FINALS Lab (Part 2)

"Obesity Classification"

William Obeng-Darko

Management Information Systems, Bowie State University

INSS662: Decision Support & Intel Systems

Dr. Sriram Srinivasan

May22nd, 2023

The dependent variable is the outcome or the target variable we try to predict or classify based on the independent variables. Therefore, in the given Obesity dataset, the dependent column is the "Label" column, which represents the obesity classification of each individual.

The independent columns in the dataset are;

- **ID:** A unique identifier for each individual. This column serves as an identifier and provides no meaningful information for predicting the obesity classification.
- **Age**: The age of the individual. It is a continuous variable that can be relevant in determining obesity classification.
- **Gender**: The gender of the individual. It is a categorical variable and can be considered an independent variable for predicting obesity.
- **Height**: The height of the individual in centimeters. It is a continuous variable that can be used as an independent variable to predict obesity.
- **Weight**: The weight of the individual in kilograms. It is a continuous variable that can be used as an independent variable to predict obesity.
- **BMI:** The individual's body mass index, calculated as weight divided by height squared. It is a continuous variable and can be considered an independent variable for predicting obesity.

To summarize, the dependent column is "Label" (obesity classification), and the independent columns are "Age," "Gender," "Height," "Weight," and "BMI." I used these independent variables to build the predictive model that classifies individuals into different obesity categories.

Here are my steps for data cleaning for the given dataset:

1. **Handling Missing/Null Values:** The dataset had no missing values. If there were, options include removing rows with missing values and filling in missing values with the mean or median.
2. **Handling Outliers:** No outliers exist in the numerical columns (Age, Height, Weight, and BMI). I recommend removing outliers or applying appropriate transformations to mitigate their effects.
3. **Data Type Conversion:** Check the data types of the columns. Ensure that they are correctly assigned. Example, the ID column should be of a numerical type float or int, while Gender and Label columns should be categorical.
4. **Encoding Categorical Variables:** Since Gender and Label columns are categorical variables, I encoded them into numerical values for analysis and modeling.
5. **Removing Unnecessary Columns:** Reviewed the dataset and identified any columns that may not be relevant to the analysis or modeling task. Thereby removing the ID column to simplify the dataset and reduce noise.
6. **Handling Duplicates:** Checked to remove any duplicate rows in the dataset.
7. **Data Validation and Sanity Checks:** I performed sanity checks on the data to identify any inconsistencies/errors. I checked that some BMI values are correctly calculated based on Height and Weight. To help validate the range and logical relationships between columns/variables.

****RESULTS****

My above table displays the K-Fold and Stratified K-Fold scores for each model. Here are some observations based on the results:

1. Logistic Regression achieves a K-Fold score of 0.7562 and a slightly lower Stratified K-Fold score of 0.7431.

2. K-Nearest Neighbors (KNN) Classifier performs well with a K-Fold score of 0.8124 and a higher Stratified K-Fold score of 0.8261.

3. Gaussian Naive Bayes shows good performance with a high K-Fold score of 0.9078 and a slightly lower Stratified K-Fold score of 0.8837.

4. Linear Support Vector Classifier (LinearSVC) achieves a K-Fold score of 0.6379, indicating moderate performance, and a lower Stratified K-Fold score of 0.5935.

5. Support Vector Classifier (SVC) has a K-Fold score of 0.6967 and a higher Stratified K-Fold score of 0.7562.

6. Decision Tree Classifier performs well, showing high accuracy with a K-Fold score of 0.9183 and an even higher Stratified K-Fold score of 0.9412.

7. Random Forest Classifier achieves the highest accuracy among all models, with a K-Fold score of 0.9882 and a slightly lower Stratified K-Fold score of 0.9765.

Based on these results, the Random Forest Classifier performs exceptionally well on the dataset, followed by the Decision Tree Classifier and Gaussian Naive Bayes. However, it's important to note that these scores are based on a smaller dataset than our other labs, so we may need to provide more diverse examples for our ML models to learn complex patterns and generalize well.