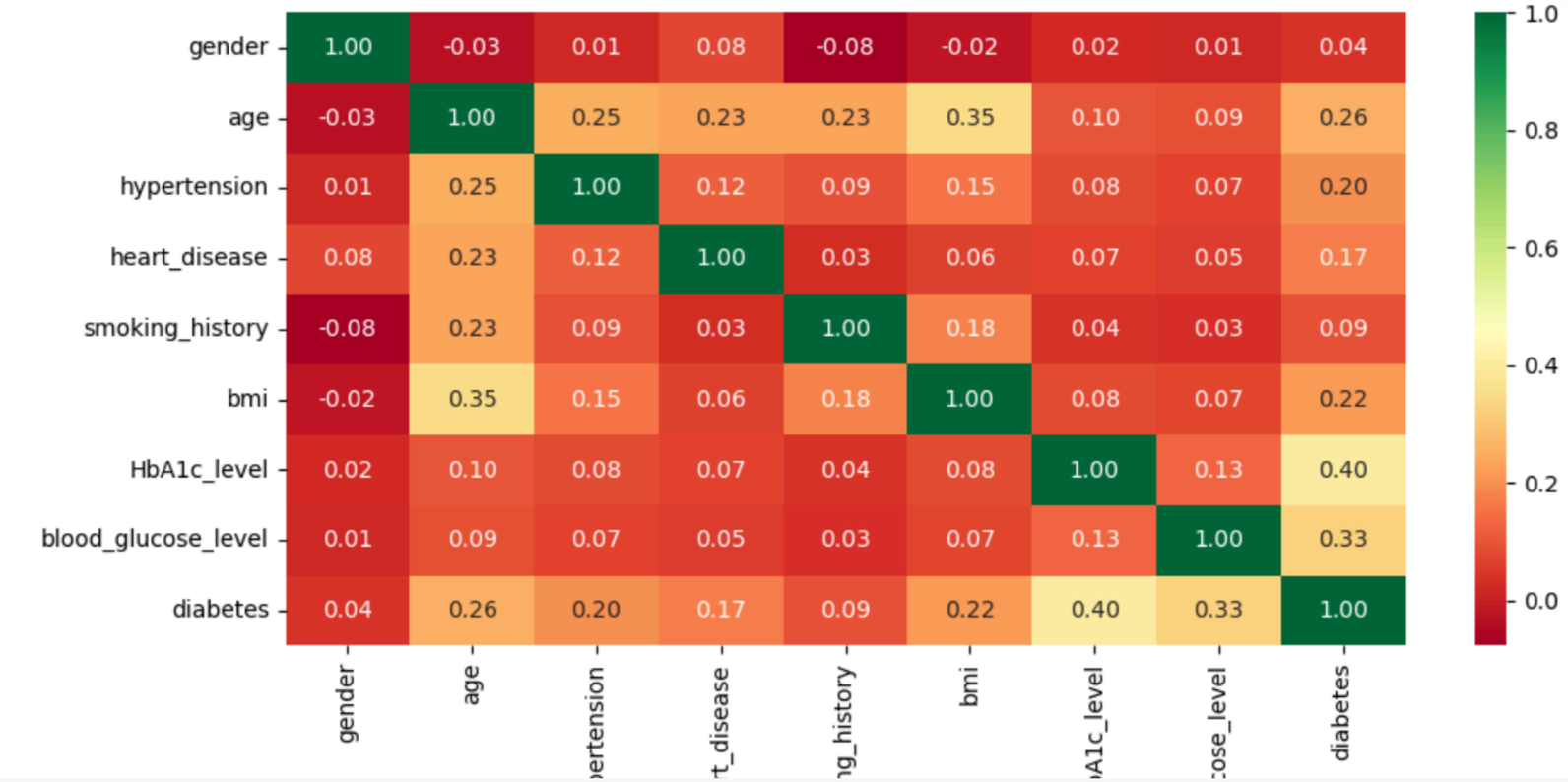
DATA VISUALISATION

PAIR PLOT

The pair plot shows the distribution of diabetes

DATA VISUALISATION

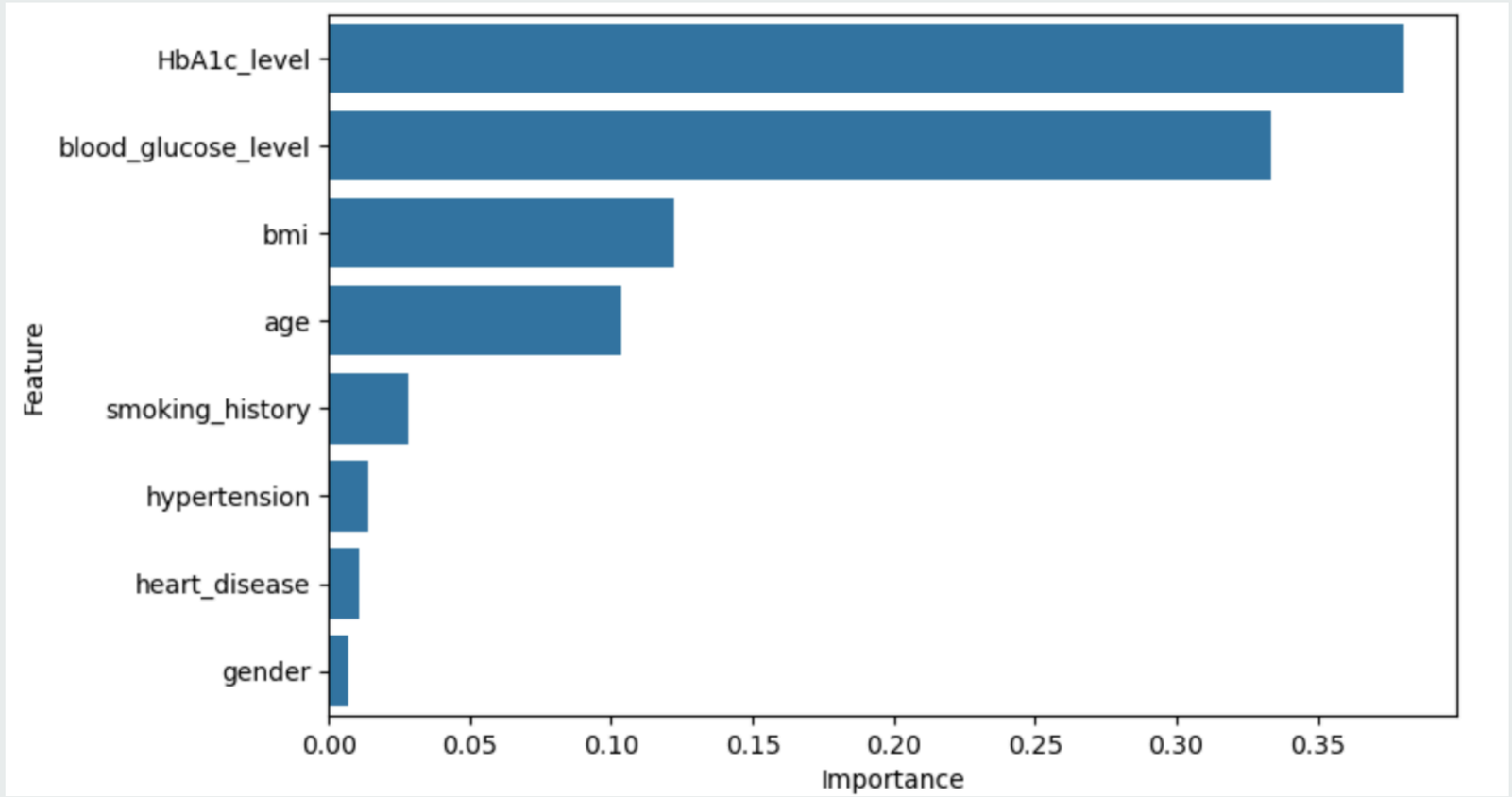
DATA CORRELATION



The Features of high correlation are HbAIC Level to diabetes, Glucose Level to diabetes

DATA VISUALISATION

FEATURE IMPORTANCE



HbAIC Level has the most relevant impact on patients having diabetes while gender has the least impact

MODELING

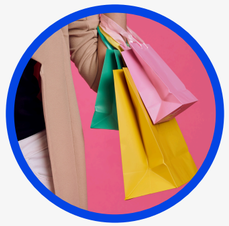
XGB CLASSIFIER

DECISION TREE

LOGISTIC REGRESSION

RANDOM FOREST

SGD CLASSIFIER

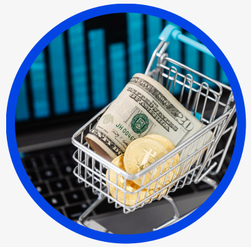


Model Training

Models were trained with dataset 80% Training and 20% Testing

Class Weight

Class weights (0:1, 1:4) were used for training to improve the model.



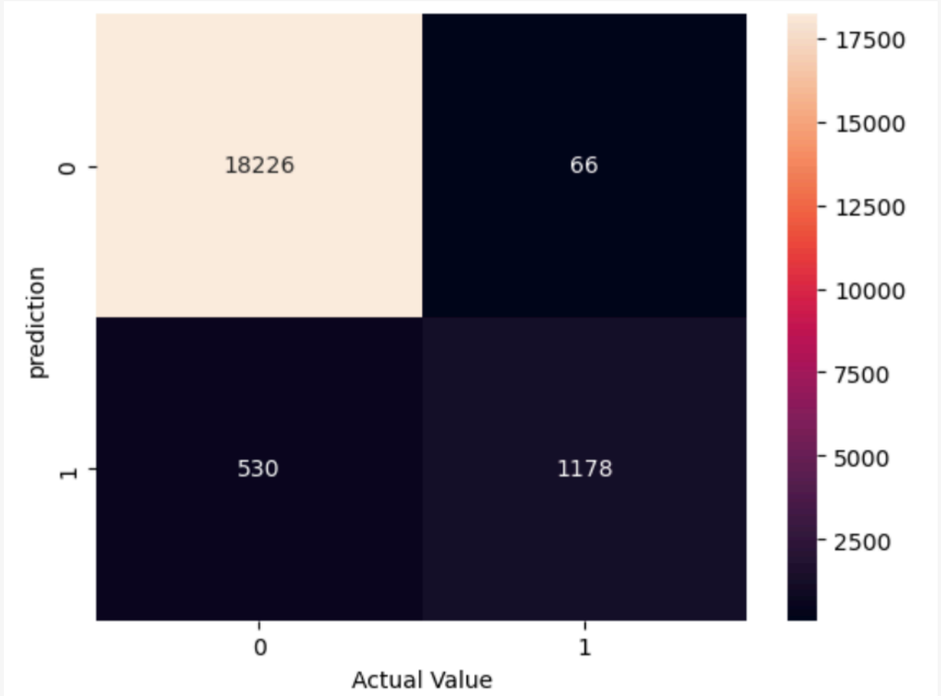
Grid SearchCV

To get the best model Grid SearchCV was to train the model

EVALUATION METRICS

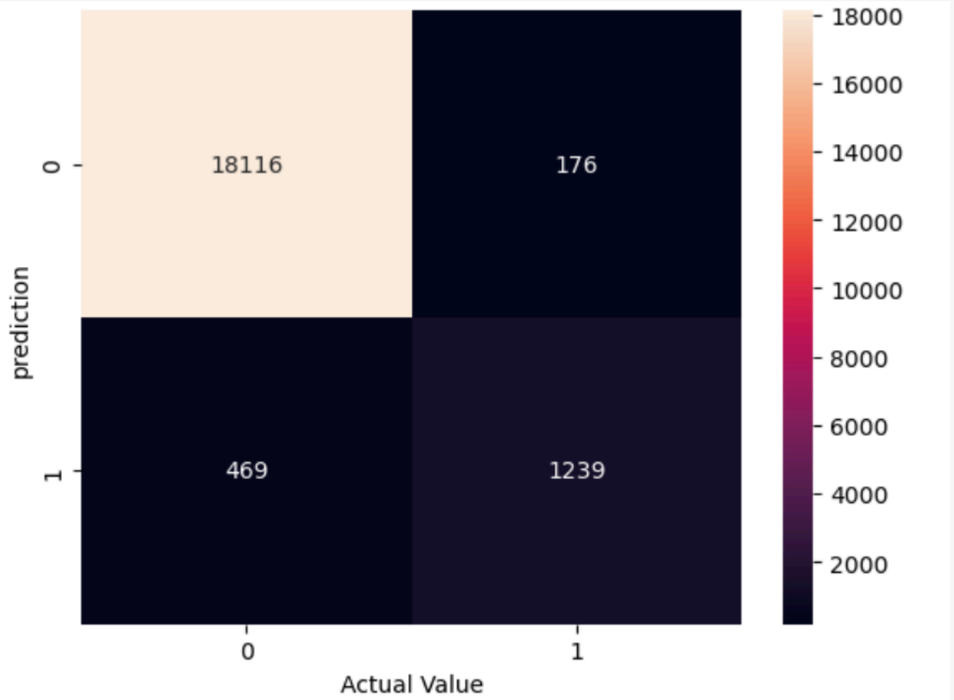
Random Forest

Train model with Class weights 1:4



	precision	recall	f1-score	support
0	0.97	1.00	0.98	18292
1	0.95	0.69	0.80	1708
accuracy			0.97	20000
macro avg	0.96	0.84	0.89	20000
weighted avg	0.97	0.97	0.97	20000

Grid SearchCV Class weight 1:4



	precision	recall	f1-score	support
0	0.97	0.99	0.98	18292
1	0.88	0.73	0.79	1708
accuracy			0.97	20000
macro avg	0.93	0.86	0.89	20000
weighted avg	0.97	0.97	0.97	20000

THANK YOU