Internship Report

Health Data Science

Name: Syeda Ifroza Ahmed
Email: ifroza.ahmed@gmail.com

Submitted To: CHIRAL Bangladesh

Introduction

- **Objective**: The goal of this analysis is to predict obesity levels in individuals based on their eating habits and physical conditions using machine learning models. The dataset includes several features such as age, gender, family history, and lifestyle choices.

- **Context**: The dataset used for this analysis comes from the UCI Machine Learning Repository and contains various attributes related to personal characteristics and behaviors, which are used to predict the level of obesity in the population.

Data Description and Preprocessing

- **Dataset Overview**: The dataset includes various features related to individuals' demographics (e.g., age, gender), health history (e.g., family history of obesity), and lifestyle factors (e.g., smoking habits, eating habits, physical activity). The target variable is NObeyesdad, which indicates the level of obesity (e.g., obesity class).

    - The dataset was accessed from the UCI repository and is in .zip format, which was extracted and read into a DataFrame.

- **Preprocessing**:

    - **Handling Missing Values**: Checked for missing values with df.isnull().sum().
    - **Duplicate Removal**: Duplicates were detected and dropped using df.drop_duplicates().
    - **Categorical Features Encoding**: Label encoding was applied to categorical features (Gender, SMOKE, family_history_with_overweight, etc.) using LabelEncoder.
    - **Feature Types**: The dataset has both numerical and categorical features. Categorical features were encoded, and numerical features were handled using appropriate methods (e.g., box-plots).

Exploratory Data Analysis (EDA)

- **Univariate Analysis**:

    - **Descriptive Statistics**: Basic statistics (e.g., mean, standard deviation, min, max) were printed for the dataset, which helps in understanding the distribution of features.

    - **Distribution of Obesity Levels**: The distribution of the target variable (NObeyesdad) was visualized using a count plot. This revealed how the data is balanced across different obesity levels.

    - **Boxplots for Numerical Features**: Boxplots were generated for Age, Height, and Weight against the obesity levels. These visualizations help to understand how these continuous features vary across different obesity levels.

- **Bivariate Analysis**:

    - **Eating Frequency and Obesity Levels**: The relationship between eating frequency (CAEC) and obesity levels was explored using a countplot, revealing patterns between eating frequency and the obesity classification.

    - **Family History vs. Obesity Level**: A countplot was used to examine how family history of obesity impacts the distribution of obesity levels.

- **Correlation Matrix**: A heatmap was generated to analyze the correlations between numerical features. This is useful to understand the relationships between features and identify potential collinearity.

Model Development and Evaluation

In this section, various machine learning models were trained and evaluated to predict obesity levels in individuals based on eating habits and physical condition. The models used were Logistic Regression, Decision Tree, and Random Forest. The performance of these models was assessed using precision, recall, F1-score, and accuracy metrics as well as ROC-AUC.

1. Logistic Regression

- **Model Summary**: Logistic Regression is a linear model used for binary and multi-class classification. We trained the model using the fit() method and evaluated it using the test data.

- **Performance Metrics**:

    - **Precision**: Precision measures the accuracy of positive predictions. For instance, for obesity level 0, the model achieved a precision of 0.83, indicating that 83% of predicted level 0 cases were true positives.

    - **Recall**: Recall measures the percentage of actual positive cases correctly identified by the model. For obesity level 1, recall was 0.58, meaning that the model correctly identified 58% of the true level 1 cases.

    - **F1-Score**: F1-score is the harmonic mean of precision and recall. For obesity level 2, the F1-score was 0.85, showing a good balance between precision and recall.

    - **Accuracy**: Overall, the logistic regression model achieved an accuracy of 82%, correctly predicting obesity levels across all classes.

2. Decision Tree

- **Model Summary**: The Decision Tree algorithm is a non-linear classifier that splits data based on feature values. It works well for handling both numerical and categorical data.

- **Performance Metrics**:

- **Precision**: The decision tree model achieved excellent precision, with the highest value of 1.00 for obesity level 4, which is ideal.

- **Recall**: The recall for most classes was very high, reaching 1.00 for obesity level 4 and 3, meaning that these obesity classes were completely captured by the model.

- **F1-Score**: The F1-scores across the classes were very balanced, with level 4 achieving a perfect F1-score of 1.00.

- **Accuracy**: The decision tree classifier performed very well, with an accuracy of 95%.

3. Random Forest

- **Model Summary**: Random Forest is an ensemble learning method that creates a collection of decision trees, typically used to improve accuracy and robustness.

- **Performance Metrics**:

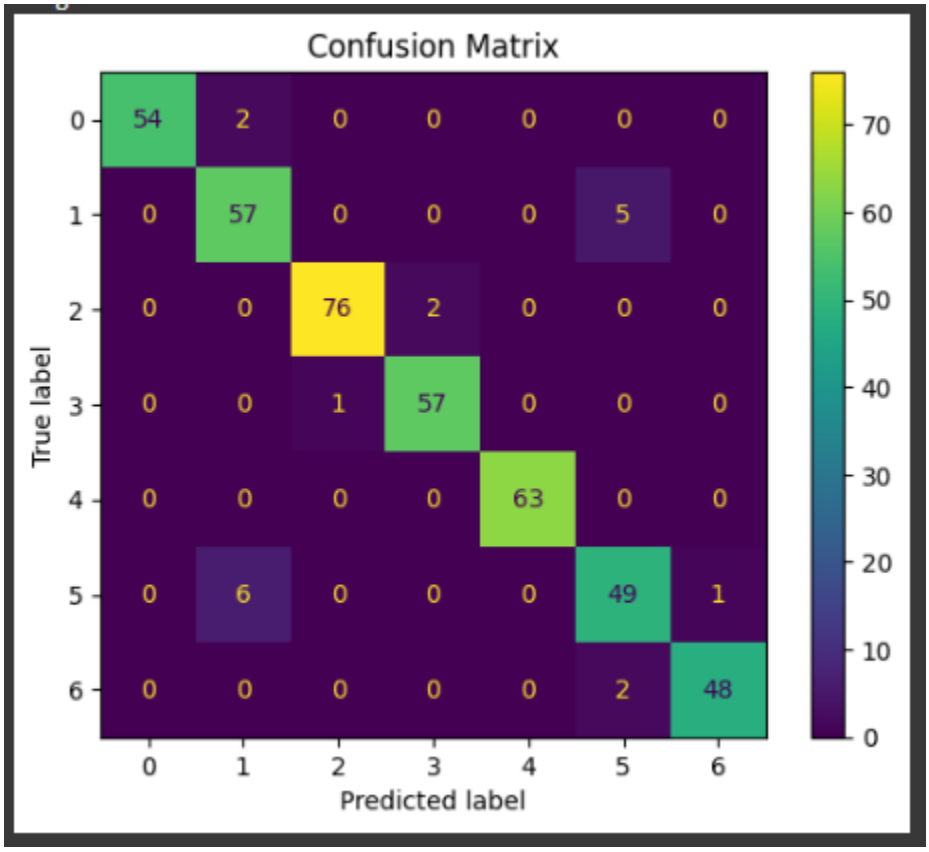
  - **Precision**: The Random Forest model performed excellently, with precision values close to 1.00 for many classes (e.g., obesity level 4).

  - **Recall**: The recall was high for all classes, with obesity levels 0, 4, and 6 showing high recall values.

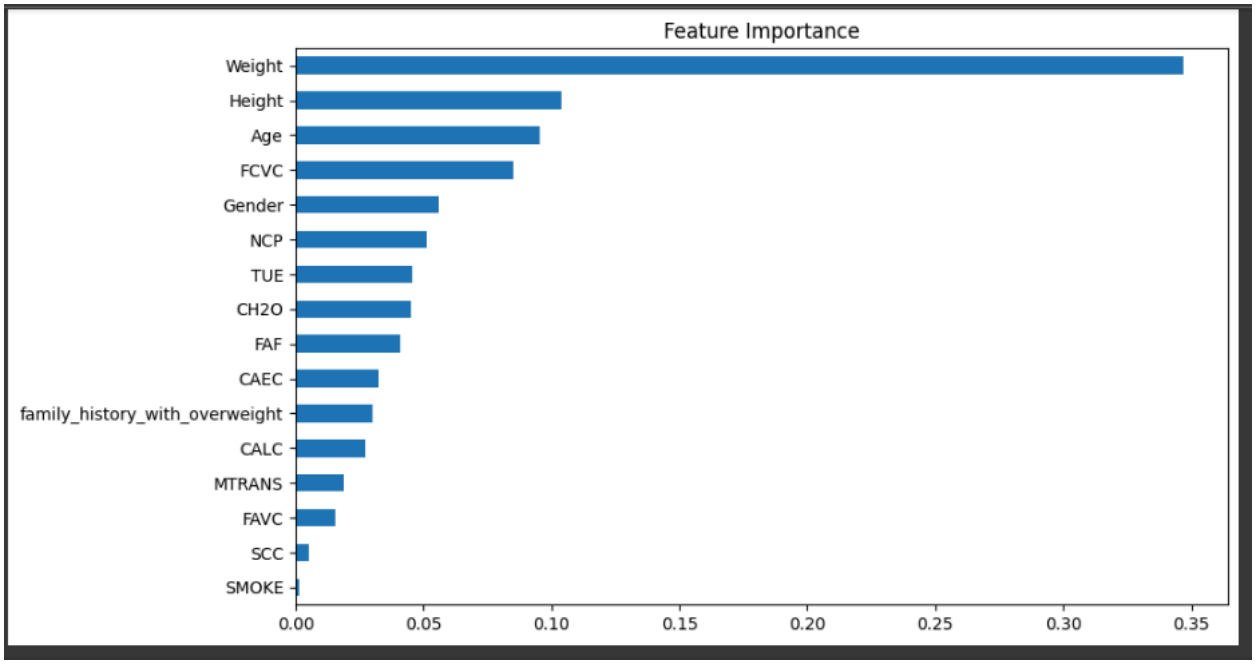
  - **F1-Score**: The F1-scores were uniformly high, reflecting good performance across all classes.

  - **Accuracy**: The random forest classifier achieved the highest accuracy of 96%, making it the best-performing model in this analysis.

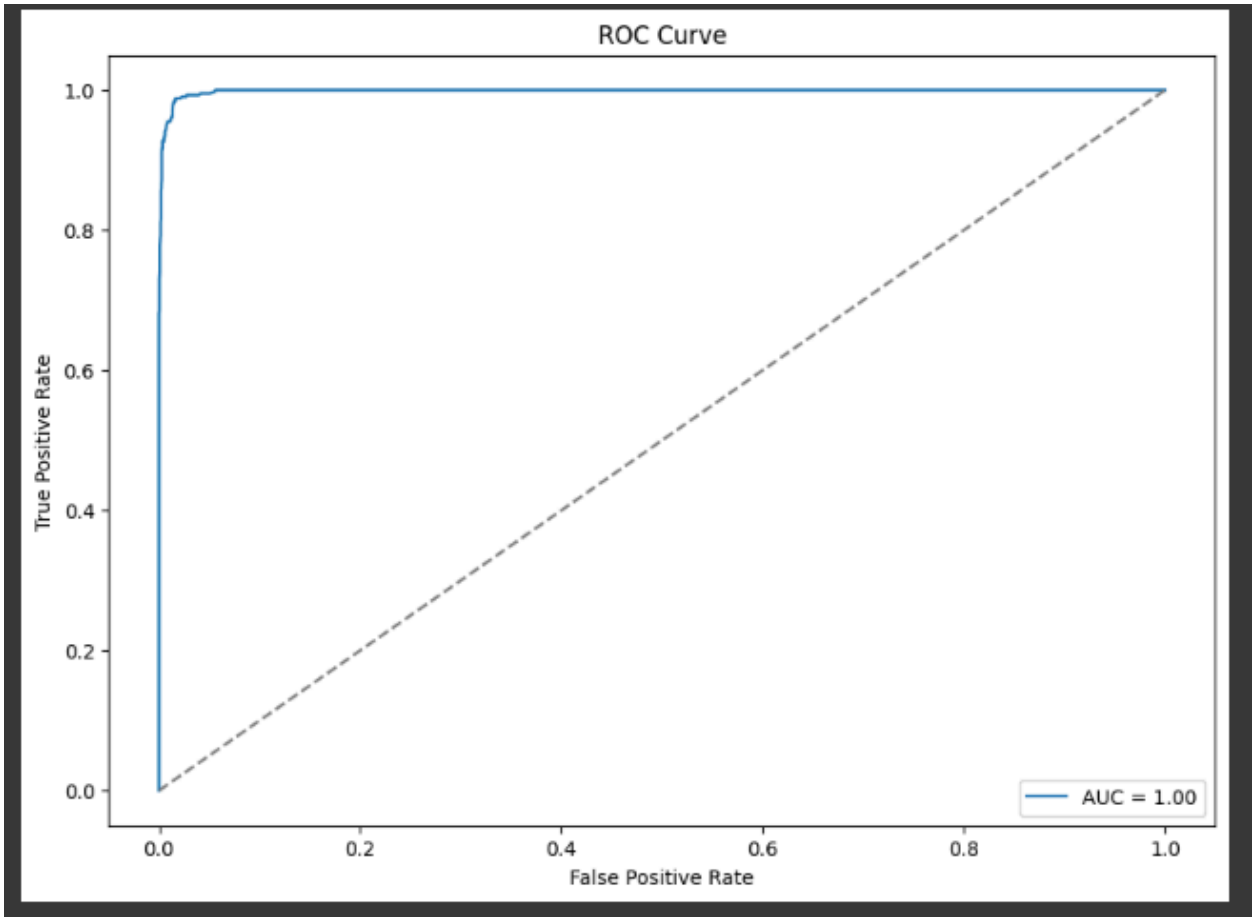
Confusion matrix - The confusion matrix for the best performing model,Random Forest, is given below.



Feature importance - The feature importance graph provides the level of importance for each of the given features, where the most important feature came out to be weight, next were height and age, respectively.



ROC-AUC- This score is a useful evaluation metric, especially for classification tasks, as it gives an aggregate measure of a model's ability to distinguish between classes. For this case, the AUC came to be a perfect 1.00 so here the classes are being distinguished very well. Meanwhile, the curve is close to the top-left corner signifying a high TPR.



Research Application:

Potential for Real-World Applications

The models developed in this study can be a valuable tool in real-world applications, especially in obesity prevention programs. The main objective of such programs is to identify individuals at high risk of obesity early and offer preventive measures to mitigate this risk. Here are some potential applications:

1. **Early Risk Identification**:
    - The models can help healthcare providers identify individuals who are at risk of obesity based on various lifestyle factors such as age, weight, height, eating

habits, and family history.

- By predicting obesity levels, healthcare systems can focus on at-risk populations and provide targeted interventions, such as personalized diet plans, physical activities, and behavioral counseling.

2. **Personalized Health Programs**:

   - The models can assist in developing personalized obesity prevention plans tailored to individuals based on the identified risk factors.

   - For example, if a model indicates that a person has a high risk due to poor eating habits, they could be enrolled in an educational program on healthy eating and exercise habits.

Limitations of the Study

While the model offers promising insights and applications, there are several **limitations** that need to be addressed:

1. **Data Quality and Representation**:

   - The dataset used in this study includes self-reported data on factors like eating habits, physical activity, and family history. Self-reported data can be biased or inaccurate, which may affect the model's performance.

   - The dataset is also limited to individuals from specific countries like Mexico, Peru, and Colombia, which may affect its generalizability to other populations or cultures with different lifestyle patterns or healthcare systems.

2. **Overfitting and Model Complexity**:

   - Despite good performance, overfitting can be a concern, especially for complex models like Random Forest. These models may perform well on the training data but fail to generalize to new, unseen data.

3. **Limited Features**:

   - The study only considers demographic and behavioral features. It overlooks other potential variables like genetic predispositions, socioeconomic factors, and

psychosocial elements (e.g., mental health), which may also play a significant role in obesity prediction.

Future Research Directions

To address these limitations and enhance the effectiveness of obesity prevention programs, several future research directions can be considered:

1. **Expanding the Dataset**:
    - Future research should aim to collect data from more diverse and representative populations to improve the model's robustness and generalizability. Including data from different countries, cultures, and socioeconomic backgrounds would help create more inclusive and universally applicable models.
2. **Incorporating More Complex Features**:
    - Incorporating additional features, such as detailed medical history, genetics, or environmental factors, could improve the model's accuracy and relevance.
3. **Implementation in Clinical Settings**:
    - Future research should focus on testing the developed models in **real-world clinical environments** to evaluate their practicality and effectiveness in helping healthcare providers make informed decisions.
    - Investigating **cost-effectiveness** and **user acceptance** of predictive tools in obesity prevention programs will be crucial for their successful implementation.

Insights and Recommendations

Insights

From the analysis and model development, several key insights have emerged:

1. **Feature Importance**:

- Features such as Weight, Height, Age, and Eating Habits play a significant role in predicting obesity levels. These insights align with common understanding of obesity risk factors, demonstrating the relevance of these variables in predicting obesity.

- The family history of overweight also has a strong correlation with obesity risk, suggesting that genetic and familial factors are crucial in determining obesity levels.

2. **Model Performance**:

    - Among the models tested, the Random Forest classifier outperformed the Logistic Regression and Decision Tree models, achieving the highest accuracy and F1 scores across all classes. This suggests that ensemble methods like Random Forest are well-suited to handling the complexity of this dataset.

    - The Decision Tree model, while performing slightly lower than the Random Forest, still delivered competitive results with good interpretability, making it a practical option for use in healthcare applications where transparency is key.

3. **Impact of Eating Habits and Lifestyle**:

    - The analysis highlighted the significant influence of eating frequency and lifestyle factors like physical activity levels in determining obesity risk. Interventions targeting these areas could be a key focus in obesity prevention programs.

Recommendations

Based on the insights drawn from the analysis, the following recommendations are made:

1. **Enhance Data Collection**:

    - To improve the model's performance, future research should focus on expanding the dataset with more detailed and diverse features.

    - Including more real-time and accurate data on physical activity levels and food intake would enhance the model's predictive power.

2. **Focus on Targeted Interventions**:

- Obesity prevention programs should focus on personalized interventions based on the predicted obesity levels from the models. For example, individuals at high risk based on their eating habits or physical condition can be provided with tailored recommendations, including diet changes or exercise routines.

3. **Long-Term Monitoring**:

    - Models like the ones developed here could be integrated into **long-term monitoring systems** where regular health assessments are conducted. By monitoring key factors over time (e.g., changes in eating habits, weight, physical activity), predictive models can help adjust intervention strategies dynamically.

Conclusion

In conclusion, this study has developed and evaluated machine learning models to predict obesity levels based on various factors such as demographic information, eating habits, and physical condition. The results indicate that **Random Forest** is the most accurate model for this task, providing reliable predictions that could possibly be applied in **obesity prevention programs**.

However, the study also highlights several limitations, including the need for more diverse datasets. By addressing the limitations and incorporating more data sources, future research can enhance the accuracy and applicability of these models, ultimately contributing to more effective obesity prevention strategies and health management.

This work opens up new avenues for integrating machine learning techniques into public health initiatives, making personalized health interventions more efficient and targeted, with the potential to improve public health outcomes globally.