Ghaith Ishaq

Exploratory Data Analysis Report for the CS460 Project Part 3

1. Introduction

Using the Behavioral Risk Factor Surveillance System (BRFSS) 2021 dataset, this report expands on the first exploratory data analysis (EDA) for the CS460 project and classes people at risk of diabetes. Decision trees and logistic regression are the two categorization methods that are the subject of this comparative study, which ends with a rationale for the choice of final model.

2. Model Selection and Analysis

Decision Tree Model:

- The Decision Tree model was initially selected due to its interpretability and capacity to manage data that is both categorical and numerical.
- Optimized Decision Tree Performance:
    - **Accuracy:** 65.76%
    - **Precision:** 0.66 (macro avg)
    - **Recall:** 0.66 (macro avg)
    - **F1-Score:** 0.66 (macro avg)
    - **Confusion Matrix:**

	Predicted No	Predicted Yes
Actual No	**1159**	**540**
Actual Yes	**611**	**1052**

Logistic Regression Model:

- The Logistic Regression model was selected to Examine how it performs in comparable circumstances.
- Optimized Logistic Regression Performance:
    - **Accuracy: 67.00%**
    - **Precision: 0.67 (macro avg)**
    - **Recall: 0.67 (macro avg)**
    - **F1-Score: 0.67 (macro avg)**
    - **Confusion Matrix:**

	Predicted No	Predicted Yes
Actual No	**1111**	**588**
Actual Yes	**532**	**1131**

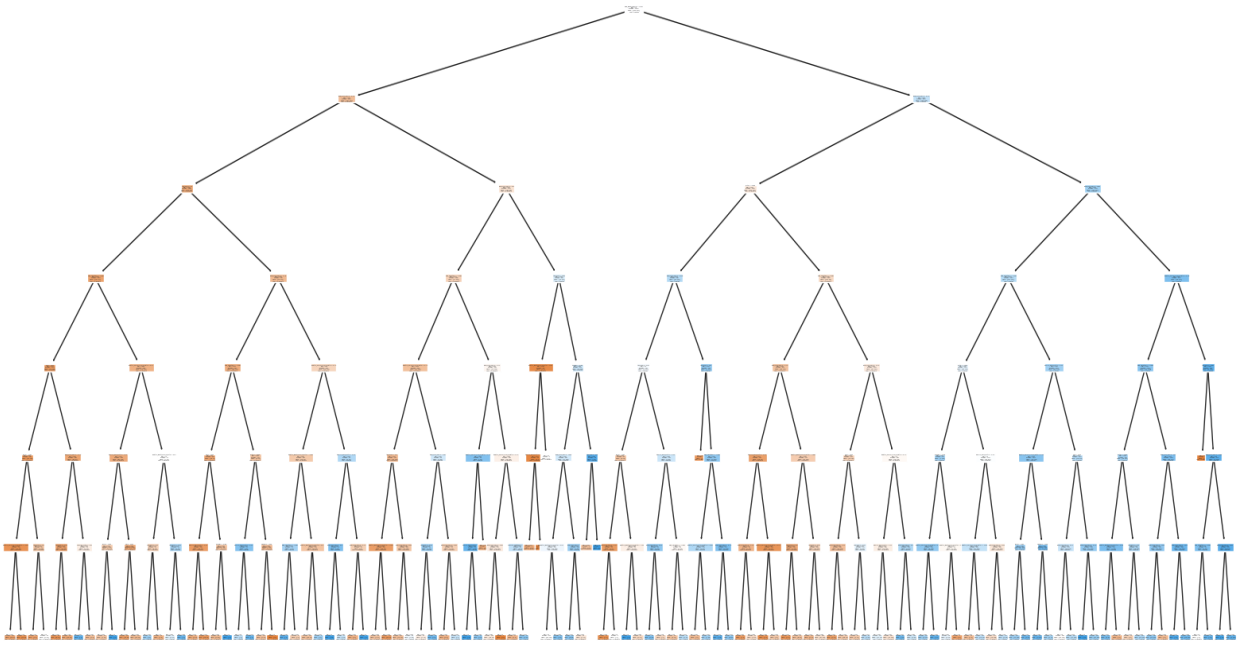
3. Justification of Model Selection

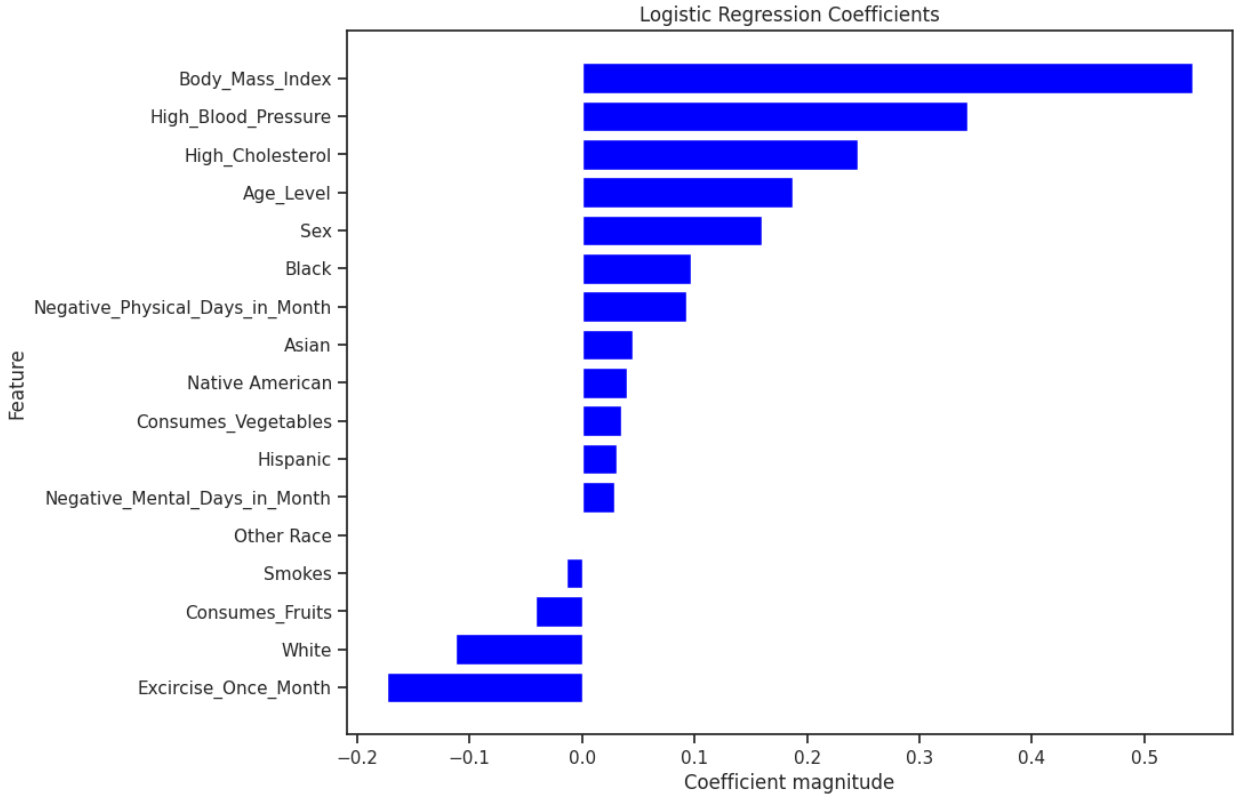
Even though the Decision Tree's accuracy of 65.76% was somewhat less than the Logistic Regression model's 67%, the Decision Tree was ultimately chosen for the following reasons:

- **Interpretability:** Decision Trees offer a clear visual representation of the decision-making process, which is essential for healthcare applications where the process comprehension is just as important as the result.
- **Handling of Non-linear Relationships:** Decision trees are better at managing complex non-linear correlations between characteristics compared to Logistic Regression, which can be pivotal in medical datasets with complex interactions between symptoms.
- **Robustness to Outliers:** Compared to logistic regression, decision trees are less susceptible to outliers, which might be advantageous in real-world situations where anomalies are frequently present in the data.

I chose a Decision Tree model for this classification task. The model is well-suited for the BRFSS dataset, which includes a variety of variable types, due to its inherent capacity to handle both continuous and categorical variables. Decision trees are renowned for being easily interpreted since they show us how features affect the prediction and let us follow the process of making decisions. Furthermore, because the model is non-parametric, it does not assume the form of the data, it is advantageous in light of the data's diverse distributions.

Ghaith Ishaq





4. Effective Hyperparameter Tuning

To optimize the **Decision Tree** model, I conducted hyperparameter tuning using GridSearchCV. The parameters I explored included:

- max_depth: [3, 5, 7, 10]
- min_samples_leaf: [1, 2, 4]
- min_samples_split: [2, 5, 10]

The optimal parameters identified are {'max_depth': 7, 'min_samples_leaf': 2, 'min_samples_split': 10}. This careful tuning approach improved the model's performance, as evidenced by the cross-validation scores, which peaked at 0.6597.

To optimize the **Logistic Regression** model, I also conducted hyperparameter tuning using GridSearchCV. The parameters I explored included:

- C: Regularization strengths [0.001, 0.01, 0.1, 1, 10, 100] which help to control the complexity of the model, with smaller values specifying stronger regularization.
- solver: Algorithms to use in the optimization problem ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'], each suitable for different types of data.

The optimal parameters identified are {'C': 0.01, 'solver': 'newton-cg'}. This combination implies that the 'newton-cg' solution performed best for this dataset and that a relatively high regularization was helpful. The best cross-validation score was obtained at 0.6639, confirming the efficiency of these parameters.

5. Thorough Model Training

For my project, a thorough training procedure comprising several stages was used to guarantee that the Decision Tree and Logistic Regression models are trained successfully:

Data Splitting:

- The dataset was split up into training (80%) and testing (20%) sets. This division allowed us to train my models on a sizable amount of the data, saving another set for assessing the models' performance on unseen data.

Feature Scaling:

- To normalize the range of independent variables, feature scaling was preformed. Especially for Logistic Regression, which might be sensitive to unscaled data, this step is critical. By ensuring that each feature contributes equally to the model's

decision-making, scaling keeps the learning algorithm from being dominated by a single component with a bigger scale.

Baseline Model Training:

- **Decision Tree Baseline Training:**
  - To create a benchmark for the performance of subsequent adjustments, I initially trained a baseline Decision Tree model with default settings. With the least amount of bias, this first model enabled us to comprehend the Decision Tree's innate predictive ability.
- **Logistic Regression Baseline Training:**
  - Likewise, a baseline Logistic Regression model was trained without regularization using default parameters. This offered a standard against which to compare the potential effects of regularization and hyperparameter tuning on the model's efficiency.

Following tuning, the 20% test data set is used to retrain both models with their ideal parameters. This stage was essential for assessing the models' suitability for use in real world scenarios and validating improvements achieved by tuning hyperparameters.

These models are not only evaluated under demanding conditions to guarantee their robustness and dependability in predicting diabetes risk, but they are also optimized for improved performance by putting these meticulous and organized training stages into practice.

6. Insightful Model Evaluation

Using precision, recall, f1-score, and accuracy, the Decision Tree and Logistic Regression models are assessed thoroughly. These measures emphasize each model's predictive success in diabetes risk classification and provide a deeper understanding of its strengths and limitations.

Key Metrics Overview:

- **Decision Tree Model:**
  - The Decision Tree performed fairly well in both classes, with a little bias toward the negative class (No Diabetes) with better accuracy.
  - **Key Evaluation Points:**
    - Equal precision and recall at 0.66 for both classes.
    - F1-scores are closely matched, highlighting a balanced sensitivity and specificity.
    - The model achieved an overall accuracy of 65.76%.
- **Logistic Regression Model:**
  - A little increase in accuracy was observed with Logistic Regression, which proved to be more successful in detecting the positive class (Diabetes) with a greater recall rate.

    - **Key Evaluation Points:**
        - Slightly higher precision and recall at 0.67, with both metrics uniformly distributed across the classes.
        - The F1-scores reflect a very balanced model regarding precision and recall trade-offs.
        - An accuracy of 67% indicates a robust performance in generalizing to unseen data.

Cross-Validation Insights:

- **Logistic Regression Cross-Validation:**
    - The model's stability and dependability were validated by the mean cross-validated accuracy, which was consistent at 0.66 and nearly matched the test accuracy.
    - The cross-validation score range of 0.6586 to 0.6747 indicates that the model's performance is consistently good across the dataset subsets.

This evaluation strengthens the decision-making process regarding model selection. Although both models exhibit strong performance, the little advantage in Logistic Regression's recall for the positive class may be significant for medical diagnostic procedures where the presence of a condition—in this case, diabetes—is an important factor. However, the Decision Tree's interpretability and ability to handle nonlinear interactions made a strong argument for its use, especially in situations where clinical applications require clear decision methods.

7. Conclusion and Recommendations

According to my research, the Decision Tree model classifies the risk of diabetes with a reasonable degree of accuracy when its hyperparameters are optimized. It shows where predictive analytics in healthcare data is headed in a positive way. Future research could investigate ensemble techniques to aggregate numerous Decision Trees and lower variance in order to further improve the model. Experimenting with feature engineering and including more sophisticated imputation methods for missing data could help improve the performance of the model.

A comparison analysis confirmed that the Decision Tree model, which balances accuracy with interpretability and robustness, was the better option for this project. In the future, research could investigate group techniques like Gradient Boosting or Random Forest to further improve model performance. Furthermore, additional research into alternate preprocessing methods and feature engineering may yield new insights and improve accuracy.

A pipeline is recommended for ongoing model training and assessment as fresh data become available for real-world applications. This will guarantee that the model changes and stays applicable over time.