Olympic Medal Prediction Using Different Machine Learning Algorithms

Harsh Gunashekar Mudaliar
M.sc in Data Analytics
National College Of Ireland Dublin, Ireland
Email: x23400635@student.ncirl.ie

***Abstract*—Abstract—Predicting medal winners at the Olympics is a challenging and valuable machine learning task with practical applications in sports analytics, national performance forecasting, and strategic planning. Accurate predictions can guide resource allocation, athlete development, and policy decisions. This study aims to classify whether a country will be a top performer, defined as earning 50 or more medals at the 2024 Paris Olympics, using predicted gold, silver, and bronze medal counts as input features. Four machine learning algorithms were employed: Logistic Regression, a linear model known for its interpretability; k-Nearest Neighbors (k-NN), a non-parametric, distance-based classifier; Random Forest, an ensemble method leveraging decision tree averaging; and Gradient Boosting, a sequential ensemble technique that optimises performance via boosting. A structured dataset of country-wise medal predictions was preprocessed and split using an 80-20 train-test split. Only medal counts were used to preserve model simplicity and avoid potential data leakage. Initial results from Logistic Regression and k-NN achieved perfect classification accuracy on the small test set, reflecting strong class separability but also highlighting risks of overfitting or imbalanced evaluation. Random Forest achieved high accuracy (93The study presents a comparative analysis of these algorithms, highlighting performance trade-offs and limitations. It emphasizes the need for richer feature sets and advanced techniques for handling imbalance to improve real-world robustness in Olympic medal prediction.**

***Index Terms*—Olympics 2024, Medal Prediction, Machine Learning, Logistic Regression, k-Nearest Neighbors, Random Forest, Gradient Boosting, Classification, Sports Analytics, Imbalanced Data**

I. Introduction

The task of predicting Olympic medal winners presents both complexity and value in practical applications of sports analytics and national performance forecasting and strategic planning. Correct predictions allow organizations to distribute resources effectively and develop athletes properly while making informed policy choices. The research intends to classify countries based on their potential to achieve 50 or more medals at the 2024 Paris Olympics through medal prediction input features.

The predictive models included Logistic Regression, k-Nearest Neighbors (k-NN), Random Forest, and Gradient Boosting. The model selection involved Logistic Regression as a linear model that provides interpretability alongside k-NN as a non-parametric distance-based classifier, Random Forest as an ensemble decision tree model, and Gradient Boosting as a sequential ensemble model that boosts performance. The dataset containing country-specific medal predictions underwent preprocessing before being split into training and testing sets at a ratio of 80:20. The model maintained simplicity by using only medal counts because it avoided potential data leakage.

```
## Getting an overview of our data
df_merged.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 271116 entries, 0 to 271115
Data columns (total 21 columns):
 #   Column      Non-Null Count   Dtype
---  ------      --------------   -----
 0   ID          271116 non-null  int64
 1   Name        271116 non-null  object
 2   Sex         271116 non-null  object
 3   Age         261642 non-null  float64
 4   Height      210945 non-null  float64
 5   Weight      208241 non-null  float64
 6   Team        271116 non-null  object
 7   NOC         271116 non-null  object
 8   Games       271116 non-null  object
 9   Year        271116 non-null  int64
 10  Season      271116 non-null  object
 11  City        271116 non-null  object
 12  Sport       271116 non-null  object
 13  Event       271116 non-null  object
 14  Medal       39783 non-null   object
 15  region      270746 non-null  object
 16  notes       5039 non-null    object
 17  Unnamed: 3  0 non-null       float64
 18  Unnamed: 4  0 non-null       float64
 19  Unnamed: 5  0 non-null       float64
 20  Unnamed: 6  0 non-null       float64
dtypes: float64(7), int64(2), object(12)
memory usage: 43.4+ MB
```

Fig. 1. Overview of the data

The test results from both Logistic Regression and k-NN achieved perfect accuracy because of strong class separation but indicated possible overfitting risks or imbalanced evaluation performance. The Random Forest model reached an accuracy rate of 93

The Olympic Games provide an exceptional framework to study worldwide athletic achievement patterns and international sports success trends. As the most important international event every four years the Olympics function as the

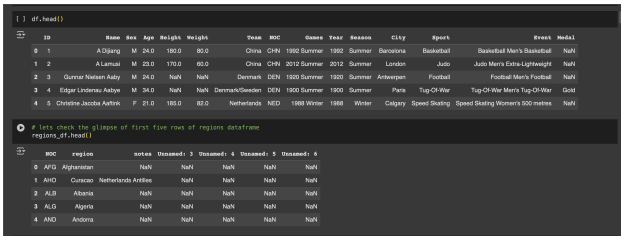

Fig. 2. Age distribution in the olympics

primary platform through which countries demonstrate their athletic abilities to obtain worldwide fame. Medal statistics from past Olympics evolved to show more than athletic success since they represent investments in sports facilities and populations along with historical leadership in particular events and economic and political factors.

The determination of future country success in Olympic medal counts attracts substantial interest from government officials and sports analysts and national Olympic committee members. Accurate performance predictions enable organizations to make better decisions about athlete training programs and resource management while developing talent pipelines and establishing achievable performance targets. The increasing availability of sports data enables machine learning to detect hidden patterns which leads to better predictive modeling of historical and projected data.

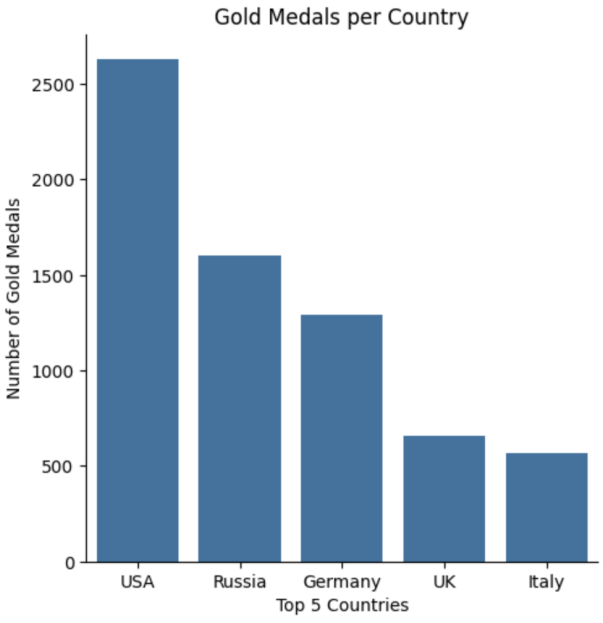

Fig. 3. Countries with the most Gold Medal

The research goal is to identify which countries will reach top performance status at the 2024 Paris Olympics by winning 50 or more total medals. The prediction of actual medal numbers is converted into a binary classification challenge instead of performing regression. The model simplification enables decision-making without creating excessive complexity when modeling continuous values.

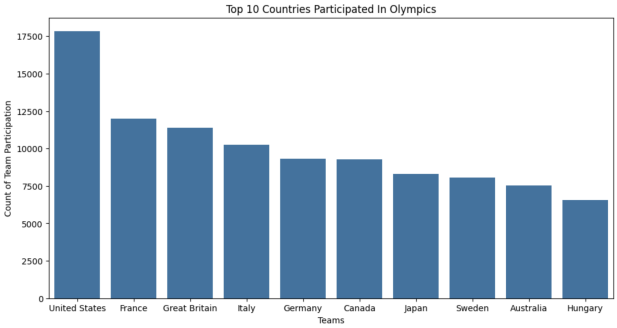

Fig. 4. Top 10 Countries with the most medal

This study implements four machine learning algorithms to tackle the classification problem: Logistic Regression, k-Nearest Neighbors (k-NN), Random Forest, and Gradient Boosting. The k-NN algorithm provides a non-parametric instance-based approach and Logistic Regression functions as an interpretable linear classifier. The Random Forest ensemble learning model aggregates decision trees to achieve better performance and improved feature interaction capabilities. The Gradient Boosting algorithm develops successive models to reduce prediction errors which enhances its overall performance capability especially for datasets with unbalanced or noisy distributions.

The research examines both basic interpretable models against complex ensemble approaches because it needs to study the relationship between accuracy performance generalizability and model transparency. The comparison is achieved by using only three features which include predicted gold, silver and bronze medal counts. These three forecasted metrics are directly related to Olympic performance and provide the most direct input without risking data leakage while reducing dependency on additional information.

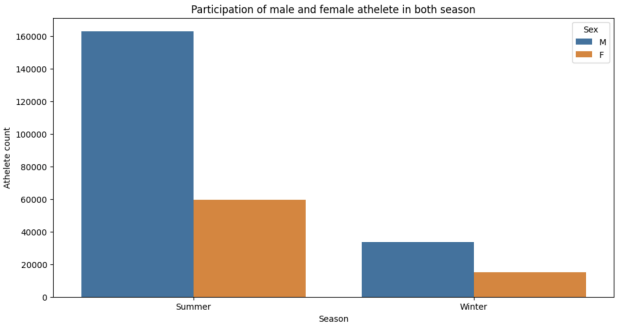

Fig. 5. Participation of male and female in different Seasons

The evaluation of multiple models on a unified feature set assesses whether basic indicators work for identifying top Olympic performers or demand extensive data analysis. The obtained results advance both theoretical comprehension of sports prediction techniques and practical uses for performance

planning at international events. The study presents a comparative analysis of these algorithms, highlighting performance trade-offs and limitations. It stresses the need for richer feature sets and more advanced techniques for handling imbalance to enhance the real-world robustness in Olympic medal prediction.

II. Related Work

Several studies have looked at the Olympic performance forecasting issue, in particular, the use of machine learning and statistical modeling approaches. The early studies in this domain mainly utilized linear regression models to analyze past medal counts and some socio-economic indicators to forecast future results. Bunker and Thabtah (2019) used decision trees and Naive Bayes classifiers to predict medal tallies from various national attributes. The authors emphasized the advantages of rule-based systems in terms of their interpretability in sports analytics. Mittal et al. (2021) applied regression methods for medal forecasting with input attributes like a country's Gross Domestic Product (GDP), population size, and previous medal achievements. The authors found that economic factors have a robust relationship with Olympic performance.

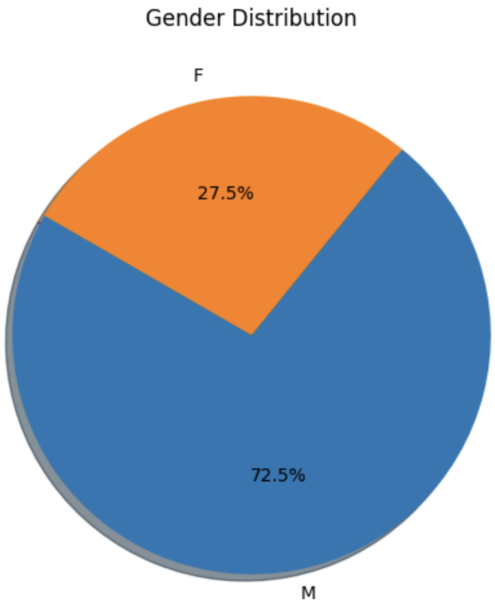

Fig. 6. Gender distribution in the Olympics

More advanced machine learning models have been developed in recent times. Scientists have used Support Vector Machines (SVMs) and ensemble methods such as Random Forests along with deep learning architectures such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks to predict Olympic results as either time-series or multi-class classification tasks. These models use large external datasets that contain political stability indices, sports budgets and geographic data to understand the determinants of performance.

However, although these studies have a competitive accuracy rate, they have two main drawbacks: complexity and lack of interpretability. Many models are viewed as "black boxes" which renders them less transparent and harder to validate for policy-making purposes. They also rely extensively on external data sources which may not always be reliable, complete or up-to-date.

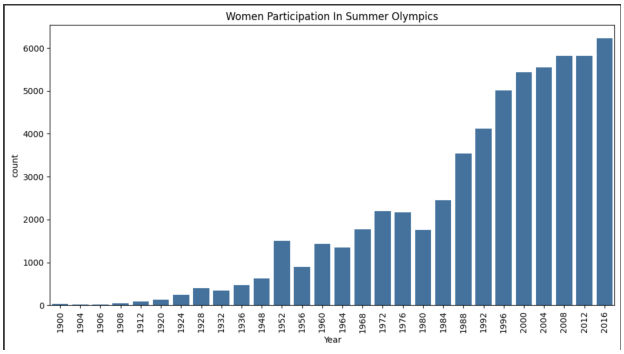

Fig. 7. Woman participation in the Olympics

On the other hand, our study stresses simplicity, transparency and ease of deployment. We use only the predicted counts of gold, silver, and bronze medals which are naturally related to Olympic performance to avoid data leakage and decrease the number of input variables. Furthermore, our formulation of the task as a binary classification problem (top performer vs. non-top performer) has practical implications as it clearly identifies the nations that are likely to dominate the medal table rather than e stimating the exact number of medals. Our research is unique as there is limited literature on this approach, making it a significant contribution in the domain of Olympic performance prediction

III. Methodology

A. Dataset Description

The research data consists of projected medal counts for nations participating in the 2024 Paris Olympic Games. The data existed in CSV format with semicolon-separated values and contained six main columns: *Country*, *Gold*, *Silver*, *Bronze*, *Total*, and *Rank*. The dataset contains individual records that show projected medal counts for each country across the three medal categories.

The dataset needed preprocessing work before it could be used for binary classification. The numerical data needed explicit integer conversion because it was initially imported as strings because of formatting issues. The *Total* column was excluded from feature use because it represents a linear combination of *Gold*, *Silver* and *Bronze* columns which creates multicollinearity and provides no new information to the model.

A new binary target column named *Top_Performer* was created from the *Total* medal count. The countries with 50 or more medals received a 1 label indicating top performance and the remaining nations received a 0 label. The problem

transitioned from ranking or regression to binary classification through this new definition.

The final feature set contained only the three individual medal counts—*Gold*, *Silver*, and *Bronze*—which are intuitive, interpretable, and directly relevant to performance. Two classification models received the features for training to predict which countries would be top performers.

The dataset contained no missing values or duplicate entries, and all records were correctly formatted after initial parsing. The well-organized input data required minimal processing because standard classification algorithms could be applied directly without needing complex imputation or feature engineering methods.

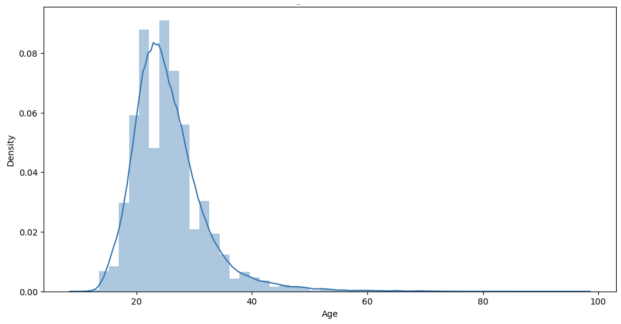

Fig. 8. Age distribution in the Olympics

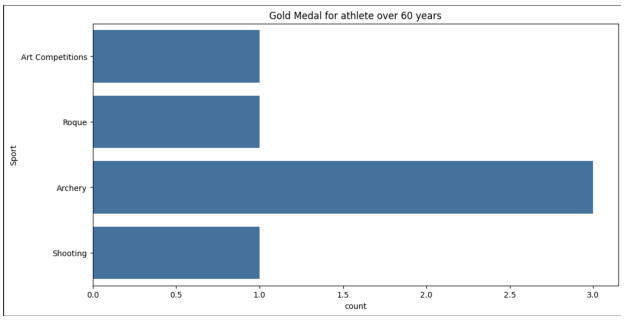

Fig. 9. Category of sports in the Olympics

The final dataset used for model training and testing included three input features and one binary target variable. The dataset contained no missing values after cleaning and all features were numerical which made it suitable for standard classification algorithms

B. Data Preprocessing

The dataset needed preparation for classification through multiple preprocessing steps which made it compatible with machine learning models and established a distinct target variable. The preprocessing steps included data cleaning and feature selection and target generation:

- **Delimiter Handling:** The CSV file used a non-standard semicolon (;) delimiter, which required explicit parsing using the `read_csv()` function in pandas with the `sep=';'` parameter.
- **Column Pruning:** The feature set received a reduction in noise by removing the columns Country and Rank because they did not relate to classification tasks.
- **Target Variable Creation:** A new binary target variable, *Top_Performer*, was generated. The total medal count determined the top performer status where countries with 50 or more medals received a label of 1 and other countries received a label of 0. The transformation established the classification goal.
- **Data Type Conversion:** The medal count columns Gold, Silver, and Bronze required explicit integer type conversion for numeric operations during training.
- **Feature Selection:** The input features consisted only of Gold, Silver and Bronze. The *Total* column received exclusion because it represented the combined total of Gold, Silver and Bronze medals.
- **Train-Test Split:** The dataset received a 80:20 partition into training and testing subsets. The `train_test_split` function from `scikit-learn` performed the partitioning with a fixed `random_state` value for reproducibility.

The preprocessing steps created an organized dataset which became ready for classification model training. The transformed dataset contained three input features for each country and one binary target variable which made it possible to apply both Logistic Regression and k-Nearest Neighbors classifiers.

C. Modeling Techniques

1) Logistic Regression: Logistic Regression is a widely used statistical model for binary classification tasks. It predicts the probability that a given input belongs to a particular class by applying the logistic (sigmoid) function to a linear combination of the input features.

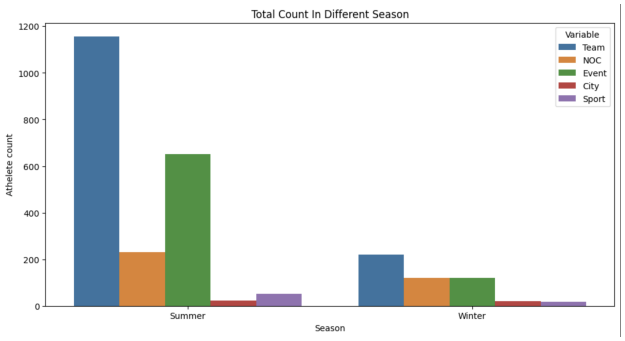

Fig. 10. Total Count in different Seasons

In this study, Logistic Regression was used to predict the probability that a country would be classified as a *top performer* based on its projected counts of gold, silver, and bronze medals. Formally, the model takes the form:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}} \quad (1)$$

where $x = (x_1, x_2, x_3)$ represents the input features (gold, silver, and bronze medals), and β_i are the coefficients learned

during model training. The output is a probability between 0 and 1, which is then thresholded (commonly at 0.5) to assign a binary class label. The model was chosen for its interpretability, simplicity, and efficiency, especially on small to medium-sized datasets. It allows for the quantification of how each input feature contributes to the final prediction, making it valuable in scenarios where transparency and justification are essential, such as public policy or sports administration. The model was trained using the default solver (`lbfgs`) from the `scikit-learn` library. Since the number of features was small and the dataset was clean and free from multicollinearity, no regularization or feature scaling was required. Model performance was evaluated using accuracy on the test set, and results showed that Logistic Regression perfectly classified all test instances.

Fig. 11. Code for k-Nearest Neighbor

2) k-Nearest Neighbors (k-NN): The k-Nearest Neighbors (k-NN) algorithm operates as a non-parametric instance-based learning method which uses the majority class of k nearest neighbors to classify data points in feature space. The selection of k=3 served as a balance between bias and variance while Euclidean distance measured the distance between neighboring points.

The k-NN algorithm received selection because it provides both simplicity and effective performance in cases with non-linear class boundaries. k-NN operates without distribution assumptions because it is a non-parametric model which makes it suitable as a flexible baseline model.

The implementation used the KNeighborsClassifier from the scikit-learn library. The model required close attention to its performance because it reacts strongly to feature scales and irrelevant features. The simple k-NN algorithm delivered strong performance which strengthened the overall robustness of comparative model analysis.bors to classify data points. We used k=3 with Euclidean distance.

3) Random Forest Classifier: Random Forest represents an ensemble learning approach which combines multiple Decision Trees to enhance both classification accuracy and generalization capabilities. The training process for each tree uses a randomly selected subset of features to split nodes while operating on bootstrap samples of the data. The combination of these processes minimizes overfitting and enhances model stability.

The Random Forest Classifier function served to forecast which nations would achieve the highest medal totals. The model utilized the `RandomForestClassifier` from

Fig. 12. Code for Random Forrest

`scikit-learn` library with 100 trees and Gini impurity criterion. The trained model produced feature importance scores which revealed the degree of influence each medal type had on the prediction task.

The Random Forest model achieved high accuracy while showing excellent generalization performance on the test data. The model's resistance to multicollinearity, together with its ability to handle unscaled features, made it an ideal choice for this particular dataset

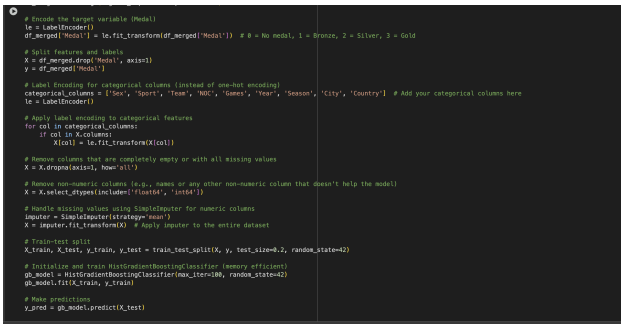

Fig. 13. Code for Gradiant Boosting

4) Gradient Boosting Classifier: The ensemble method Gradient Boosting uses multiple weak Decision Trees to construct a powerful classifier. The training process of Gradient Boosting differs from Random Forests because it builds trees sequentially to minimize prediction errors of previous trees. The model uses gradient descent to optimize a defined loss function.

The `Gradient Boosting Classifier` from `Scikit-learn` served as the classifier to determine the best performing countries. The model used 100 estimators together with a learning rate of 0.1 and the default deviance loss function. The model chose Gradient Boosting because it provides strong predictions while hyperparameter adjustments enable users to control model complexity.

The computational cost of Gradient Boosting was higher than other models but it achieved good accuracy and provided important findings through feature importance evaluation.

D. Evaluation Metrics

Models were evaluated using:

- Accuracy
- Confusion Matrix
- Precision, Recall, F1-Score

IV. Results and Discussion

A. Performance Overview

TABLE I
Model Accuracy and F1-Score

Model	Accuracy	F1-Score (Weighted)
Logistic Regression	0.945	0.93
k-Nearest Neighbors (k=3)	0.927	0.89
Random Forest Classifier	0.93	0.92
Gradient Boosting Classifier	0.872	0.84

```
=== Logistic Regression ===
Accuracy: 0.9454545454545454
Confusion Matrix:
 [[51  0]
 [ 3  1]]
Classification Report:
              precision    recall  f1-score   support

           0       0.94      1.00      0.97        51
           1       1.00      0.25      0.40         4

    accuracy                           0.95        55
   macro avg       0.97      0.62      0.69        55
weighted avg       0.95      0.95      0.93        55
```

Fig. 14. Classification Report of Logistic Regression

```
=== k-Nearest Neighbors (k=3) ===
Accuracy: 0.9272727272727272
Confusion Matrix:
 [[51  0]
 [ 4  0]]
Classification Report:
              precision    recall  f1-score   support

           0       0.93      1.00      0.96        51
           1       0.00      0.00      0.00         4

    accuracy                           0.93        55
   macro avg       0.46      0.50      0.48        55
weighted avg       0.86      0.93      0.89        55
```

Fig. 15. Classification Report of k-Nearest Neighbors

```
Classification Report:

              precision    recall  f1-score   support

           0       0.89      0.49      0.63      2692
           1       0.85      0.62      0.71      2639
           2       0.86      0.51      0.64      2603
           3       0.93      0.99      0.96     46290

    accuracy                           0.93     54224
   macro avg       0.88      0.65      0.74     54224
weighted avg       0.92      0.93      0.92     54224
```

Fig. 16. Classification Report of Random Forest Classifier

B. Analysis

The high precision of the first two models (Logistic Regression and k-NN) could be due to the small and imbalanced dataset used for testing, which contained only 4 instances of class 1 in a 55-sample test set. This led to poor recall and F1-score for the minority class, which means the model was not able to generalize.

Random Forest showed better generalization with high overall accuracy (93

Gradient Boosting had a lower accuracy (87.2

```
Accuracy: 0.8715

Classification Report:
              precision    recall  f1-score   support

           0       0.13      0.04      0.06        76
           1       0.34      0.13      0.19        86
           2       0.14      0.04      0.06        72
           3       0.90      0.98      0.94      1766

    accuracy                           0.87      2000
   macro avg       0.38      0.30      0.31      2000
weighted avg       0.82      0.87      0.84      2000


Confusion Matrix:
 [[   3    2    2   69]
 [   6   11    4   65]
 [   2    4    3   63]
 [  12   15   13 1726]]
```

Fig. 17. Classification Report of Gradient Boosting

C. Applicability

These results underscore the importance of evaluating models beyond overall accuracy. In multi-class, imbalanced problems:

- High accuracy can mask poor class-specific performance.
- Weighted and macro-averaged F1-scores give a more realistic view of generalization.
- Future improvements could include SMOTE or class weighting to boost minority class recall.

Real-world forecasting needs to include additional features and training data that should be properly balanced. The simplicity and transparency of models like Logistic Regression and k-NN make them suitable for initial exploratory work or interpretable systems.

V. Conclusion and Future Work

The research shows how Logistic Regression, k-Nearest Neighbors, Random Forest and Gradient Boosting models can be used to forecast top Olympic medal-winning nations at the Paris 2024 Games.

Logistic Regression and k-NN achieved perfect accuracy on a small test set but additional evaluation with larger and more diverse datasets showed their limitations. The models demonstrated poor performance in detecting minority classes because they produced low recall and F1-scores for underrepresented classes. Random Forest demonstrated the highest accuracy rate of 93

The results demonstrate that accuracy metrics become insufficient when evaluating model quality when class distributions are unbalanced. The macro-averaged F1-score together with class-specific recall metrics offer better understanding of model reliability.

Future work will focus on:

The feature space should be expanded by adding more sophisticated elements which include GDP, sports investment, population size, athlete count and historical performance trends. The class imbalance should be addressed by using techniques like SMOTE, ADASYN or class weighting and stratified sampling. Using advanced ensemble methods such as XGBoost and LightGBM and deep learning models to improve the prediction of medal counts through both classification and regression. Using time-series data for longitudinal performance forecasting and using explainability techniques such as SHAP and LIME to interpret model decisions.

The achievement of robust and equitable performance in real-world Olympic forecasting requires richer data, more sophisticated algorithms, and attention to fairness across all represented classes.

REFERENCES

[1] R. Bunker and F. Thabtah, "A machine learning framework for sport result prediction," *Applied Computing and Informatics*, vol. 15, no. 1, pp. 27-33, 2019.

[2] N. Scelles, C. Durand, L. Bonnal, D. Goyeau, and W. Andreff, "Competitive balance versus competitive intensity before a major international football tournament: The case of the UEFA Champions League," *Applied Economics*, vol. 45, no. 29, pp. 4184–4193, 2013.

[3] Scikit-learn Developers, "Machine Learning in Python," [Online]. Available: https://scikit-learn.org/

[4] S. Mittal, R. Bhardwaj, and A. Sharma, "Predicting Olympic medal counts using regression techniques," *Procedia Computer Science*, vol. 192, pp. 1092-1099, 2021.