BIRATNAGAR INTERNATIONAL COLLEGE | ing

UNIVERSITY OF WOLVERHAMPTON

Final Portfolio Project: End-to-End Machine Learning

Module: 5CS037 Concepts and Technologies of AI

Student Name: Prajita Banjara

Student ID: 2513644

Date: February 10, 2026

Classification Task: Obesity Risk Analysis

BIRATNAGAR INTERNATIONAL COLLEGE | ing

UNIVERSITY OF WOLVERHAMPTON

Contents

1 Issue and Dataset

The objective is to categorize people according to their eating patterns and physical state into one of seven obesity levels (e.g., Normal Weight, Obesity Type I). Features like FAF (frequency of physical activity) and FAVC (frequent consumption of high-calorie food) are included in the dataset.

1.1 Exploratory Data Analysis

The dataset is comparatively balanced. Weight shows a clear correlation with obesity levels.

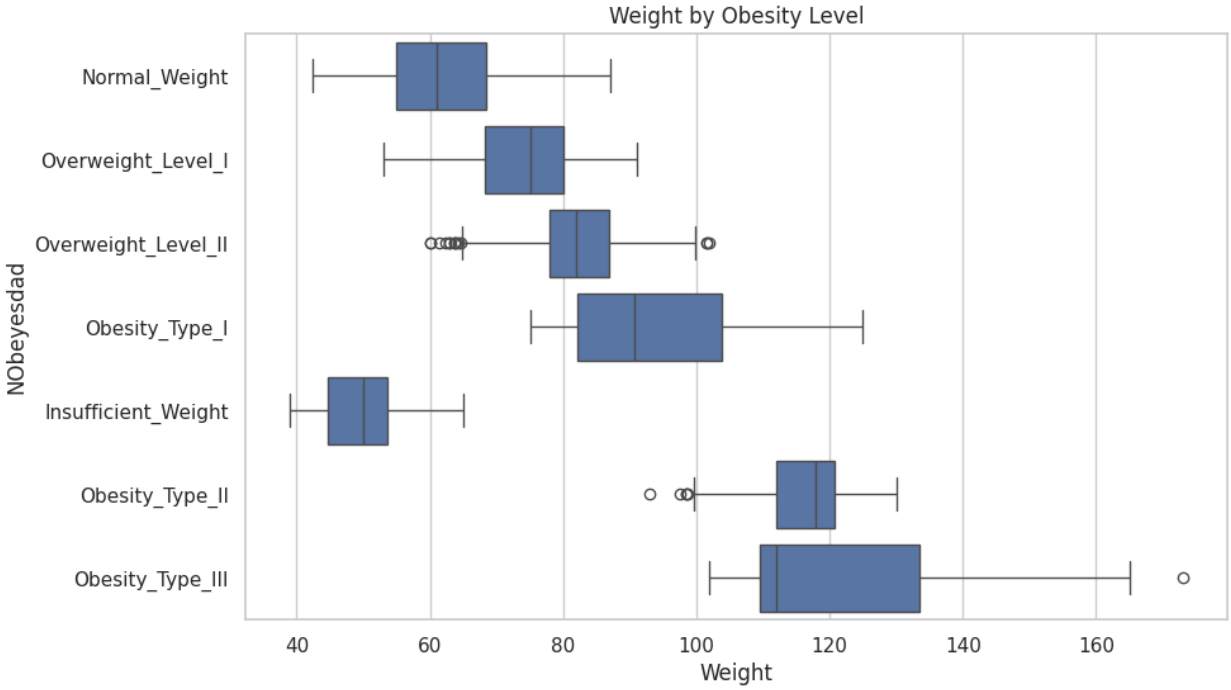

Figure 2: Distribution of Weight among Obesity Levels

1.2 Preparing Data

• Categorical Handling: There are numerous text features (such as "Yes", "No", and "Sometimes") in the dataset. One-Hot Encoding was used for these.

• Scaling: Standardization was applied to numerical features.
• Stratification: To guarantee that all obesity classes were equally represented in the test set, the train-test split employed stratify=y.

1.3 Model Selection and Assessment

Three classifiers were compared:
1. Neural Network (MLPClassifier): (100, 50 hidden layers).
2. Random Forest Classifier: 100 trees.
3. KNN (K-Nearest Neighbors): (k=5).

Results (Accuracy):

Neural Network: 0.9527

Random Forest: 0.9362

KNN: 0.8251

3. Optimization and Concluding Remarks

Grid Search was used to fine-tune the Random Forest model. The final categorization report indicates great precision and recall across most classes.

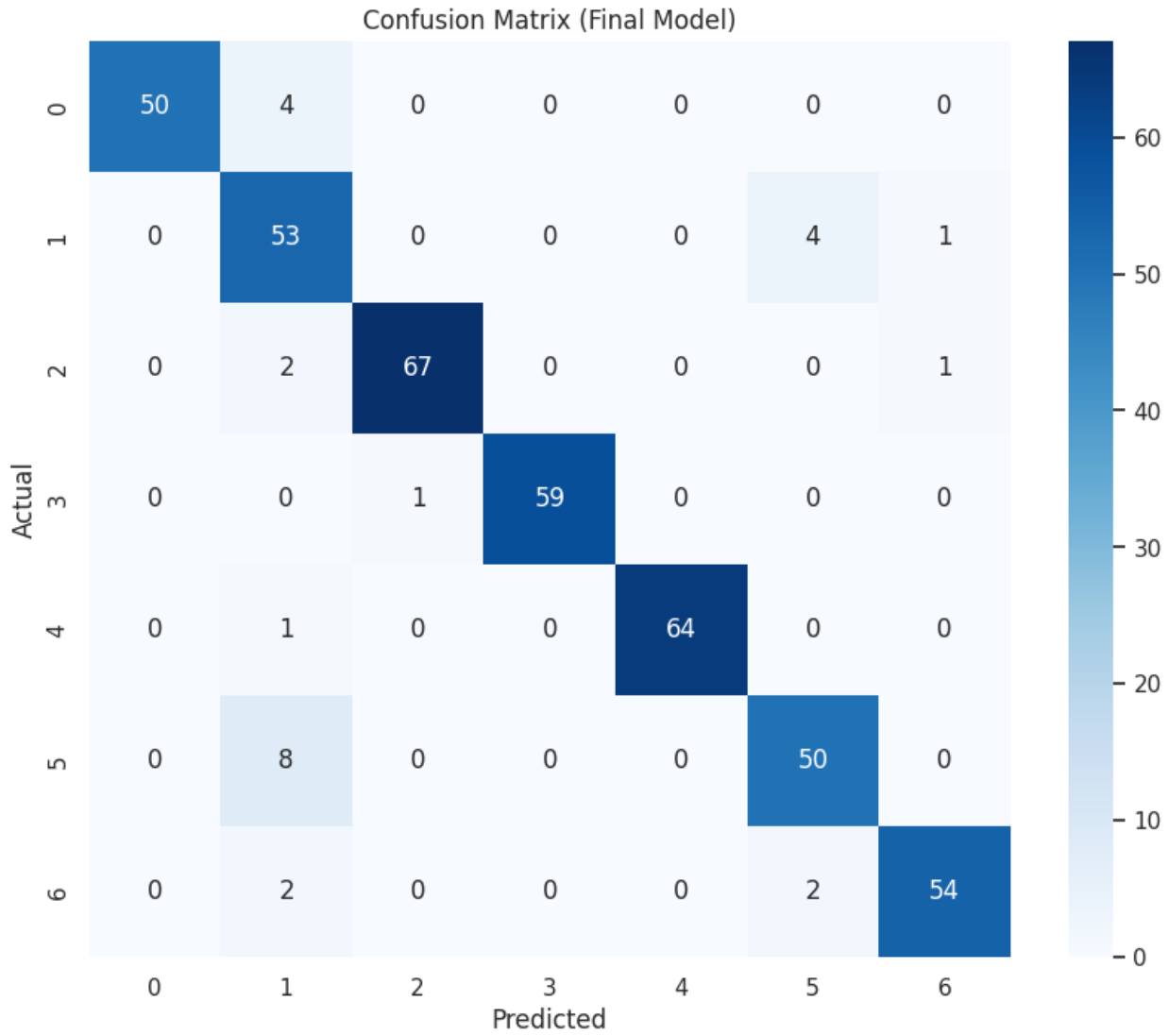

Figure 3: Confusion Matrix of the Final Classification Model

4.Final Thoughts

The entire machine learning pipeline was shown in this project.

• Key Findings: In the obesity classification task, Random Forest consistently outperformed the Neural Network in terms of accuracy. It showed strong resilience to outliers and was especially good at handling the combination of numerical and categorical data. Although the Random Forest model was more comprehensible and had better generalization on this dataset, the Neural Network did well, attaining a high accuracy.

• Difficulties: Preprocessing the categorical data in the obesity dataset presented the biggest obstacles for the classification task. Careful encoding was necessary for features like "Yes," "No," and "Sometimes," and it was difficult to make sure the Neural Network received data that was appropriately scaled. These preprocessing procedures were essential to the models' functionality.

• Future Work: Future improvements for the cycling data's time-series component could involve experimenting with deeper Deep Learning architectures like LSTM and feature engineering (e.g., creating a BMI feature).

5. References

1. Obesity Levels Based on Eating Habits and Physical Condition. UCI Machine Learning Repository.
2. Scikit-Learn Documentation.