Estimation of Obesity Levels Based on Eating Habits and Physical Condition

1. INTRODUCTION

Obesity has emerged as a significant public health concern, influenced by various factors including eating habits and physical condition. Understanding the relationship between dietary choices and physical activity levels is crucial for assessing obesity levels within populations. This study aims to explore how specific eating patterns and the overall physical condition of individuals contribute to the prevalence of obesity. By analyzing these factors, we can gain insights that may inform effective interventions and promote healthier lifestyle choices.

2. DESIGN AND IMPLEMENTATION

2.1 Data Collection

In this study, the approach use a dataset known as the "ObesityDataSet_raw_and_data_sinthetic" acquired from HubbleMind. The dataset contains a total of 16 feature variables and 1 target variable. The dataset comprises of 17 various attributes covering 8 numerical and 9 nominal categories. A total of 2111 instances are used for prediction purposes, where all the instances available in the dataset are utilized.

To gain deeper insights into the underlying patterns and relationships within the data, the dataset is visualized using various plots and charts, enabling a clearer understanding of feature distributions, correlations, and potential anomalies that could impact the model's performance.

2.2 Data Preprocessing

The dataset has no missing values, therefore categorical features needs to be encoded, so this study use label encoder for features like "Gender", "family_history_with_overweight", "FAVC", "SMOKE", "SCC", and "Nobeyesdad" because some of the features contains binary values(i.e. yes, no) while the rest have multi-class values but there are ordinality in them. One hot encoder was use to encode other features like "CAEC", "CALC", "MTRANS" because they do not have ordinality and they are multi-class features. Real world data contains some outliers which is due to several reasons, Figure 2.1 shows how this study use boxplot to plot the distribution graph of the continuous variable. From the graph, it shows that some of the continuous variable like "age" and "NCP" have many outliers, but the author did not deal with it because the average range of human age is 20-80. Inter-quartile range technique was used to cap "Height" and "Weight" columns and afterwards was normalize using sk-learn MinMax library.

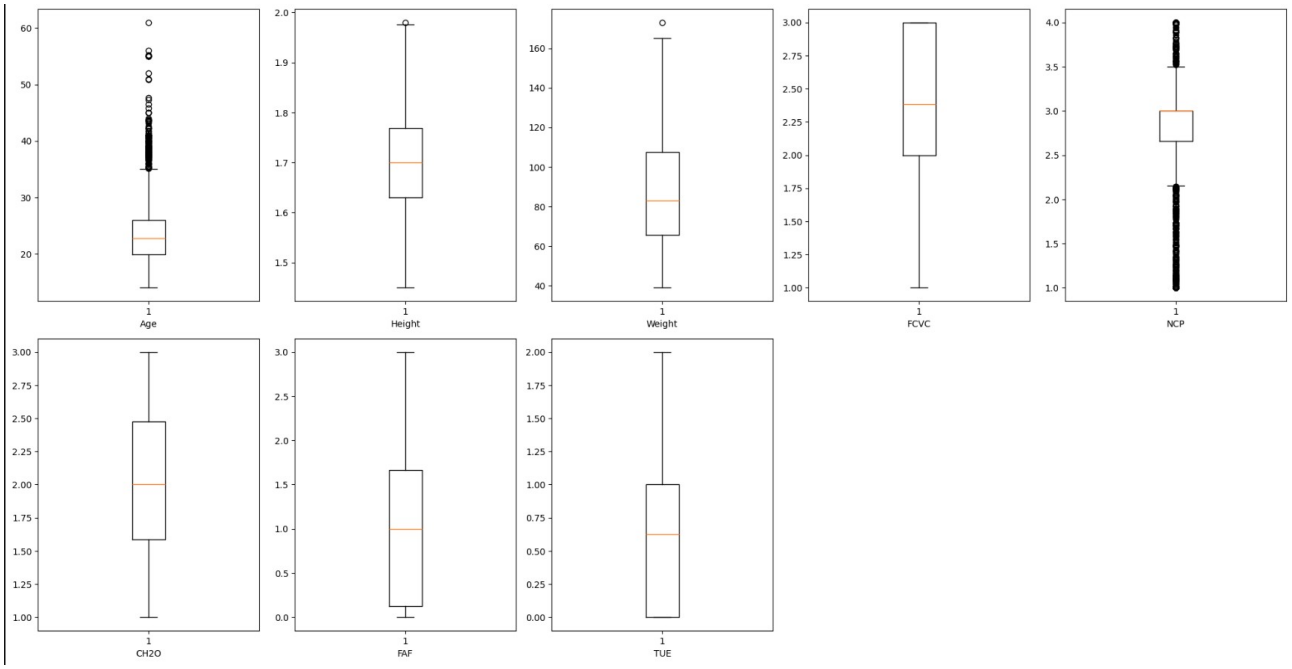

Figure 2.1 Boxplot of the continuous variable.

2.3 Exploratory Data Analysis

Figure 2.2 shows how each continuous variable is distributed, depicting the mean, mode, and the median. With the “age” column the mean is 24, median is 22 and the mode is 18, “Height” column the mean 1.75, median is 1.73, mode is 1.8, “Weight” column the mean is 85, mode is 75, median is 80, “FCVC” column the mean and the median is 2.4 and the mode is 3, “NCP” column the mean is 2.7, mode is 3, “CH2O” column the mode is 2, “FAF” column the mean and the median is 1, mode 0, and the “TUE” column the mean is 0.8, median is 0.6 and mode is 0

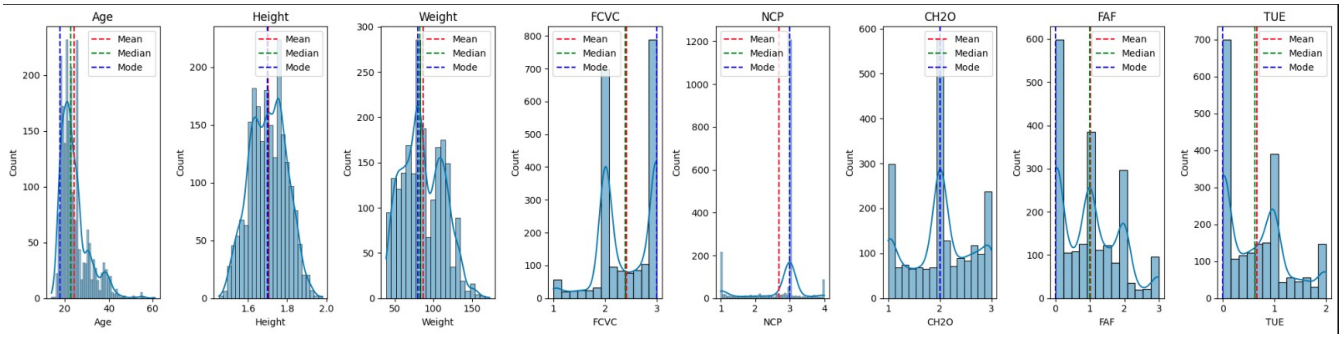

Figure 2.2 Distribution graph for continuous variable.

Figure 2.3 show how the values of each continuous variable is plotted, the graph show the relationships between the continuous features and the obesity levels. Features like “FAF” and “TUE” have no outliers in them, while “Weight”, “CH2O” and “FCVC” have mild outliers.

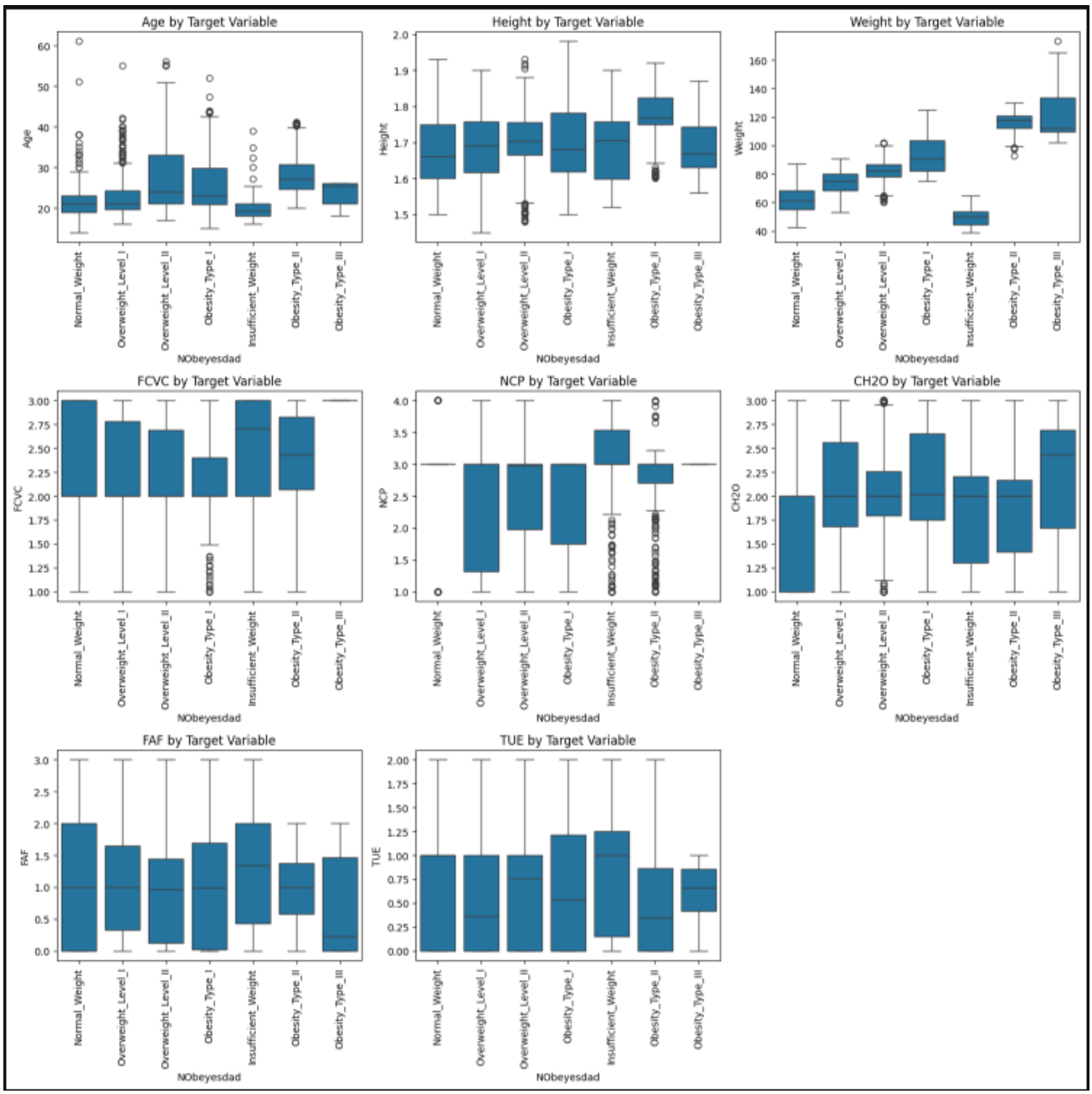

Figure 2.3 Relationships between features and obesity levels.

Figure 2.4 shows how the author use heatmap to visually depict the strength and the direction among the continuous variables, feature variables like “Age”, “Weight”, “FCVC” tends to decrease in direction and have strong relationship. Also features like “Height”, “NCP”, “CH2O”, “FAF” shows a positive relationship but the correlation is weak. Lastly “TUE” correlation is positively strong.

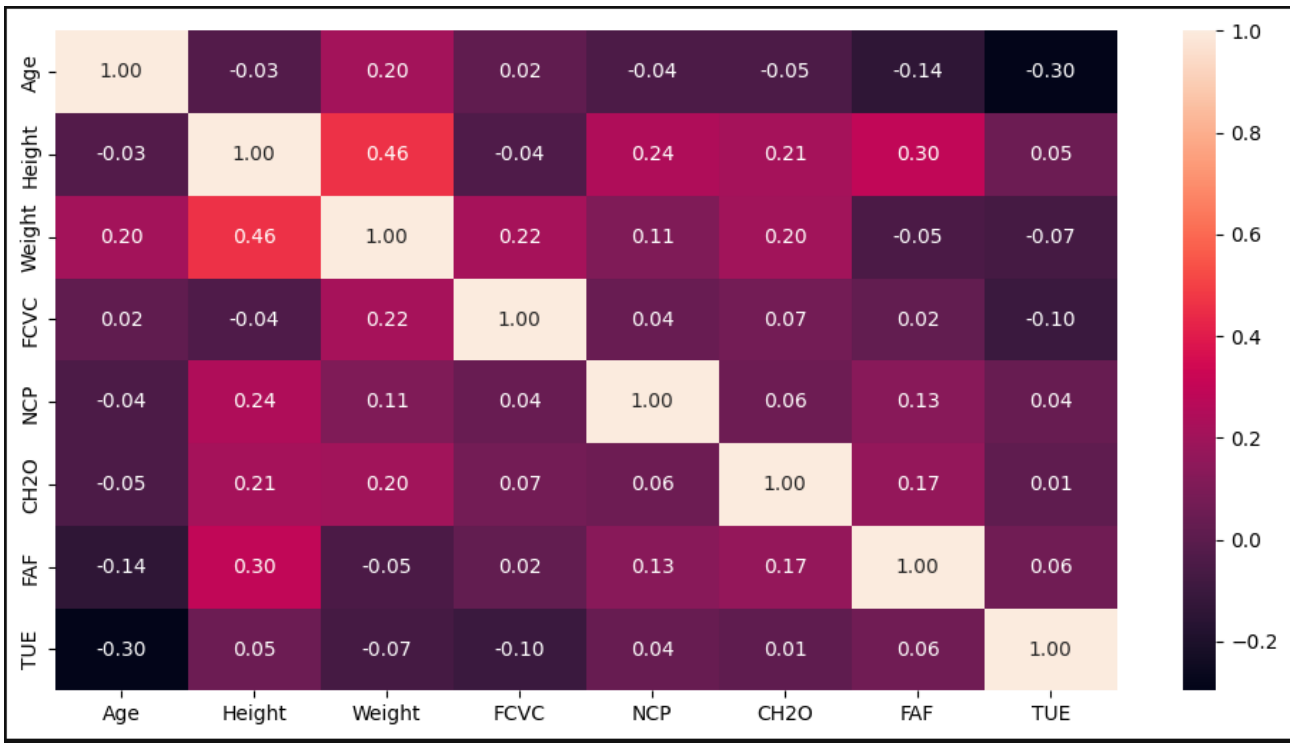

Figure 2.4 Correlation analysis among the continuous variables.

2.4 Machine Learning

Before diving into the various models used in this study, the author throw more light to feature relationships using pair plot. Pair plot is like a scatter plot which shows how values of one variable relate to values of another. Figure 2.5 represent the pair plot which show the relationships among the features, The diagonal plots represent the distribution (density or KDE plots) of individual features. Each color corresponds to a different weight category, showing how each category is distributed across each feature. “Age” and “Height” appear to have a distinct pattern, where individuals with different weight categories are separated across the diagonal. “Height” and “Weight” show a strong positive relationship, which is expected, as higher weight generally correlates with taller height.

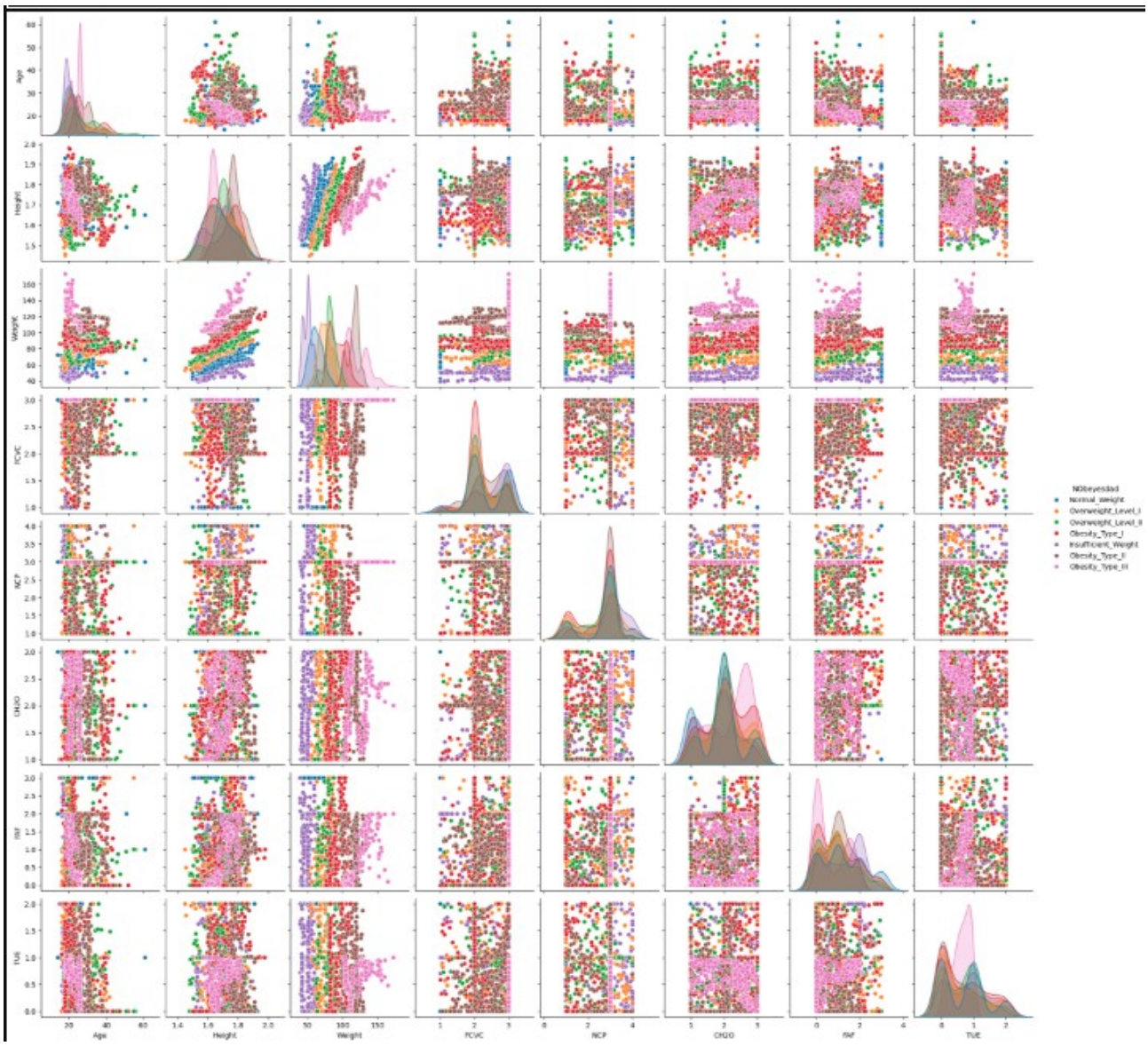

Figure 2.5 Pair plot graph to explore features relationships.

This study emphasis about which feature has the critical impact on the model. Figure 2.6 represent feature importance from Random forest classification model. From the graph, "Weight" has the most impact on the model, "Age", "Height", and "FCVC" have relatively high impact, while "MTRANS_Motorbike", "MTRANS_Bike", "CALC_Always" has no impact on the model. The rest also have mild impact.

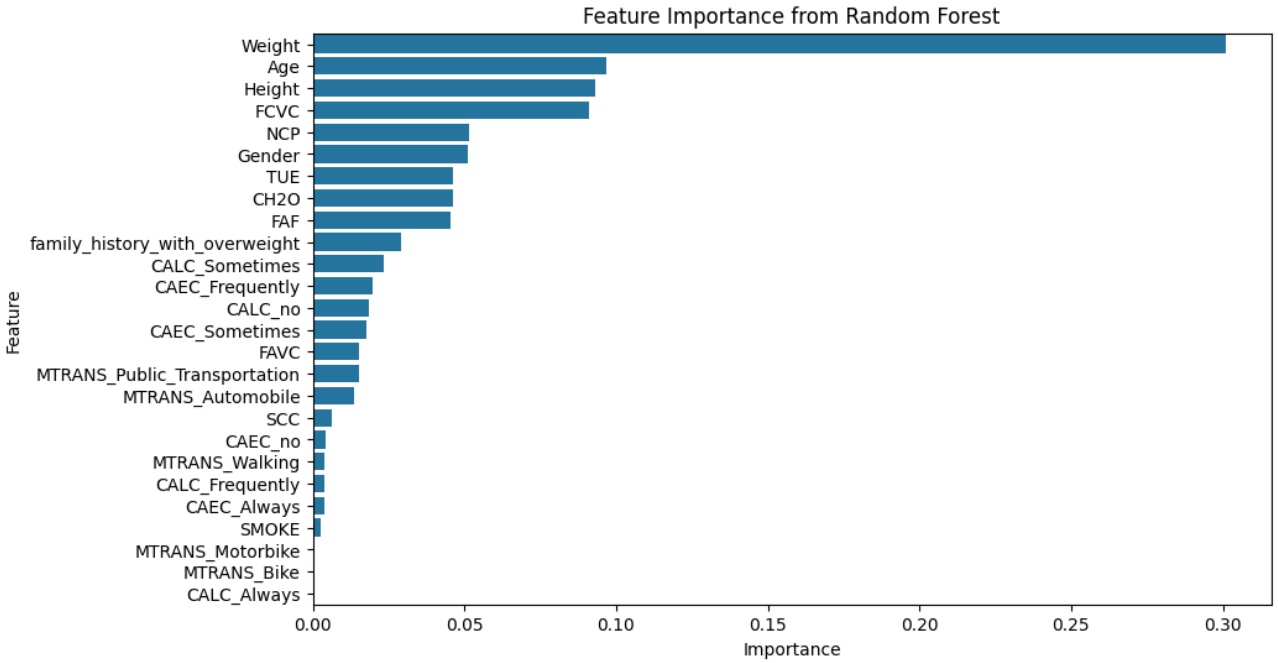

Figure 2.6 Feature Importance from Random Forest.

2.4.1 Evaluation Metrics

In this study, after the dataset has been preprocessed, it was split into training set and testing set. Training set size is 80% and the remaining 20% were used for the testing set. After this step, it was fit into the model. Below are the key metrics in which this study evaluates the models on, they are;

- Confusion Matrix: Confusion matrix is a table with combination of predicted values compared to actual values. It measure the performance of the classification problem.
- Precision: Precision is fraction of the true positive among all the examples that were predicted to be positive.
- Recall: Recall is the fraction of true positives among all the examples that were actually positive.
- F1-score: F1-score is the harmonic mean of the precision and the recall.

2.4.2 Logistic Regression

Figure 2.7 shows how well logistic regression model perform on the dataset. With 86% as the train accuracy and 84% as the test accuracy, this study concludes that this model performs well while dealing with overfitting and underfitting. In terms of precision, recall, and f1-score the performance of this model did well.

```
Training Accuracy of Logistic regression is 0.8601895734597157
Test Accuracy of Logistic regression is 0.8416075650118203

              precision    recall  f1-score   support

           0       0.82      0.98      0.89        56
           1       0.82      0.53      0.65        62
           2       0.91      0.90      0.90        78
           3       0.91      1.00      0.95        58
           4       1.00      1.00      1.00        63
           5       0.65      0.75      0.69        56
           6       0.74      0.70      0.72        50

    accuracy                           0.84       423
   macro avg       0.84      0.84      0.83       423
weighted avg       0.84      0.84      0.84       423
```

Figure 2.7 Evaluation Metrics of the Logistic Regression.

Figure 2.8 shows the confusion matrix of the Logistic regression model. The diagonal values are rate at which the model predicted the true positive and the true negative.

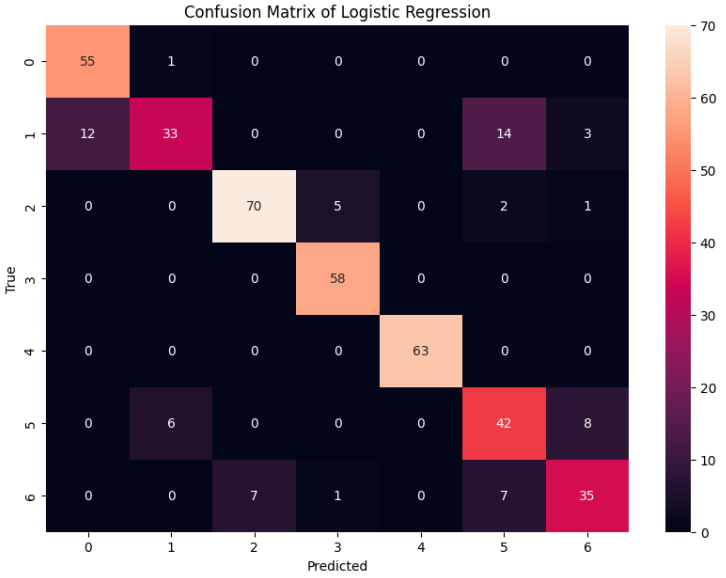

Figure 2.8 Confusion matrix of Logistic Regression Model.

2.4.3 Random Forest

Figure 2.9 shows the classification report by the Random Forest Classifier. The train accuracy is 100% which means the model was able to learn the patterns of the dataset, the test accuracy is 95% which means the model was able to perform well on the new dataset(i.e. test set). The precision, recall, and the f1-score performance is great.

```
Training Accuracy of Random Forest Classifier is 1.0
Test Accuracy of Random Forest Classifier is 0.9479905437352246

Classification Report :-
               precision    recall  f1-score   support

           0       0.98      0.96      0.97        56
           1       0.85      0.90      0.88        62
           2       0.99      0.95      0.97        78
           3       0.98      0.98      0.98        58
           4       1.00      1.00      1.00        63
           5       0.88      0.88      0.88        56
           6       0.96      0.96      0.96        50

    accuracy                           0.95       423
   macro avg       0.95      0.95      0.95       423
weighted avg       0.95      0.95      0.95       423
```

Figure 2.9 Evaluation Metrics of the Random Forest Classifier.

Figure 2.10 shows the confusion matrix of the Random Forest model. The diagonal values are rate at which the model predicted the true positive and the true negative.

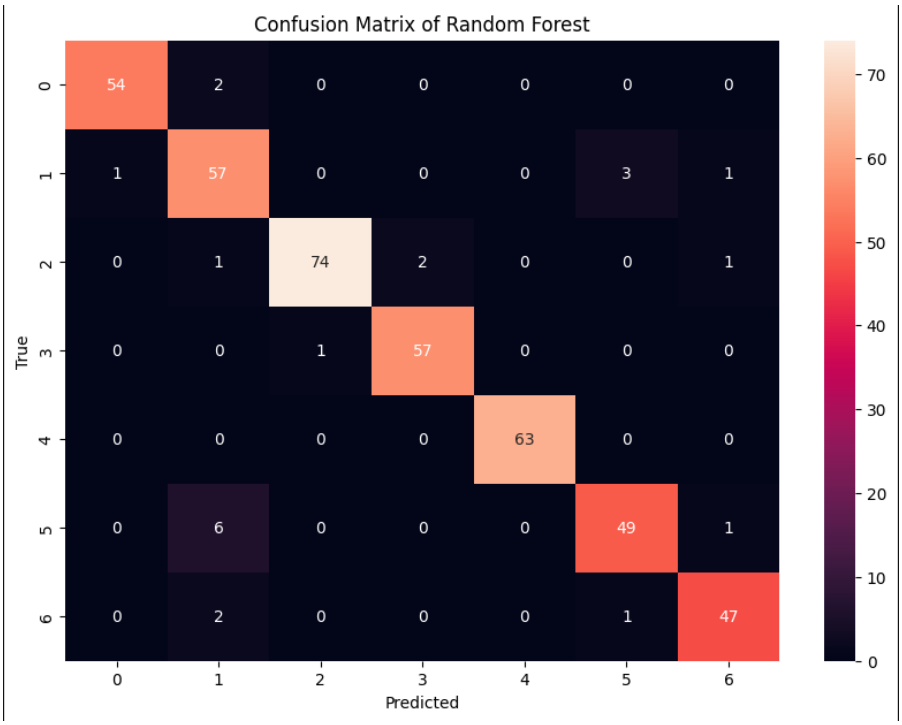

Figure 2.10 Confusion matrix of Random Forest.