DevSoc Core Assignments 2025-26

Kamal Kumar Manchenella

June 2025

Contents

1 Ensemble Learning

1.1 Overview of Ensemble Learning

Ensemble learning is a machine learning paradigm where multiple base models are strategically generated and combined to solve a particular computational intelligence problem. The main idea is that a group of weak learners can come together to form a strong learner.

There are three main types of ensemble methods:

- **Bagging (Bootstrap Aggregating)**: Trains multiple versions of a model on different random subsets of the training data and averages their predictions. Example: Random Forest.
- **Boosting**: Trains models sequentially, where each model tries to correct the errors of its predecessor. Later models focus more on the hard-to-classify examples. Examples: Gradient Boosting, XGBoost.
- **Voting**: Combines predictions from multiple models. In *soft voting*, the average of predicted class probabilities is used; in *hard voting*, the most frequent class label among the base models is chosen.

These techniques increase robustness and accuracy by reducing variance, bias, or improving prediction stability.

1.2 Objective

The goal of this project was to predict an individual's obesity category based on health, lifestyle, and demographic indicators using machine learning. The problem was formulated as a multi-class classification task and addressed using various ensemble learning techniques to improve performance over traditional models.

1.3 Dataset and Features

The dataset used contains 2111 records and 17 columns, including features such as `Age`, `Height`, `Weight`, `Physical Activity (FAF)`, `Daily Water Intake (CH2O)`, and categorical variables like `Gender`, `Smoking`, and `Family History`. The target variable was `NObeyesdad`, representing one of seven obesity categories.

1.4 Preprocessing

All categorical features were label-encoded or one-hot encoded as appropriate. Numerical features were scaled using standardization. The data was then split into 80% training and 20% testing subsets. The target class distribution was balanced and stratified.

1.5 Models Used

We implemented and evaluated the following models:

- Logistic Regression (Baseline)
- Random Forest (Bagging)
- Gradient Boosting
- XGBoost
- Voting Ensemble (Soft)

Confusion Matrix Terminology

Term	Meaning
TP (True Positive)	Correctly predicted positive class
TN (True Negative)	Correctly predicted negative class
FP (False Positive)	Incorrectly predicted as positive
FN (False Negative)	Incorrectly predicted as negative

Table 1: Basic terms used in confusion matrices

1.6 Evaluation Metrics

Performance was evaluated using:

- Accuracy
- Macro-Averaged F1 Score
- Confusion Matrices (for class-wise comparison)

Metric Definitions

The following standard classification metrics were used:

- **Accuracy**:
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
- **Precision**:
$$Precision = \frac{TP}{TP + FP}$$
- **Recall**:
$$Recall = \frac{TP}{TP + FN}$$

- **F1 Score**:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

- **Macro F1 Score**:

$$Macro - F1 = \frac{1}{C} \sum_{i=1}^{C} F1_i$$

where C is the number of classes.

1.7 Results and Discussion

Figure 1 shows the performance comparison across all models in terms of Accuracy and Macro F1 Score.

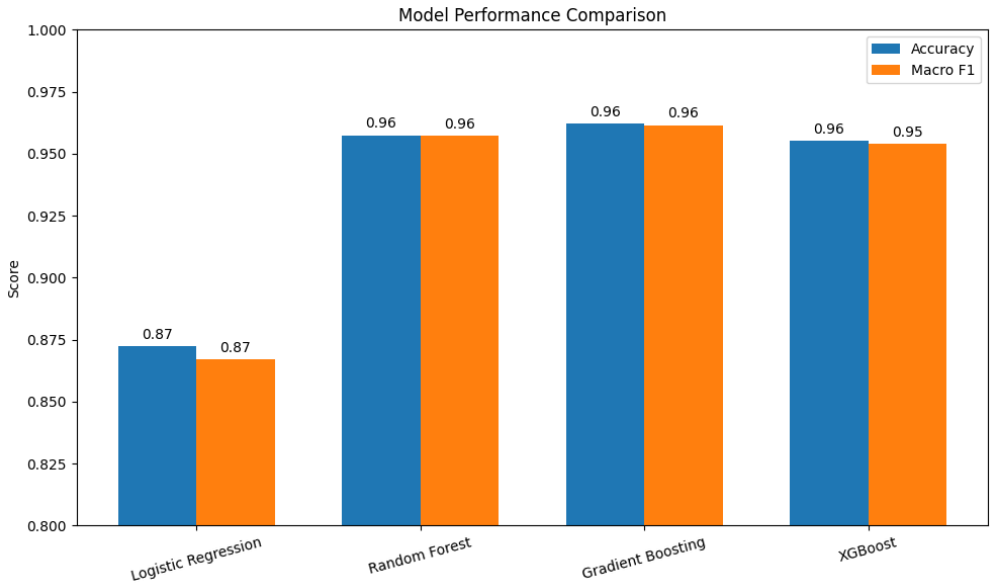

Figure 1: Accuracy and F1 Score Comparison across Models

Baseline (Logistic Regression): Achieved an accuracy of 87%. However, it struggled with mid-range classes like `Overweight_Level_I`, as seen in Figure **??**.
We evaluated four individual models before ensembling: Logistic Regression (baseline), Random Forest, Gradient Boosting, and XGBoost. All ensemble-based models outperformed the baseline, achieving high classification performance with 96% accuracy on average. Figure 6 presents their confusion matrices for comparison.

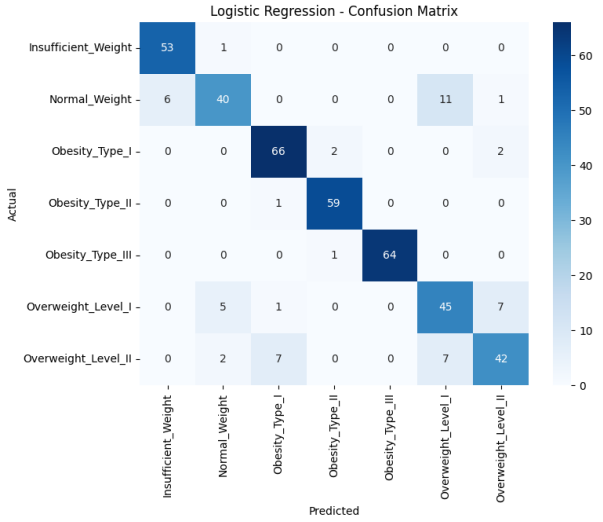

Figure 2: * Logistic Regression

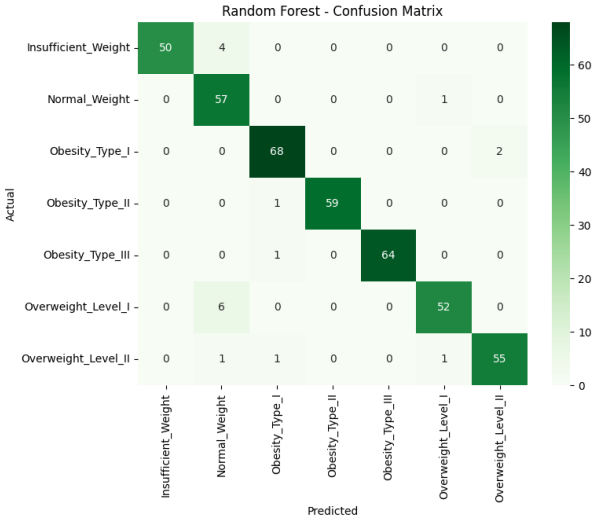

Figure 3: * Random Forest

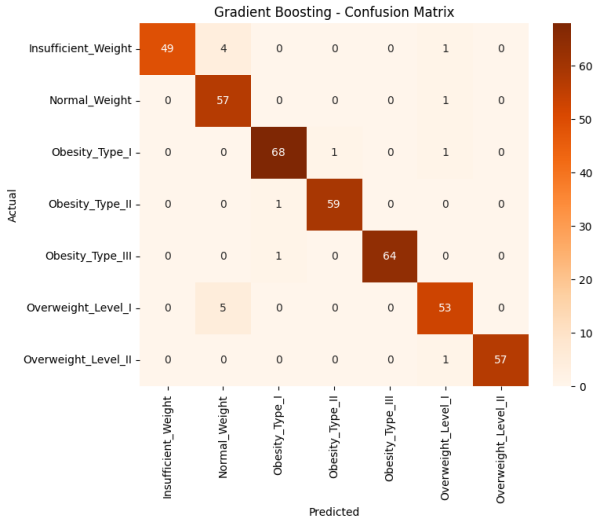

Figure 4: * Gradient Boosting

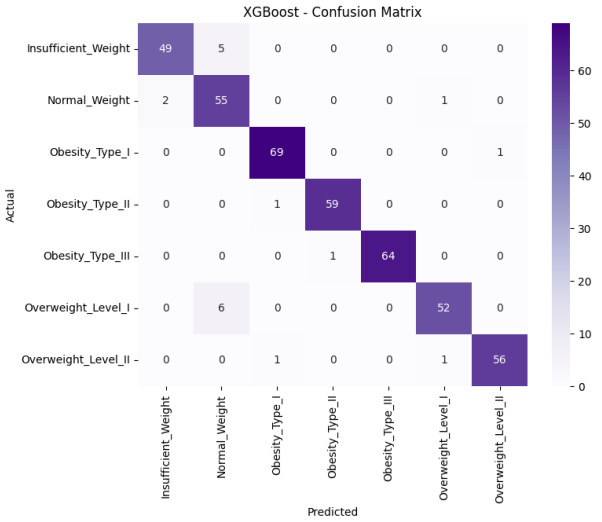

Figure 5: * XGBoost

Figure 6: Confusion Matrices for Singular Models: Logistic Regression, Random Forest, Gradient Boosting, and XGBoost

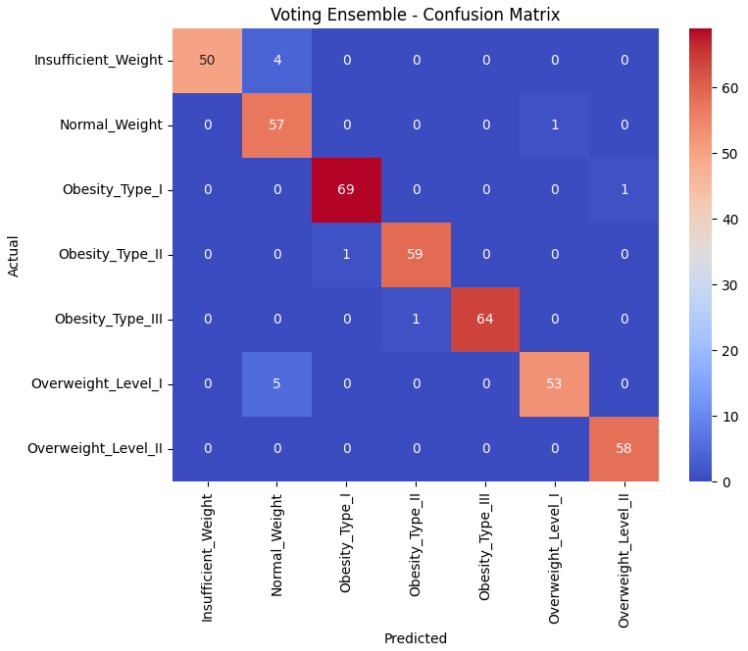

Figure 7: Voting Ensemble - Confusion Matrix

1.8 Conclusion

Ensemble models, particularly the Voting Classifier, significantly outperformed the baseline. This highlights the robustness of combining multiple algorithms in structured healthcare classification tasks. Future work could include hyperparameter optimization, feature importance ranking, and deployment using model serialization.