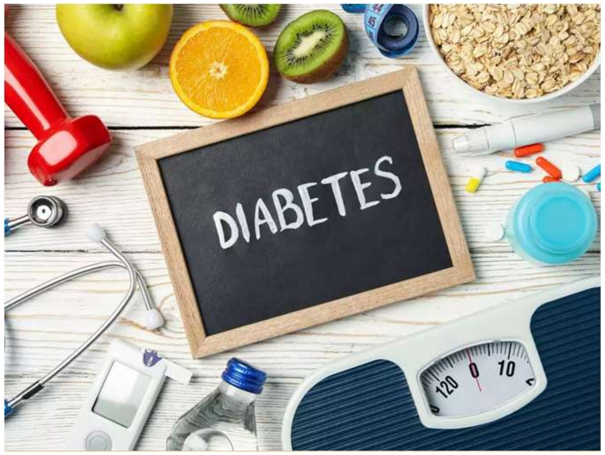
Predictive Model
for
Lifestyle-Diabetes Association



UNIVERSITY of
SOUTH FLORIDA

Data Wranglers

1. Pavan Teja Gupta Allenki

2. Gayatri Mohan Kshirsagar

3. Manvitha Nagandla

4. Pradeep Chand Potturi

5. Sneha Panjala

UNIVERSITY of SOUTH FLORIDA

INTRODUCTION

Diabetes is a growing public health concern in the United States, affecting millions of individuals and imposing a significant economic burden on the healthcare system.

Lifestyle factors, such as age, sex, and mental health, physical health, are known to influence the risk of developing diabetes. However, the intricate interplay between these factors and the prevalence of diabetes remains a complex and insufficiently explored issue.

This data mining project aims to investigate and model the relationship between lifestyle factors and the prevalence of diabetes in the United States.

DATASET OVERVIEW

- Features

HighBp, HighChol, CholCheck, BMI, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthCare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education, Income

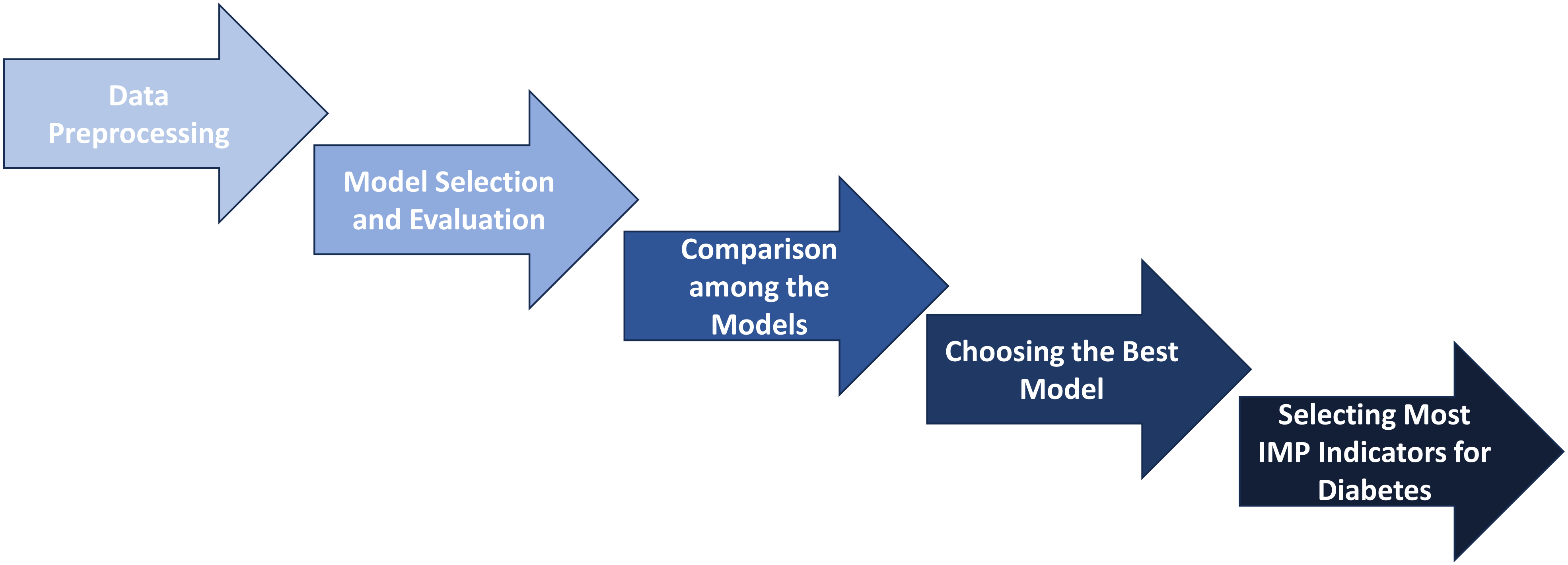
- Target

Diabetes

Number of Instances: 253,680

Number of Features: 21

FLOW



Data Preprocessing

Model Selection and Evaluation

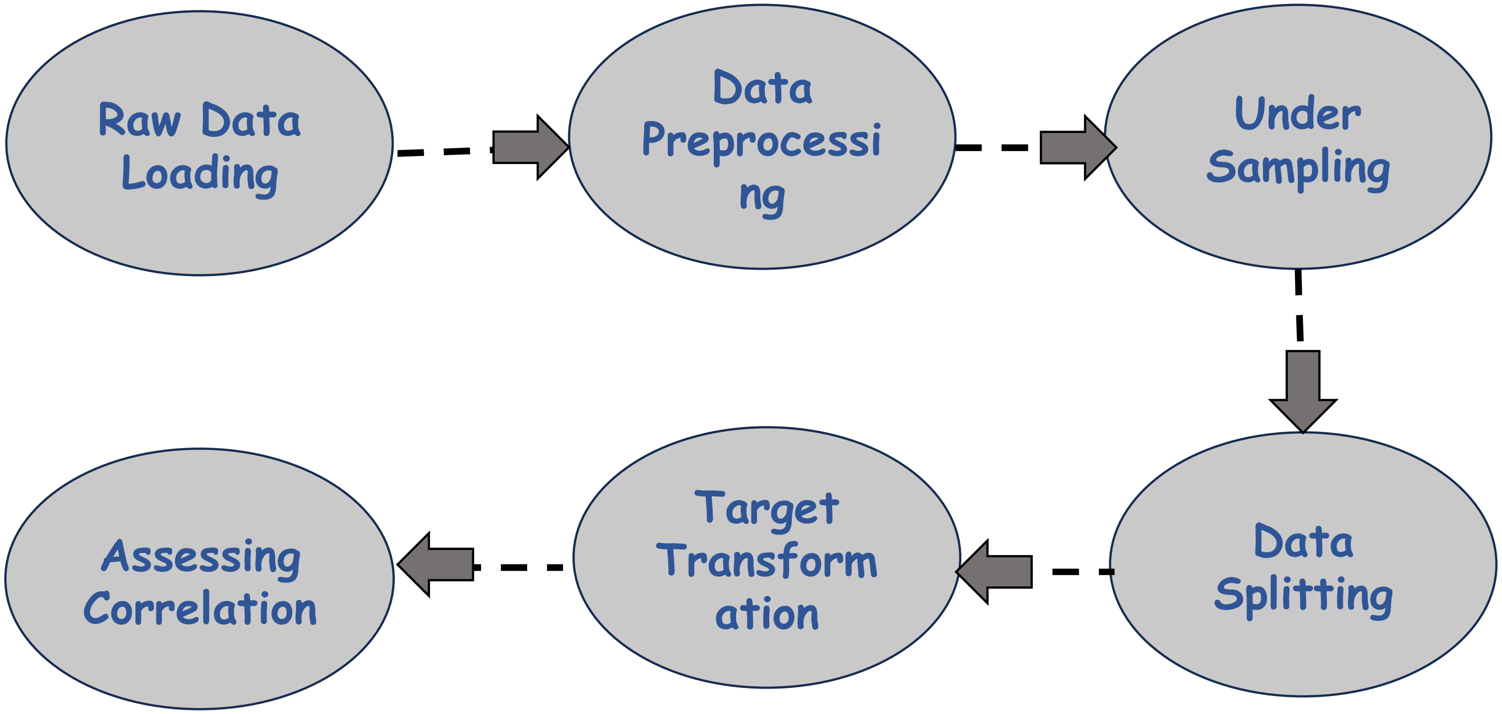
Comparison among the Models

Choosing the Best Model

Selecting Most IMP Indicators for Diabetes

UNIVERSITY of SOUTH FLORIDA

DATA PREPROCESSING



Raw Data Loading

Data Preprocessi ng

Under Sampling

Data Splitting

Target Transform ation

Assessing Correlation

Balancing the data using Under Sampling

Splitting the Data into Train and Test Sets

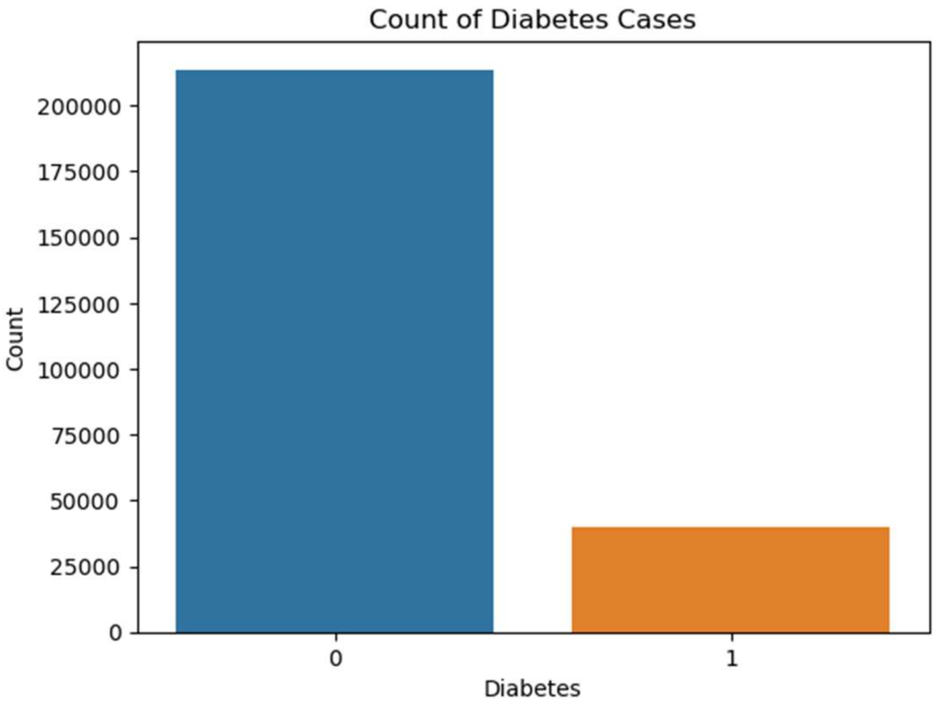
Transforming multiclass “Target” variable into binary class

Assessing target variable correlation with other columns to identify and drop irrelevant features.

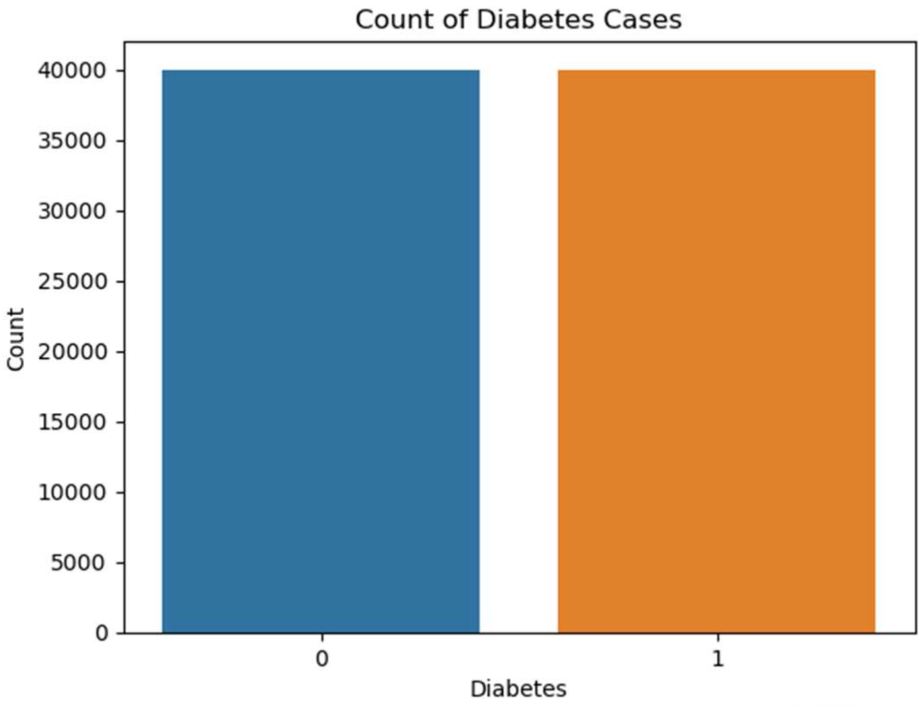
UNIVERSITY of SOUTH FLORIDA

COUNT PLOT OF TARGET

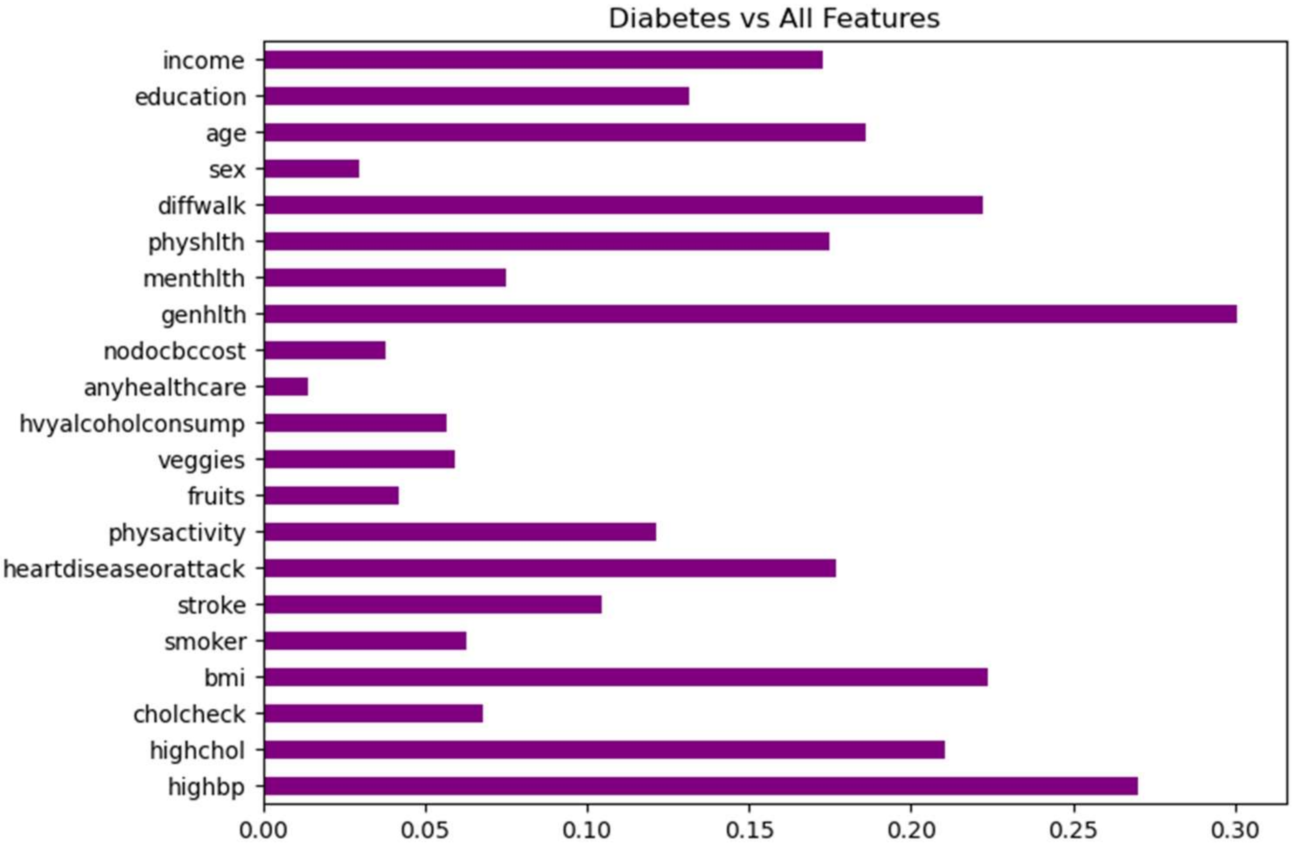
Imbalanced Data



Balanced Data after Under Sampling

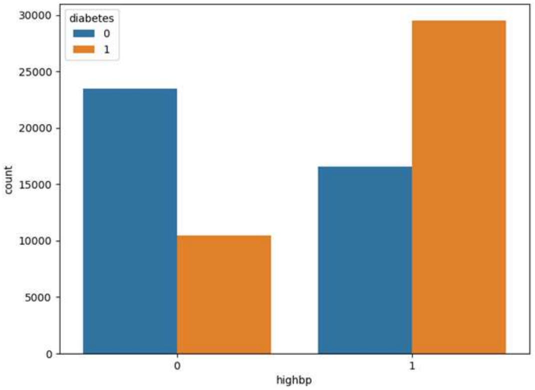


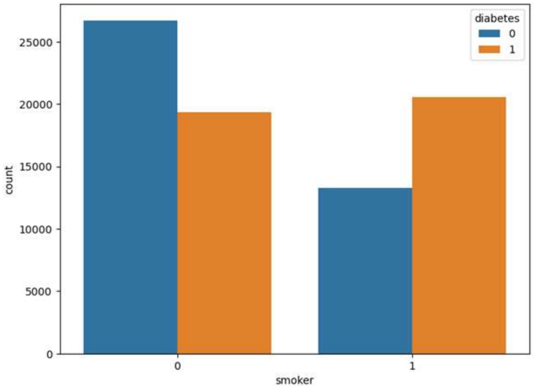
CORRELATION GRAPH



EXPLORATORY DATA ANALYSIS

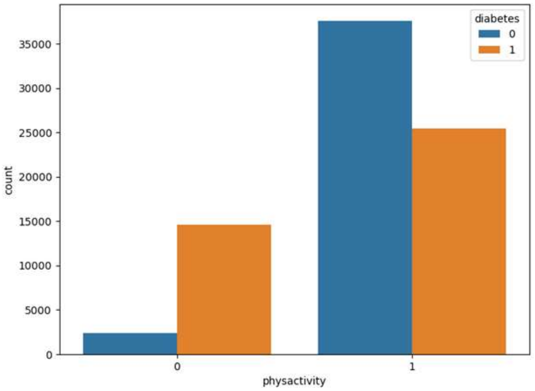
highbp

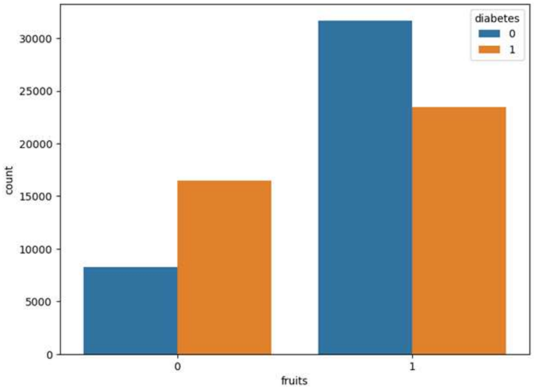




smoker

physactivity

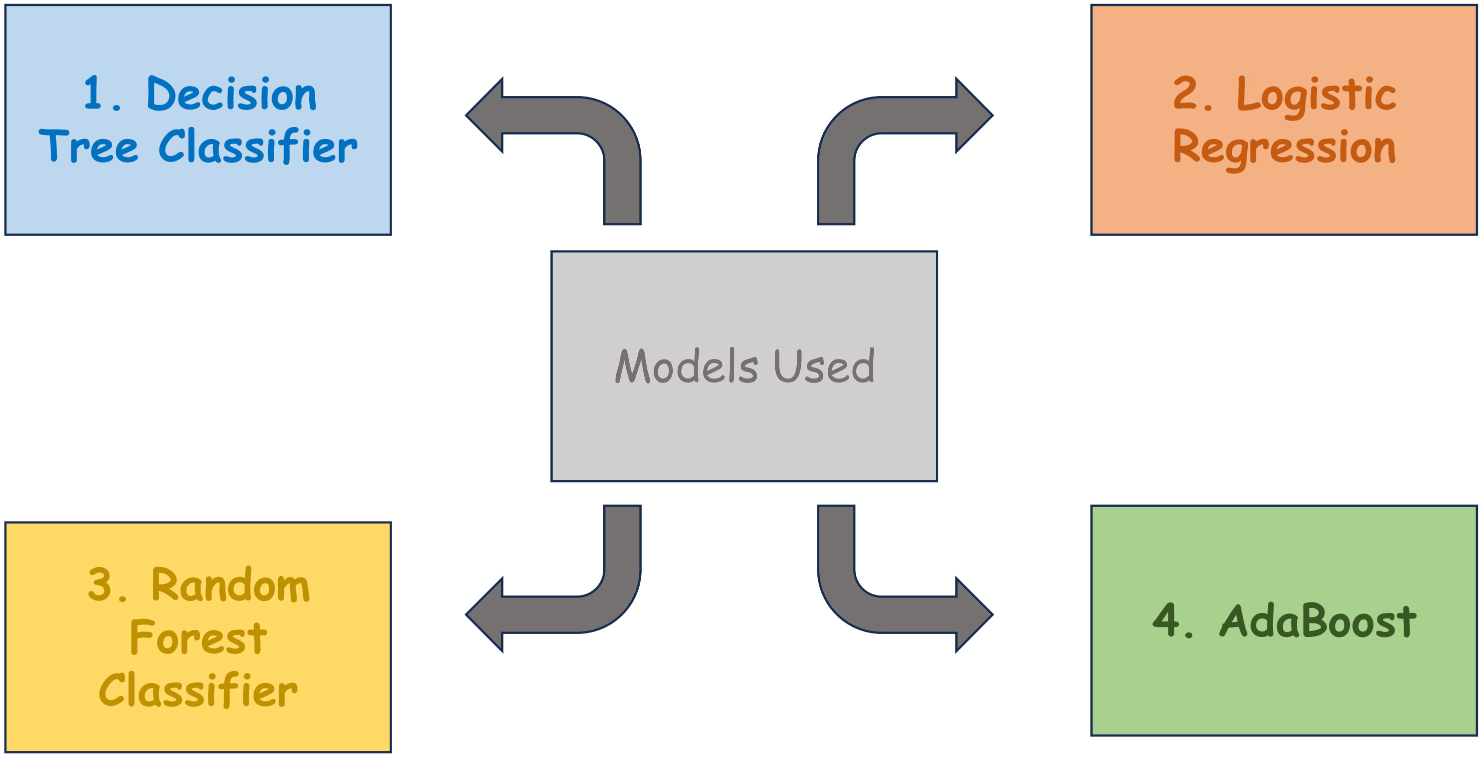




fruits

UNIVERSITY of SOUTH FLORIDA

MODELS EVALUATED



Models Used

1. Decision Tree Classifier

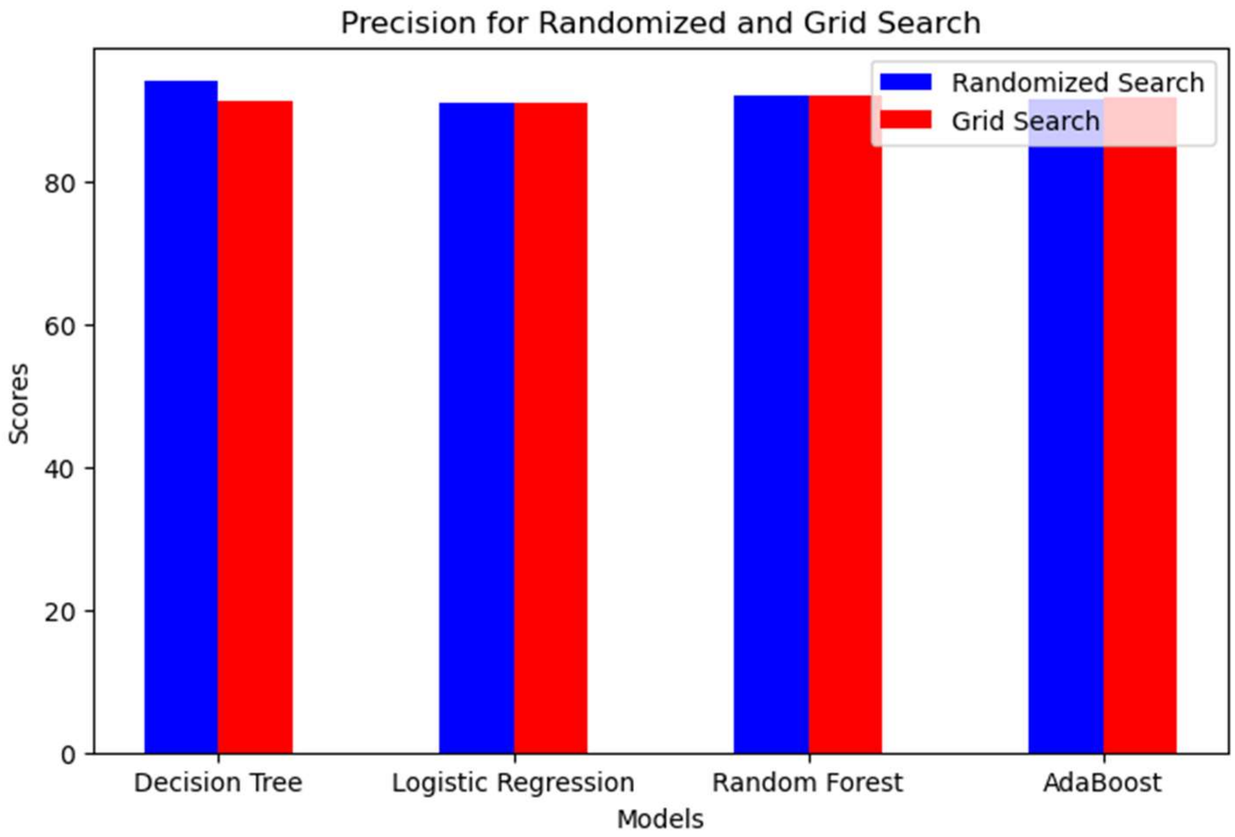
2. Logistic Regression

3. Random Forest Classifier

4. AdaBoost

UNIVERSITY of SOUTH FLORIDA

PRECISION COMPARISON FOR RANDOM AND GRID SEARCH



PERFORMANCE METRICS SELECTION

- High precision ensures accurate positive diagnoses in medical settings.
- It reduces false positives, preventing unnecessary treatments and interventions.
- Precision fosters trust in medical decisions and healthcare providers.
- Ultimately, it improves patient outcomes by reducing the risk of incorrect diagnoses and unnecessary procedures.

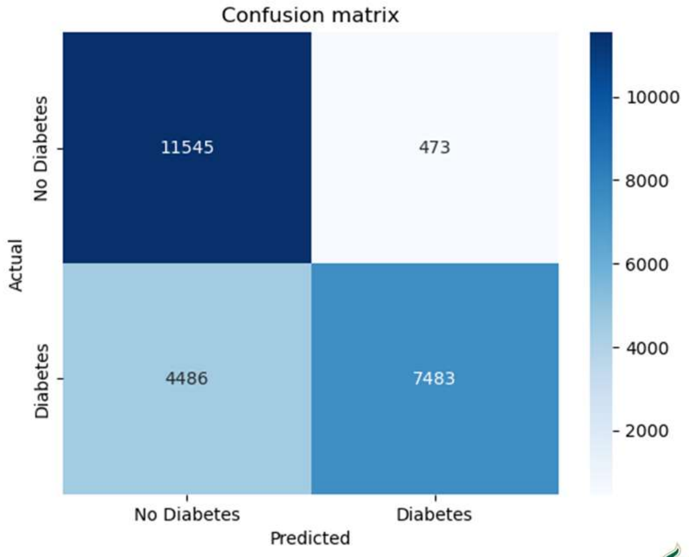
RESULTS

- **Decision Tree Classifier after using Randomized search**

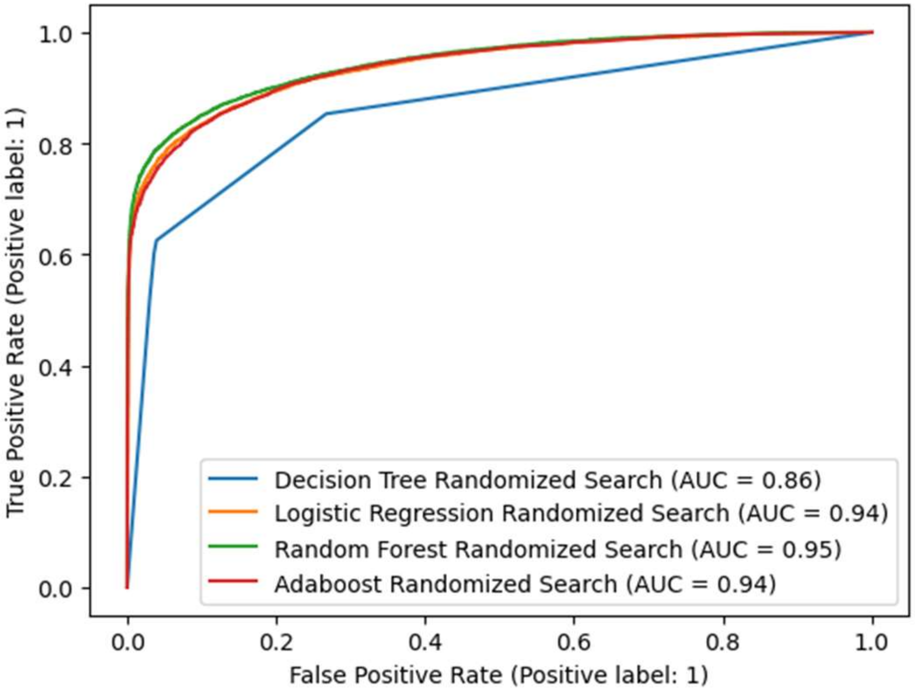
✓ Precision Score: 0.94

✓ Confusion Matrix

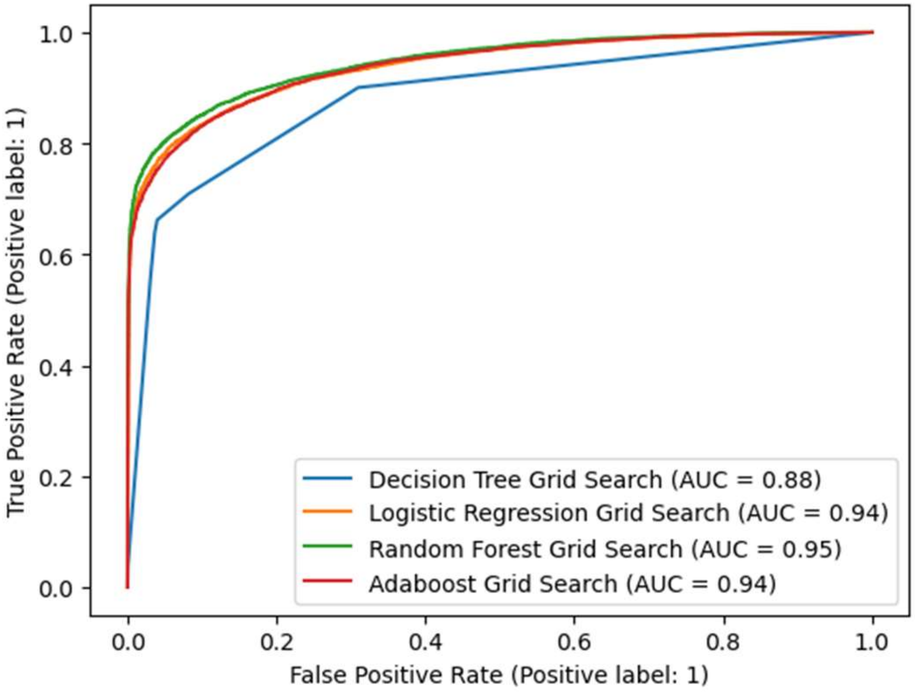
- ❑ 11545 : True Negative
- ❑ 473: False Positive
- ❑ 4486: False Negative
- ❑ 7483: True Positive



ROC CURVES FOR MODELS WITH RANDOMIZED AND GRID SEARCH



Randomized Search



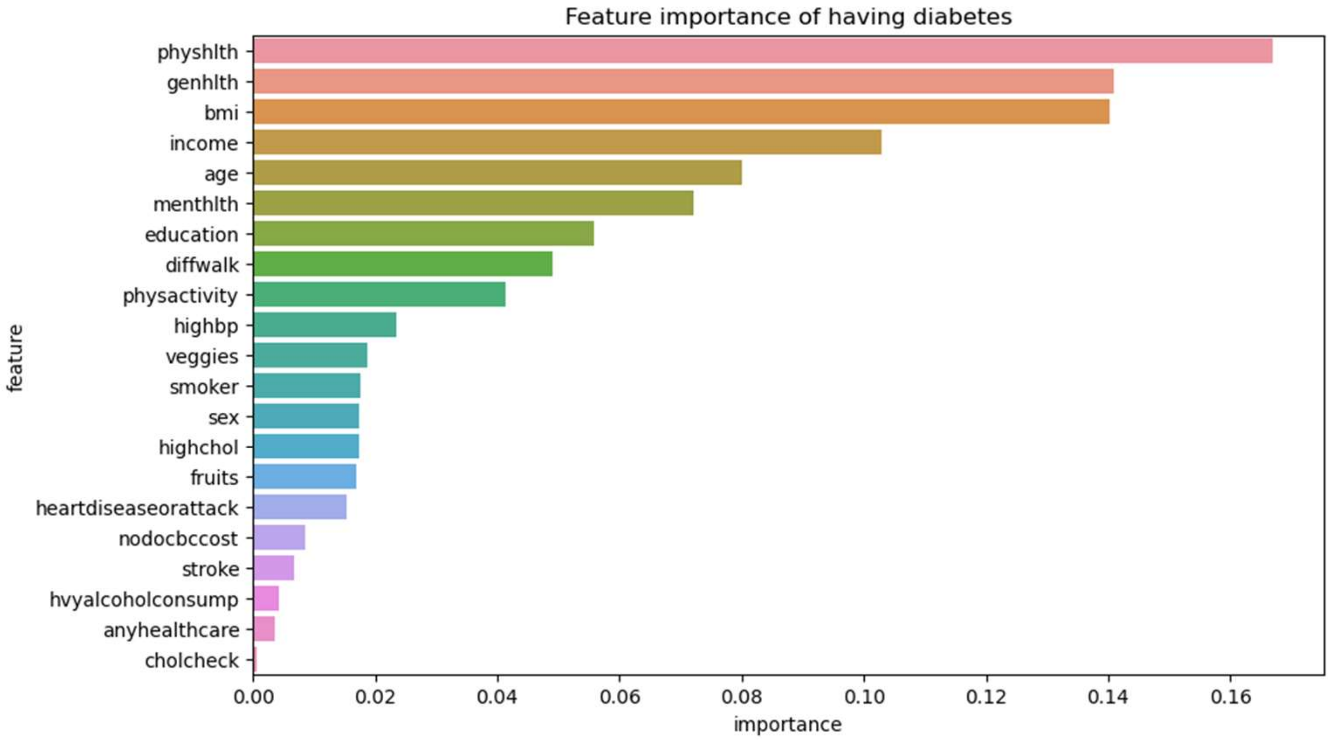
Grid Search

MODEL RECOMMENDATION

- We selected the Random Forest Classifier as our best model based on high precision and a strong ROC curve, indicating its ability to make accurate positive predictions and excellent overall model performance.
- The second-best model in contention is AdaBoost, which achieved a precision of 91.70%. Additionally, it showed a high AUC (Area Under the Curve) score of 94%, highlighting its capability to distinguish between classes effectively.
- Our choice of Random Forest is justified by its robust precision and ROC performance, making it a reliable option for tasks where accurate positive diagnoses are crucial, such as medical diagnosis.
- AdaBoost's strong precision and AUC also make it a valuable alternative for applications where a slightly different trade-off between precision and overall classification performance is acceptable.

MOST IMPORTANT INDICATORS FOR DIABETES

- Physical Health
- General Health
- BMI
- Income
- Age
- Mental Health
- Education
- Difficulty Walking
- Physical Activity



UNIVERSITY of SOUTH FLORIDA

Thank You