Diabetes Prevention

Eric Tysinger & Leo Zheng

Problem Statement

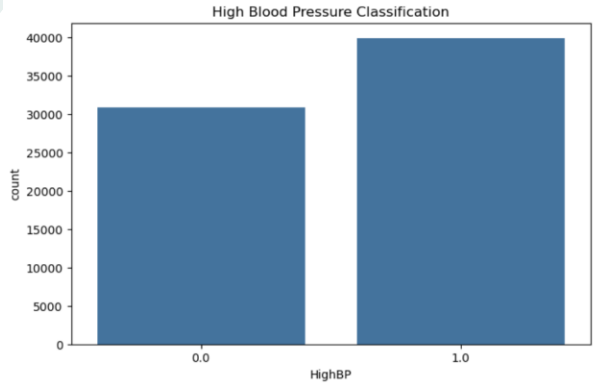
- Diabetes is a chronic health condition that affects how the body turns food into energy. It is a growing public health concern, with millions of individuals affected worldwide. Early detection and prevention are key to managing this condition and reducing its long-term health consequences. Through predictive modeling and machine learning, we aim to identify significant indicators of Diabetes for individuals. The ultimate goal is to provide insights that support public health initiatives and promote healthier lifestyles.

Dataset

- We sourced our dataset from the **UCI Machine Learning Repository**
- The dataset originated from a 2014 survey conducted by the U.S. Center for Disease Control (**CDC**)
- The dataset contains **70692 rows** and **22 columns**
- There is 1 target variable (**Diabetes_binary**), and **21 independent variables**

Methodology

- Data Collection and Preparation:
  - Data pre-processing:
    - Imported the data, observed its structure, and computed summary statistics
    - Checked for missing values
  - Exploratory Data Analysis:
    - Created histograms for univariate analysis
    - Created confusion matrices for bivariate analysis
    - Created a heatmap for multivariate analysis
    - Checked for missing values



Methodology contd.

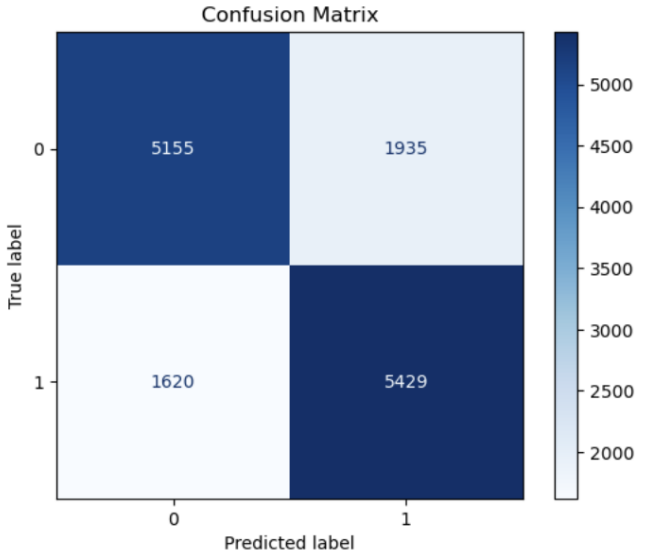
- Models Tested:
  - Logistic Regression:
    - We trained the model using the feature in the dataset (includes BMI, physical health, general health indicators), then evaluated its accuracy, precision, recall, F1 score, and ROC AUC.
  - Decision Tree:
    - Using the same features as the logistic regression model, we created a decision tree with 4 nodes (High Blood Pressure (Yes or No), General Health (1-5 scale, 1 being the best) Age, BMI, and High Cholesterol)
  - Random Forest:
    - We trained the model using all the features and created a feature importance graph to supplement our findings

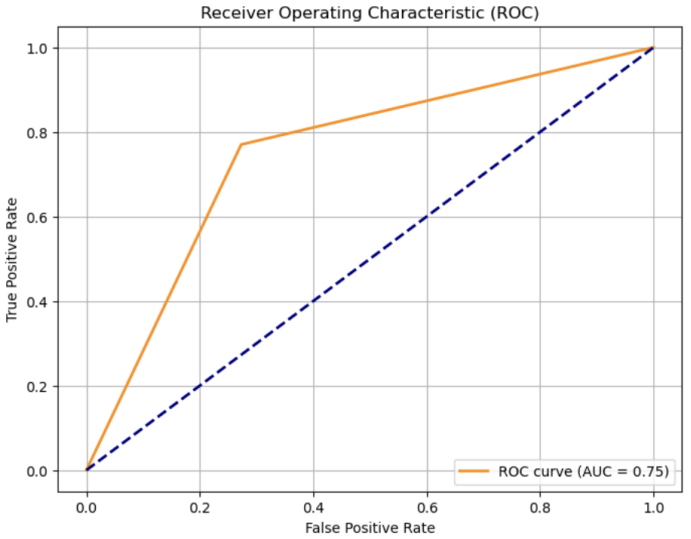
Logistic Regression

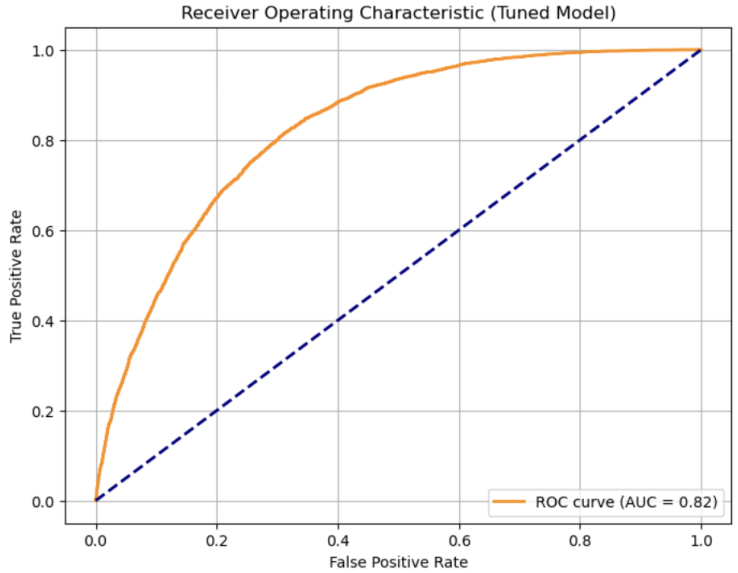
We trained the model using the features in the dataset (includes BMI, physical health, general health indicators), then evaluated its accuracy, precision, recall, F1 score, and created an ROC Curve.

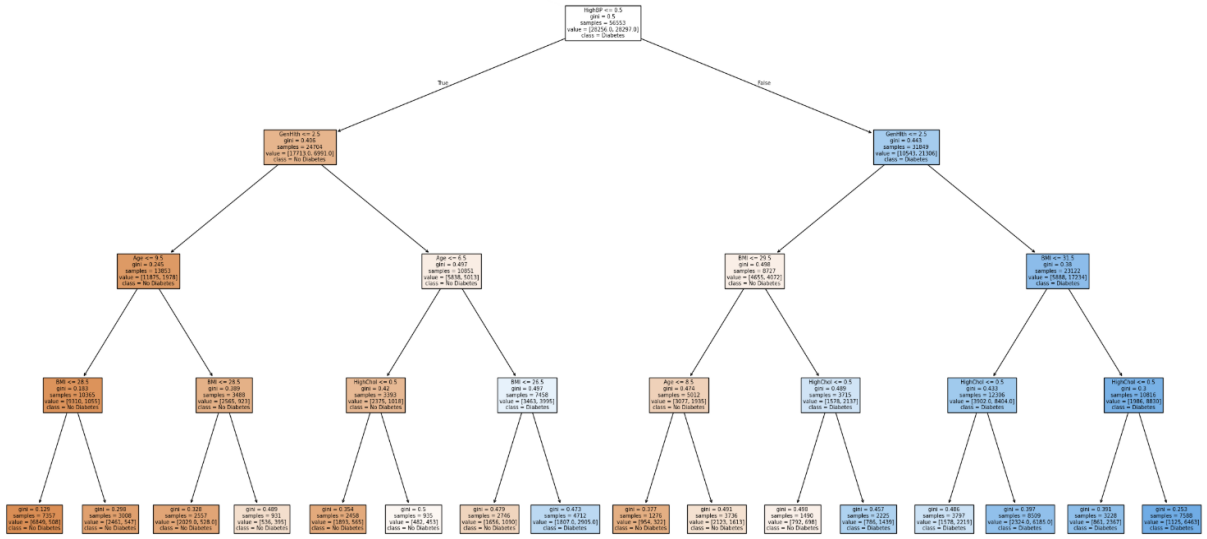
Hyperparameter tuning:

- Grid Search:
  - Achieved AUC of 0.82 opposed to 0.75 after hyperparameter tuning









```
DecisionTreeClassifier
DecisionTreeClassifier(max_depth=7, min_samples_leaf=20, min_samples_split=50,
                       random_state=42)
```

Decision Tree

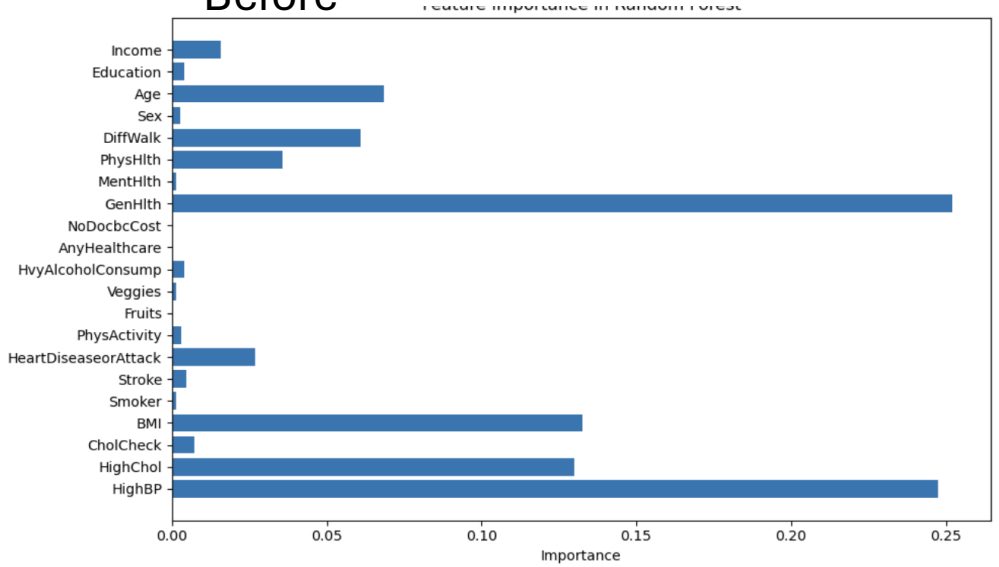
- Using the same features as the logistic regression model, we created a decision tree with 4 nodes:
  - High Blood Pressure
  - General Health (1-5 scale, 1 being the best)
  - Age
  - BMI
  - High Cholesterol

- Hyperparameter tuning:
  - tuned parameters such as max depth, min samples split, and min samples leaf
  - Optimized model: max_depth = 7, min_samples_leaf = 20, and min_samples_split = 50
  - The accuracy was 0.74 which is a slight improvement over the previous decision tree which had an accuracy of 0.728.

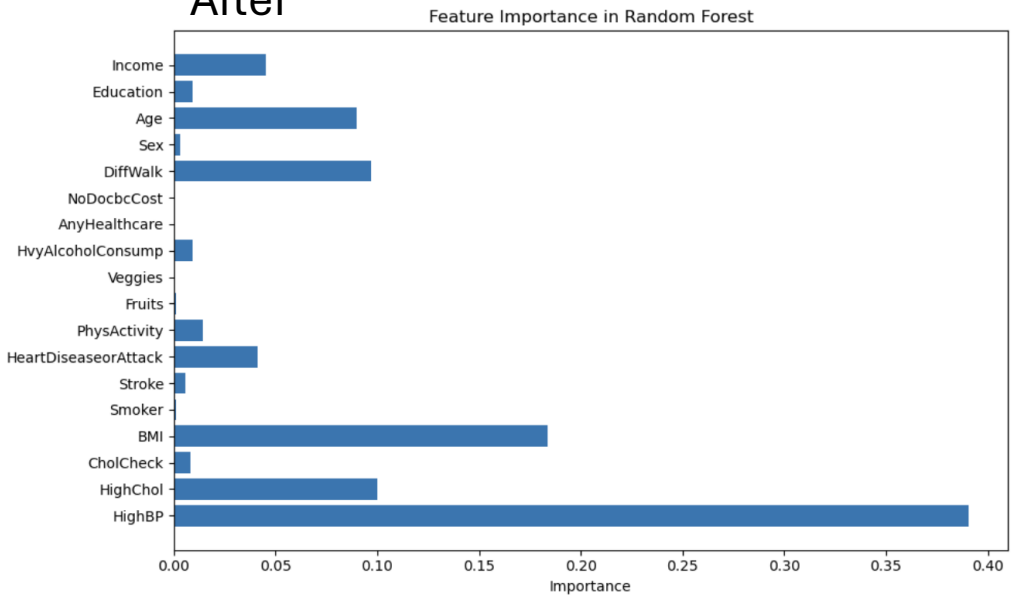
Random Forest

- Created a subset of variables that removes arbitrary factors such as General, Physical, and Mental Health that may show collinearity with other variables to give more weight to non-arbitrary classifiers.
- By decreasing and removing arbitrary estimators the model may highlight other significant predictors that are classifiable and easily identifiable
- We found that by removing variables from the model, the accuracy score did fall from 0.74 to 0.73.

Before



After



Model Selection: Random Forest

- Based on the evaluation metrics, the Logistic Regression model barely outperforms the other models. The Logistic Regression has an accuracy score of 0.75, while the Decision Tree and Random Forest models have accuracies of 0.73 and 0.74 respectively. We believe the minute differences in accuracy are negligible as another run of the models could show different results. When looking at recall, precision, and f1-score the same story holds true. Therefore, we elect to choose the **Random Forest** model as it's likely to have lower bias than the other models.

High **BMI, Cholesterol,** and **Blood Pressure** are the most significant risks and indicators of Diabetes

- **High BMI** is a strong indicator of diabetes risk — managing BMI at a healthy level can significantly reduce the risk for diabetes.
- **High cholesterol levels** often co-occur with diabetes — patients with elevated cholesterol may benefit from early screenings to prevent this.
- **High blood pressure** is frequently observed in diabetic individuals — monitoring and controlling hypertension can serve as an early intervention strategy.