Diabetes Prediction Analysis

Presented by : Team 12

Dataset

- This dataset comprises medical and demographic information from patients, including their diabetes status (positive or negative).
- Dataset Size: 100,000 data points. (Imbalanced)
- Data Distribution: Training set (80%), Testing set (10%), and Validation set (10%). (performed by us)
- Key features of the dataset:
  - Age
  - Gender
  - Body Mass
  - Hypertension
  - Diabetes
  - Heart Disease
  - Smoking History
  - HbA1c Level
  - Blood Glucose Level

Research Questions

Primary Question :

- How can identifying the most relevant features from a dataset contribute to the accuracy of diagnosing diabetes?

Secondary Questions :

- How does lifestyle factors, such as smoking history, interact with physiological indicators like BMI and blood glucose levels in predicting diabetes ? Is there a correlation between hypertension and diabetes?
- Does a history of heart disease increase the risk of diabetes?
- How do demographic factors (gender, age) interact with other variables in predicting diabetes risk?
- Does the presence of hypertension or heart disease exacerbate the effects of other risk factors on diabetes incidence?

Model Design

DatasetSelection (source: Kaggle)

01

02

Data Preprocessing (Data Cleaning and Transformation, Split of Data)

Exploratory Data Analysis (EDA)

03

04

Feature Insights (Using Pearson Correlation Matrix)

Model Training (supervised learning)

05

06

Model Evaluation (Confusion Matrix, RMSE, F1-score, precision)

Feature Importance

07

Data Preparation

1.Data Cleaning

- Removing null values
- Removing duplicate data points
- Removing unrelated Values (‘Other’ value of Gender column, Age with value 0)

2. Data Transformation

- Performing Normalization of data
- One hot encoding for Categorical Column(Smoking_history, Gender)

3. Split of Data

- Train (80%), Test (10%) and Validation (10%)

4. Feature Selection

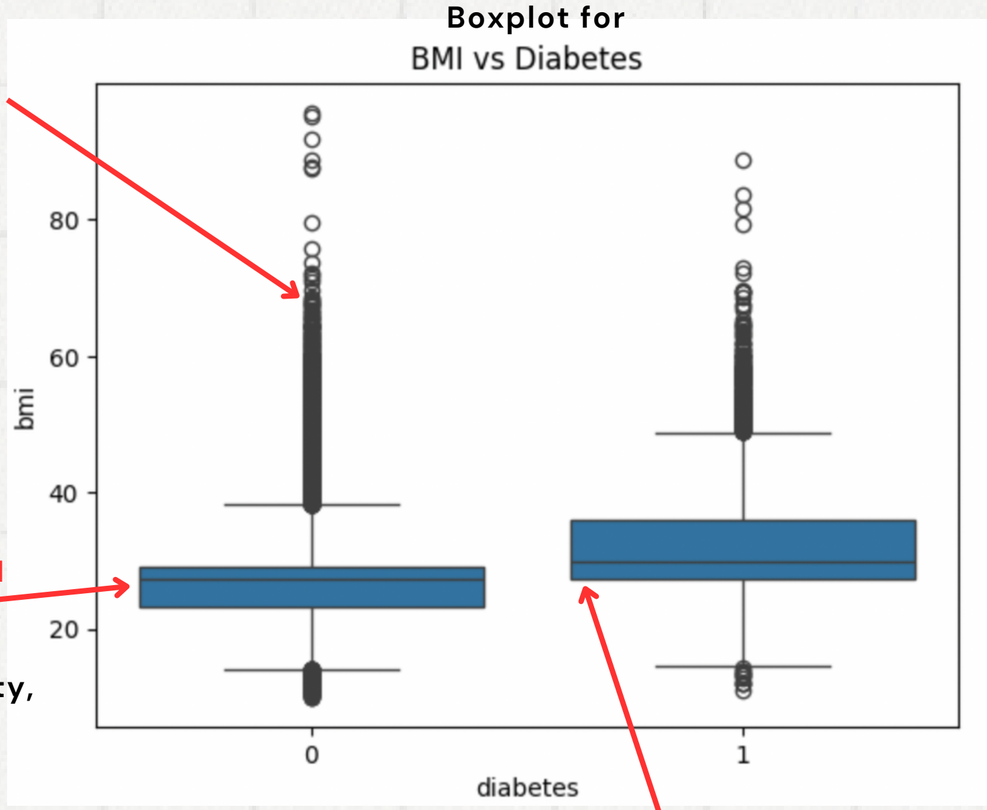
- Finding correlation of columns using Pearson Correlation Matrix

Exploratory Data Analysis

Secondary Question:

1. How does BMI (Body Mass Index) affect the likelihood of having diabetes? Are there differences in diabetes prevalence based on demographic factors such as gender and age, and does age and gender influence the likelihood of having diabetes?

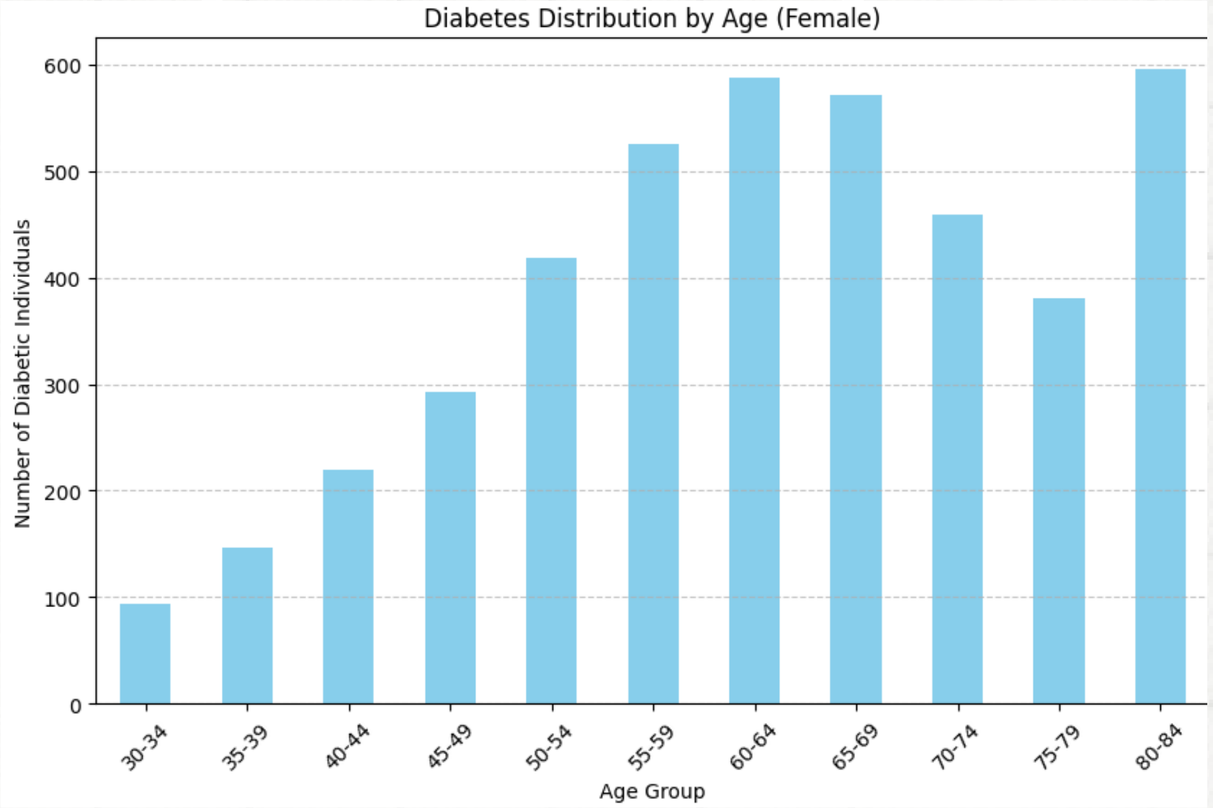
Answer: Part one of this question can be infered by observing the results of the following plot, shows that, the People with higher BMI are at higher risk of Diabetes.

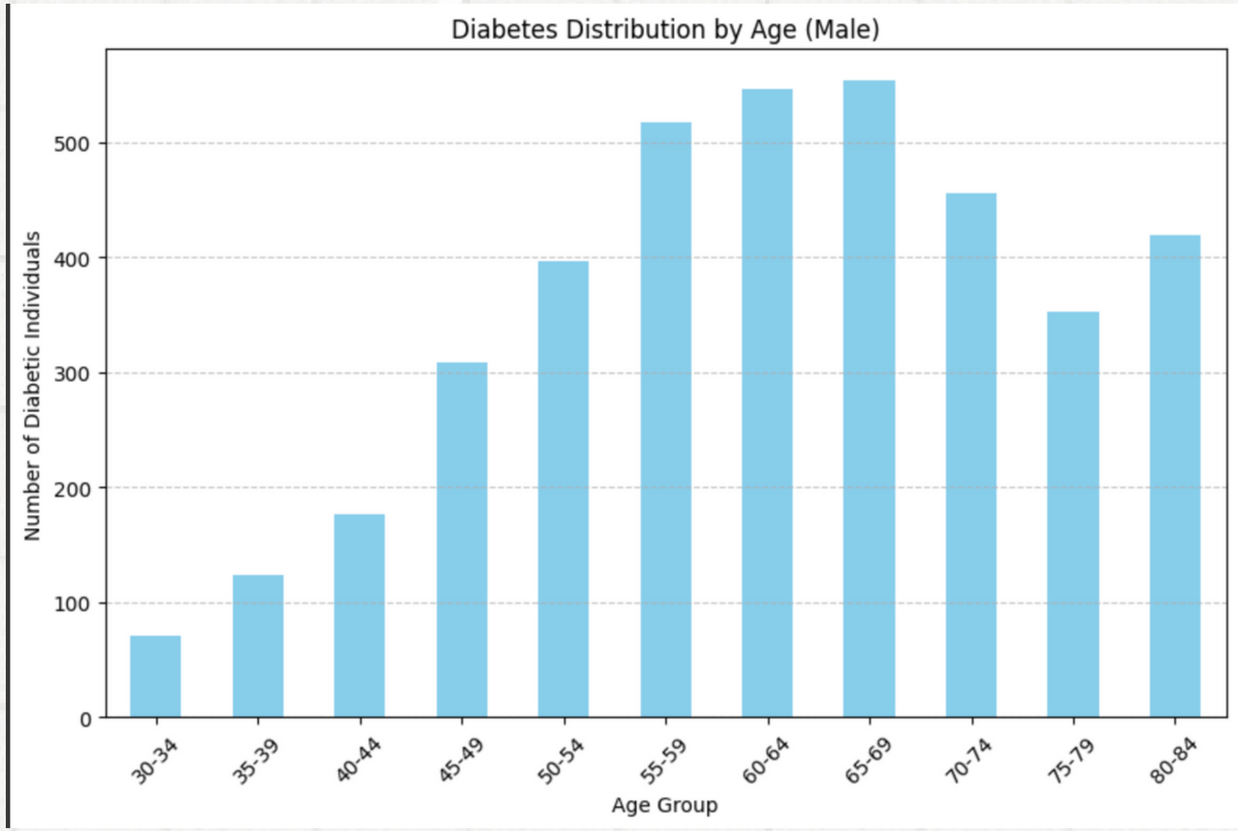


Elevated BMI levels in non-diabetic individuals (i.e above the median) may suggest overweight or obesity, which are known risk factors for developing type 2 diabetes.

BMI levels below the median for the diabetic category suggest that this subset of diabetic individuals has relatively lower BMI levels compared to the median.
Lower BMI levels in diabetic individuals may indicate better weight management or control among diabetic individuals.

Exploratory Data Analysis

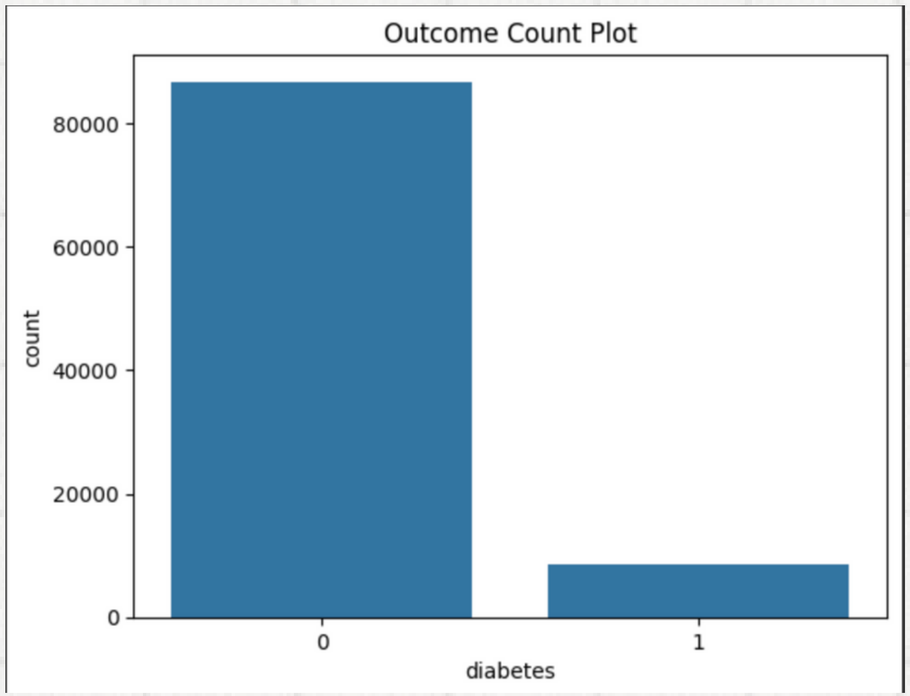




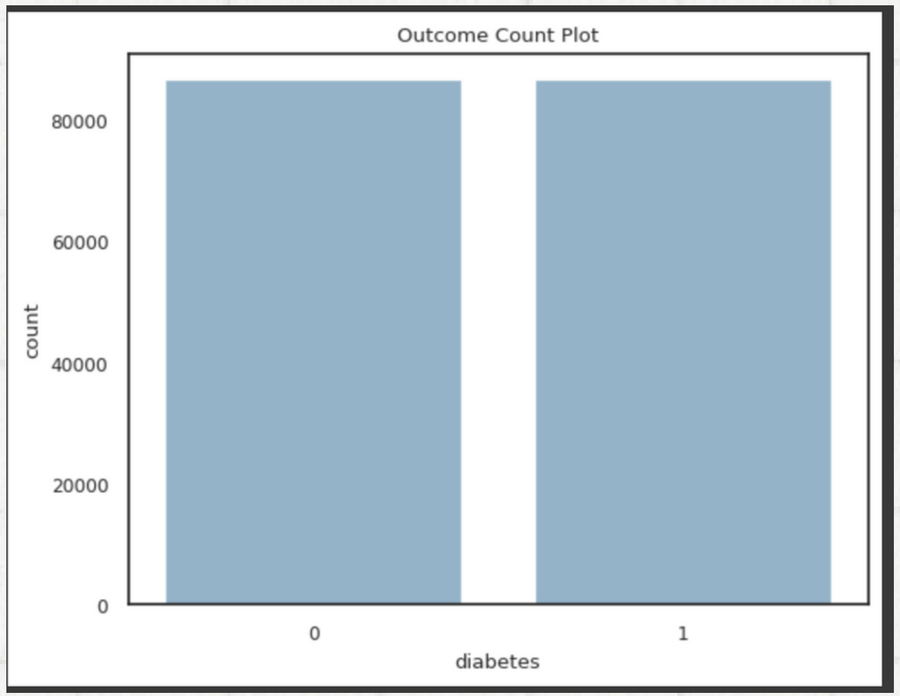
Secondary Question:

2. How do demographic factors (gender, age) interact with other variables in predicting diabetes risk?
Answer: The part two of the previous secondary research question and the above secondary research question can be answered by observing the above EDA, which shows that Male and Female are at equal risk of Diabetes but the age plays a vital role where individuals irrespective of their gender above the age of 40 are at a higher risk.

Class Imbalance



Before performing SMOTE



After performing SMOTE

Model Implementation

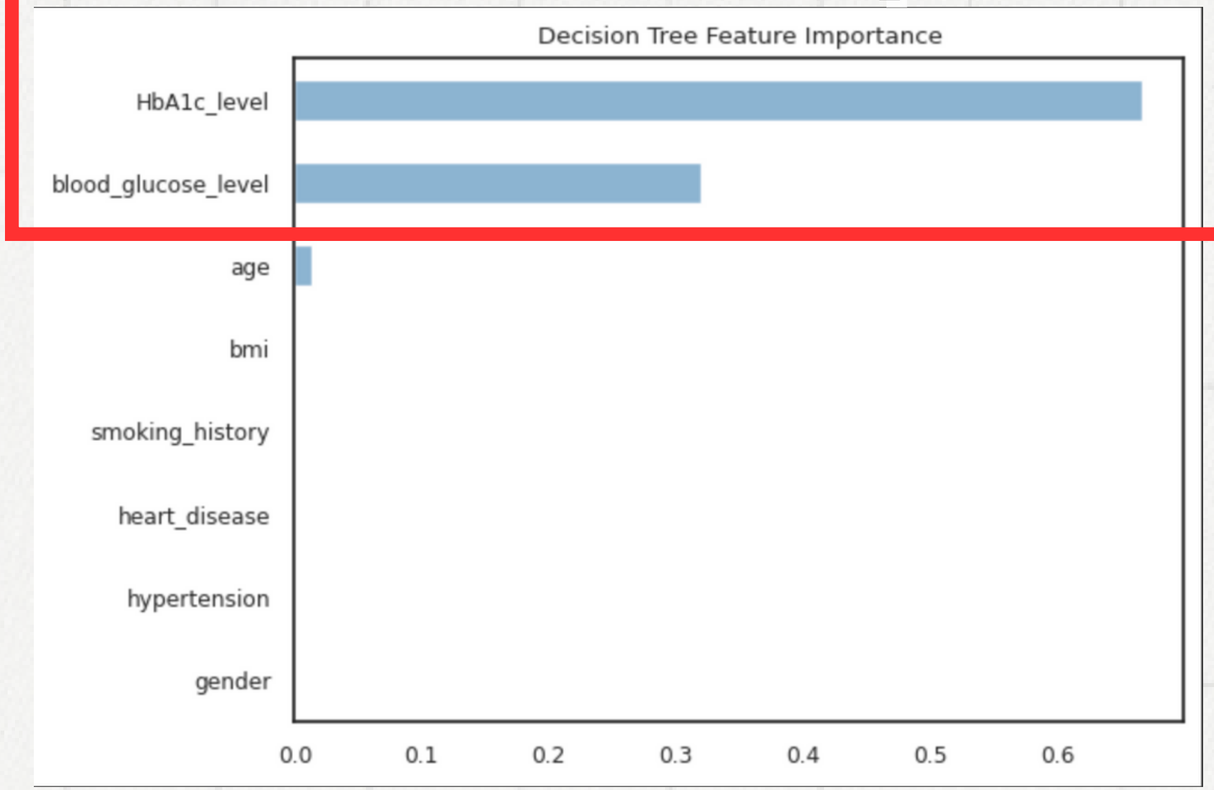
Logistic Regression

- Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set.
- Used Sklearn to implement Logistic Regression.
- Training, Validation, Test Ratio - (80:10:10)
- Evaluation Metrics implemented for model comparison and performance
  1. Classification Report to generate F1-Score
  2. Precision
  3. Recall
  4. Confusion Matrix
  5. RMSE
- Carried out for imbalanced dataset as well on balanced dataset generated using SMOTE

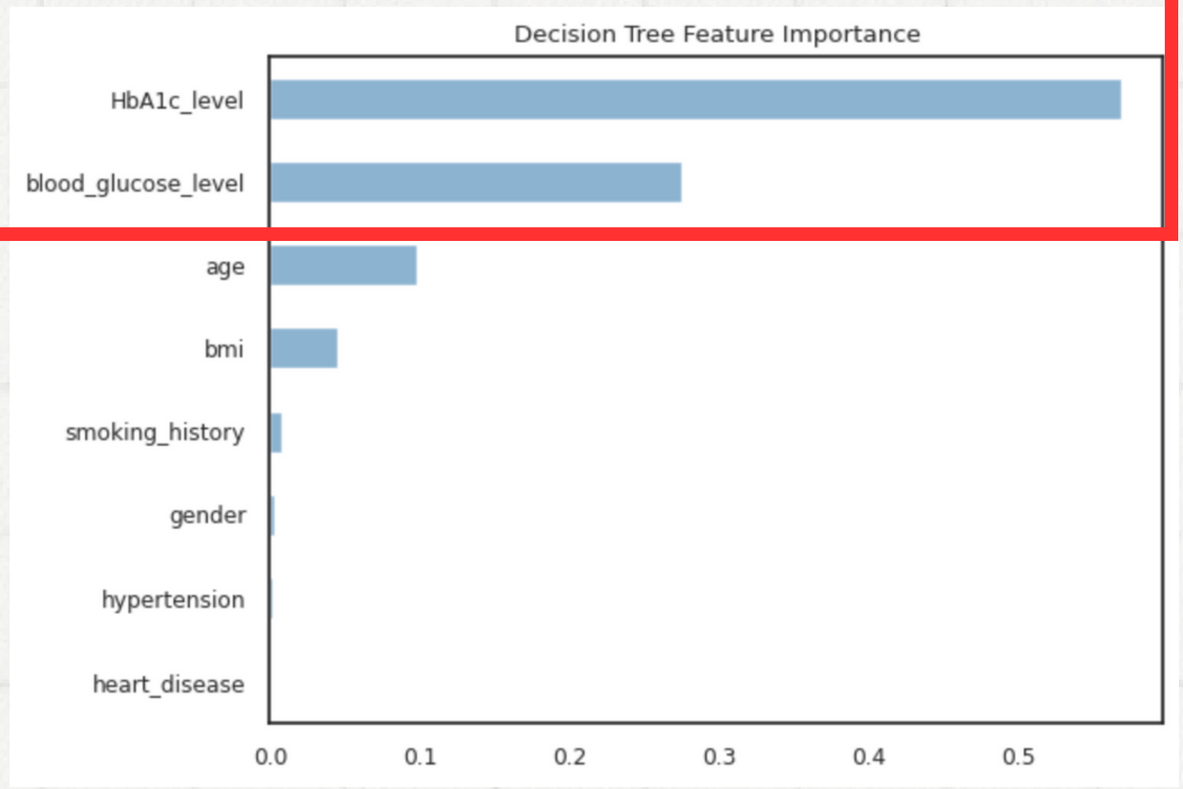
Decision Tree

- Used Decision Tree to show Feature Selection
- Used Sklearn to implement Decision Tree.
- Training, Validation, Test Ratio - (80:10:10)
- Implemented GridSearchCV using training and validation data for hyperparameter tuning
  1. Max_depth : 3
  2. Min_Sample_leaf : 2
  3. Min_Sample Split: 5
- Evaluation Metrics implemented for model comparison and performance
  1. Classification Report to generate F1-Score
  2. Precision
  3. Recall
  4. Confusion Matrix
  5. RMSE
- Carried out for imbalanced dataset as well on balanced dataset generated using SMOTE

Feature Importance



Feature Importance Before SMOTE



Feature Importance After SMOTE

<u>Primary and Secondary Research Questions:</u>

PQ1. : How can identifying the most relevant features from a dataset contribute to the accuracy of diagnosing diabetes?
Answer: HbA1c and blood_glucose_level are the features which are most relevant from the dataset that contribute to the accuracy of diagnosing diabetes. (accuracy can be observed by comparing F1 scores in the further slides).

2. Does the presence of hypertension affects the diabetes ? 3. Does a history of heart disease increase the risk of diabetes?
Answer: The above feature importance before SMOTE as well as after SMOTE shows that as per the dataset it can be infered that Hypertension and Heart diseases do not contribute to the risk of diabetes.

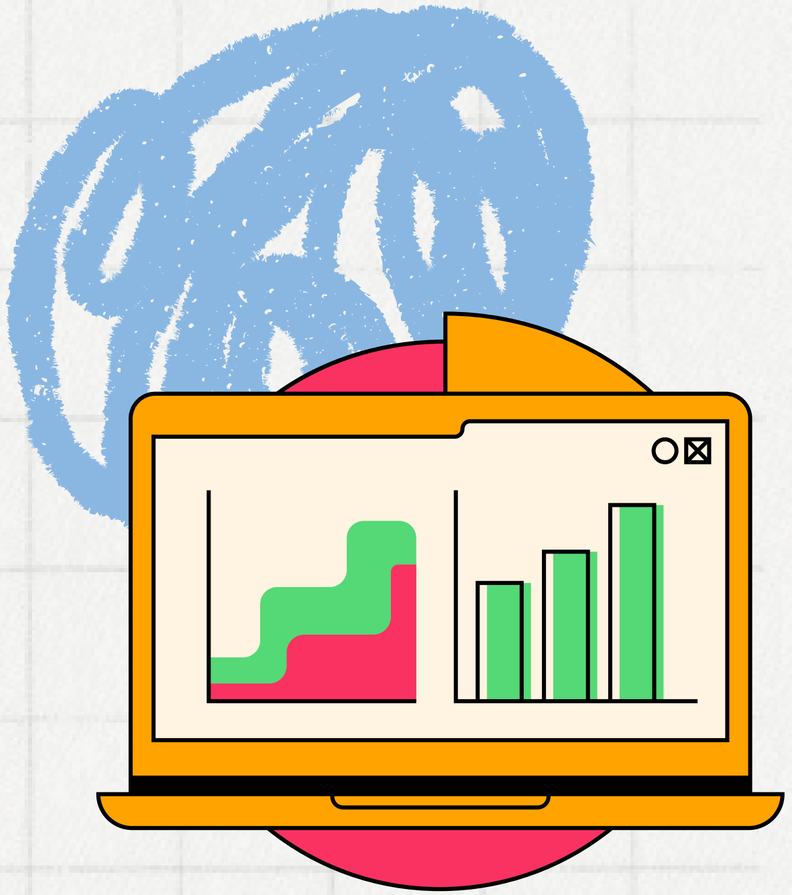
Model Evaluation and Discussion

Logistic Regression

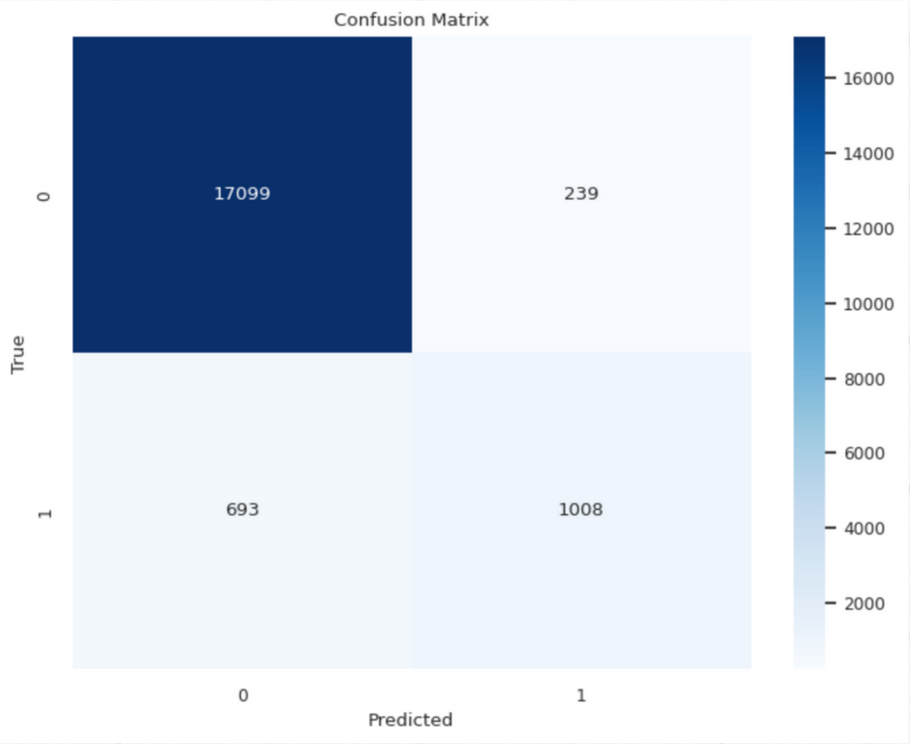
Decision Trees

Logistic Regression after SMOTE

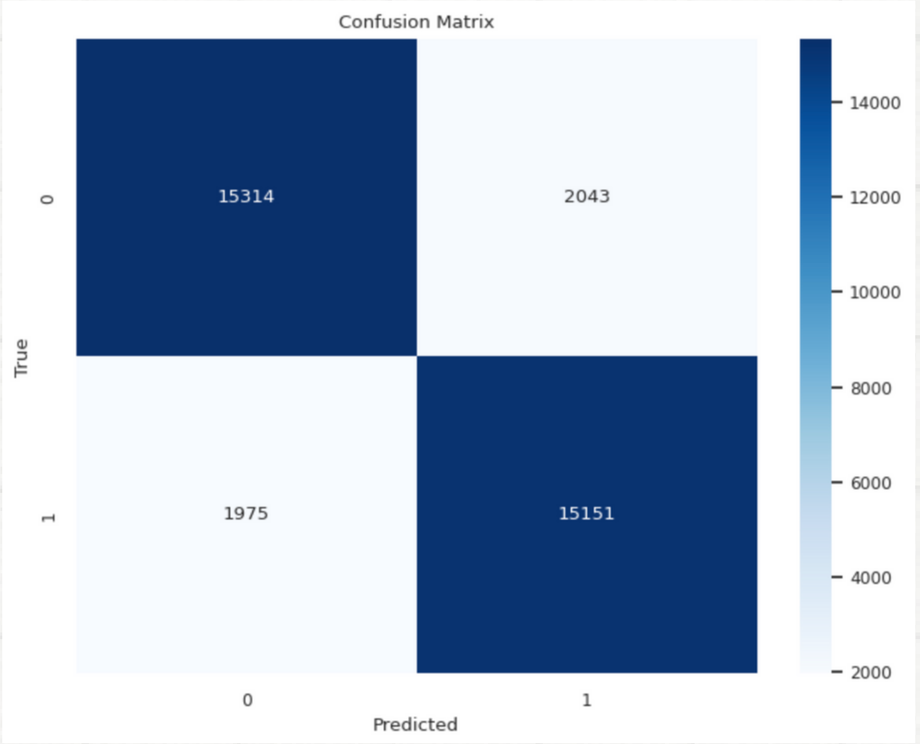
Decision Trees after SMOTE



Model Evaluation and Discussion: Logistic Regression



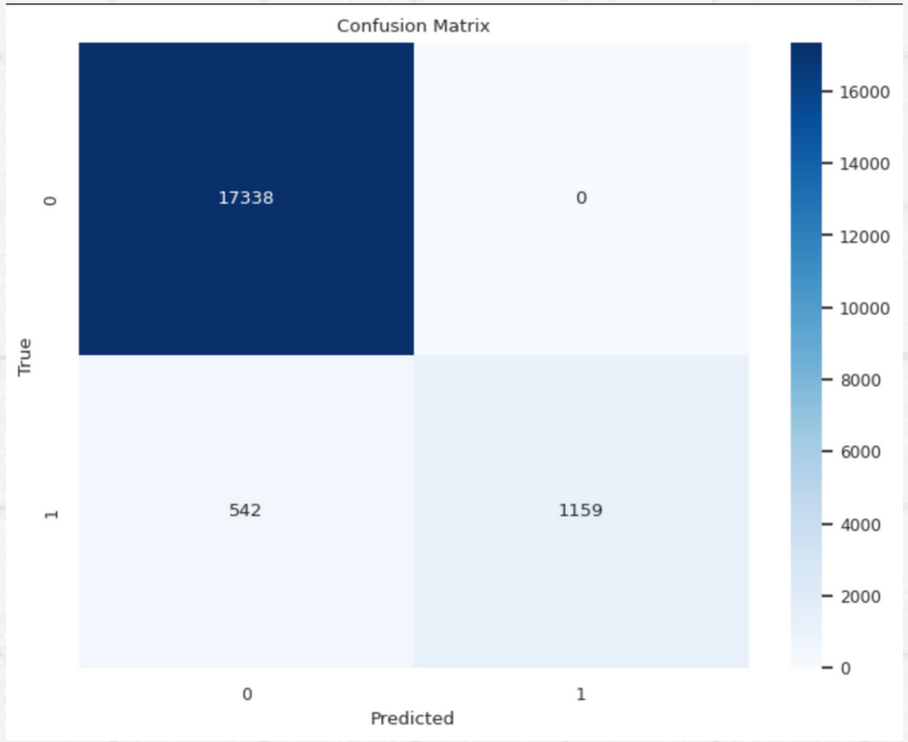
Confusion Matrix before SMOTE



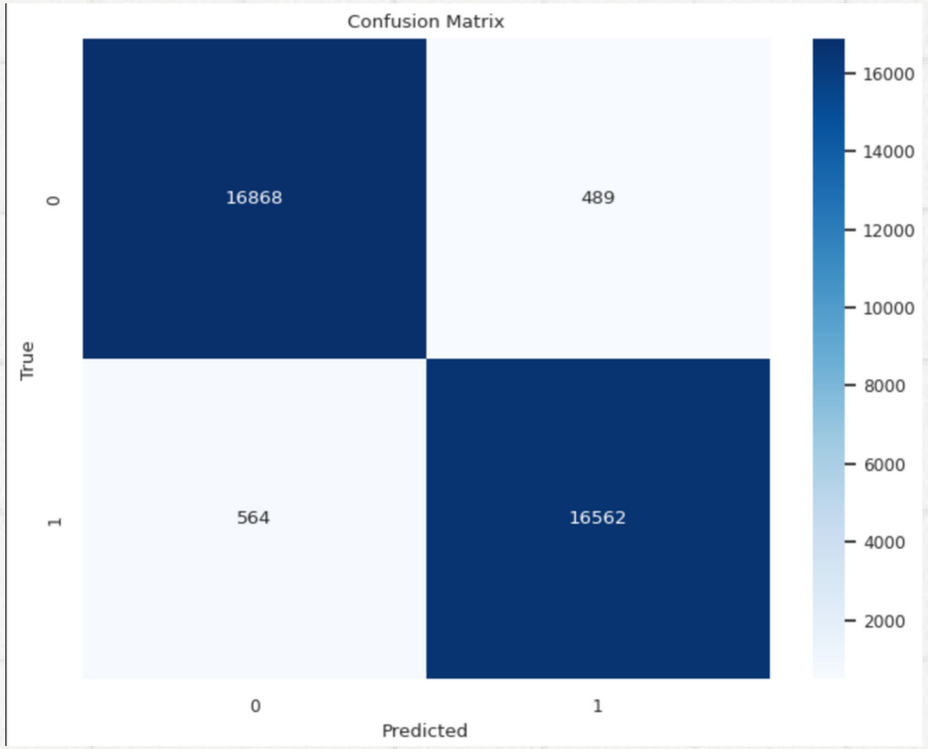
Confusion Matrix after SMOTE

	Before SMOTE		After SMOTE	
Evaluation Metric	**Logistic Regression**	**Decision Tree Classifier**	**Logistic Regression**	**Decision Tree Classifier**
Precision	89%	98%	88%	97%
Recall	79%	84%	88%	97%
F1 Score	83%	90%	88%	97%
RMSE	0.22	0.16	0.34	0.17

Model Evaluation and Discussion: Decison Tree



Confusion Matrix before SMOTE



Confusion Matrix after SMOTE

	Before SMOTE		After SMOTE	
Evaluation Metric	**Logistic Regression**	**Decision Tree Classifier**	**Logistic Regression**	**Decision Tree Classifier**
Precision	89%	98%	88%	97%
Recall	79%	84%	88%	97%
F1 Score	83%	90%	88%	97%
RMSE	0.22	0.16	0.34	0.17

F1 score has increased significantly indicating that the model has improved both in identifying true positives and avoiding false positives

Any Questions?