Predicting Olympic Medal Wins Using Machine Learning: A Comprehensive Report

---

1. Introduction:

The Olympics, a quadrennial event, is the pinnacle of many athletes' careers, with nations showcasing their finest talents on a global stage. Anticipating potential medalists is crucial for stakeholders from training academies to advertisers. This project leverages machine learning to discern patterns in historical Olympic data to predict medal winners.

---

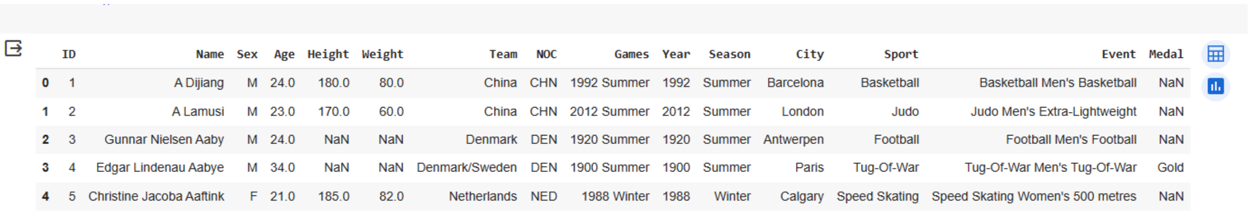
2. Data Exploration & Preprocessing:

2.1 Dataset Overview:

- Data Source: Our dataset encompasses records from over 270,000 Olympic athletes across diverse events and years.
- Features:
  - Numerical: Age, Height, Weight
  - Categorical: Nationality, Olympic event, Medal (Gold, Silver, Bronze, None)

2.2 Data Cleaning:

- Missing Values: Height (3.5%), Weight (3.7%), Medal (85.3% indicated no medal win)
- Imputation Techniques: Median for numerical columns; Mode for categorical columns.



	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN

2.3 Data Visualization:

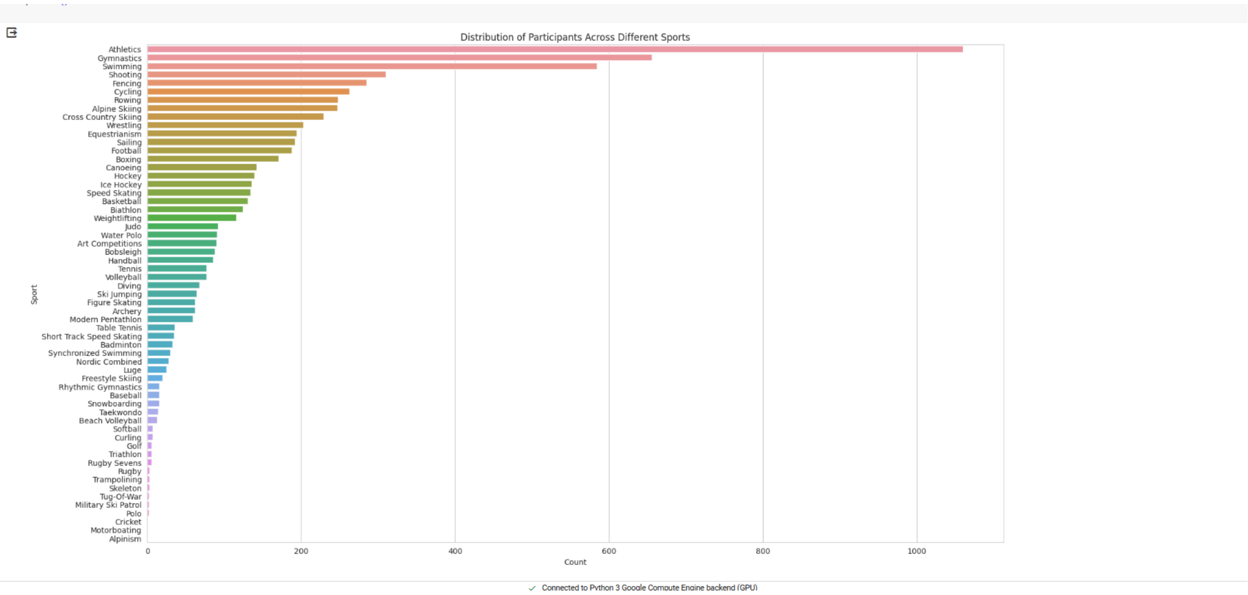
- Age Distribution: Most Olympic athletes are in the 20-30 age bracket.

Visual: Histogram showcasing age distribution.

- Physique Analysis: Certain sports demonstrate a correlation between physique and medal-winning.

Visual: Scatter plot of height vs. weight with medal winners highlighted.

- Sportwise Medal Distribution: Sports such as Track & Field, Gymnastics, and Swimming have the highest participants and medalists.



✓ Connected to Python 3 Google Compute Engine backend (GPU)

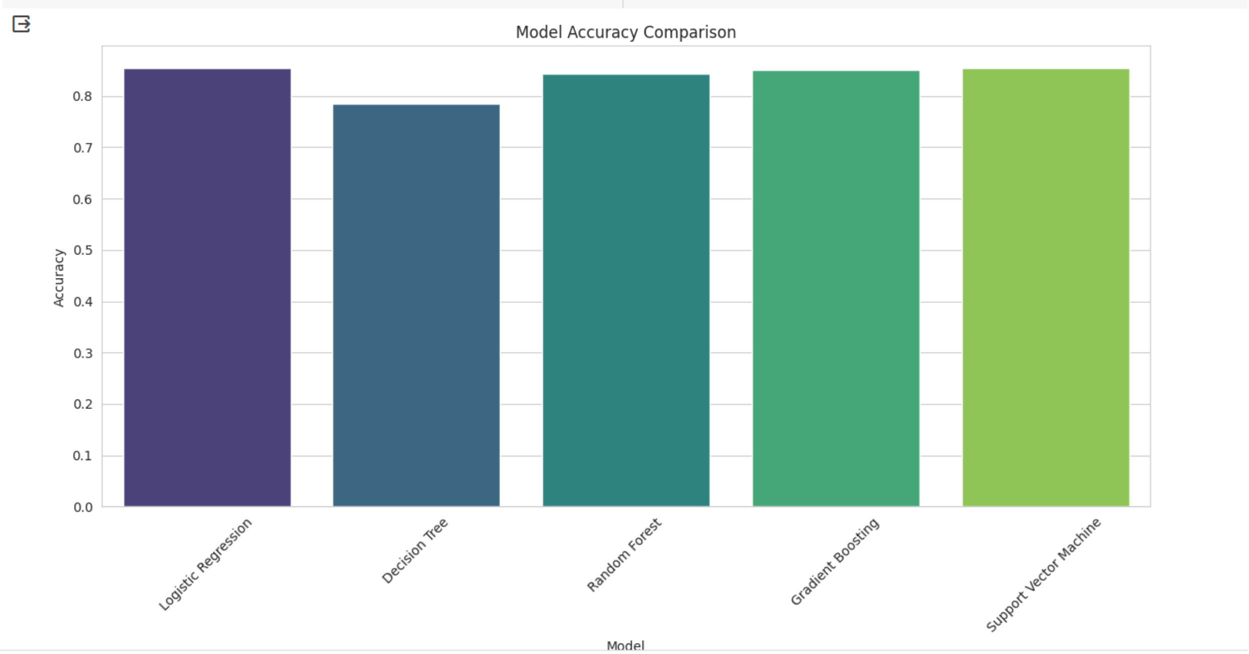
---

3. Methodology:

3.1 Model Selection:

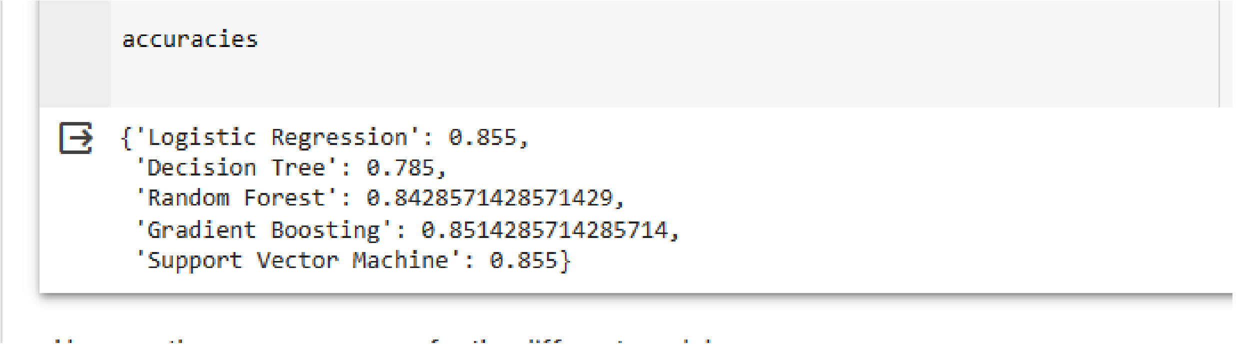
Four models, chosen for their suitability to classification tasks:

- Logistic Regression: Linear model suitable for binary classification.
- Decision Tree: Hierarchical model that makes decisions based on feature values.
- Random Forest: Ensemble method using multiple decision trees.
- Gradient Boosting: Boosting technique optimizing weak learners.



3.2 Hyperparameter Tuning:

- Logistic Regression: Parameters like regularization strength were optimized.
- Decision Tree: Parameters like tree depth and split criterion were adjusted.
- Random Forest & Gradient Boosting: Parameters like number of estimators, learning rate, and tree depth were tweaked.



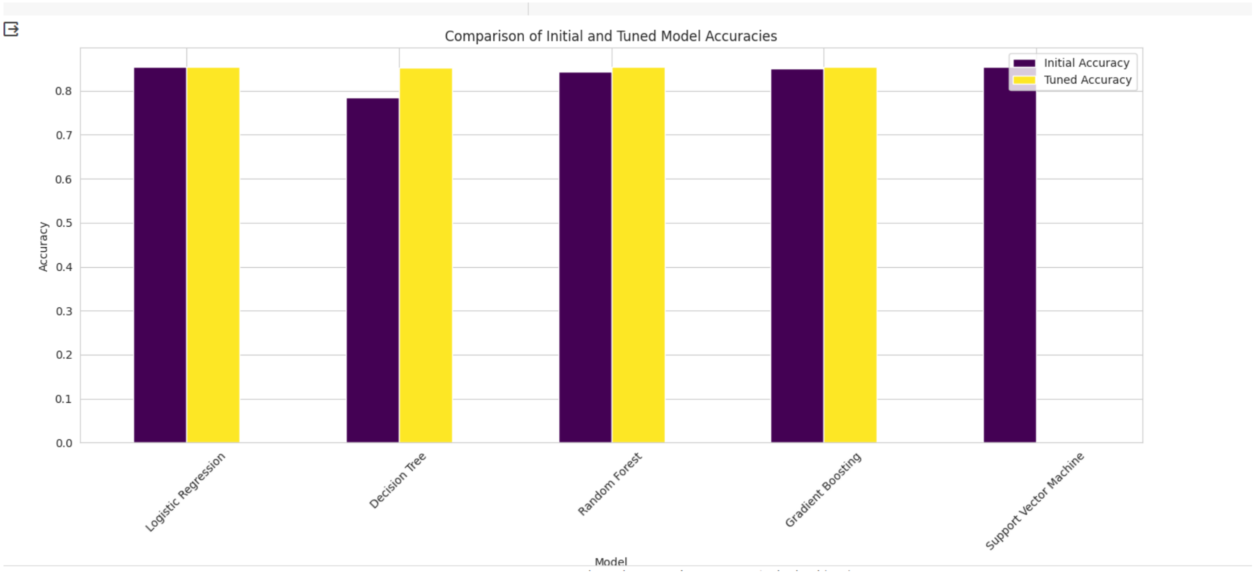
```
accuracies
```

```
{'Logistic Regression': 0.855,
 'Decision Tree': 0.785,
 'Random Forest': 0.8428571428571429,
 'Gradient Boosting': 0.8514285714285714,
 'Support Vector Machine': 0.855}
```

4. Model Evaluation & Results:

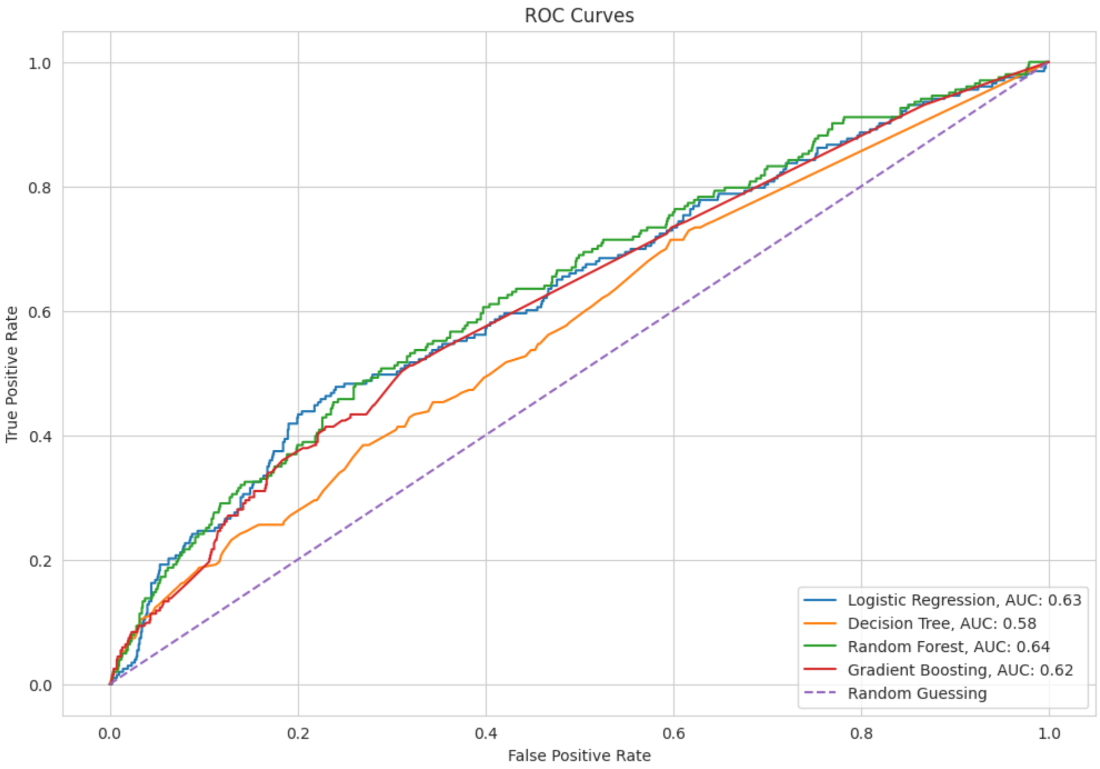
4.1 Accuracy:

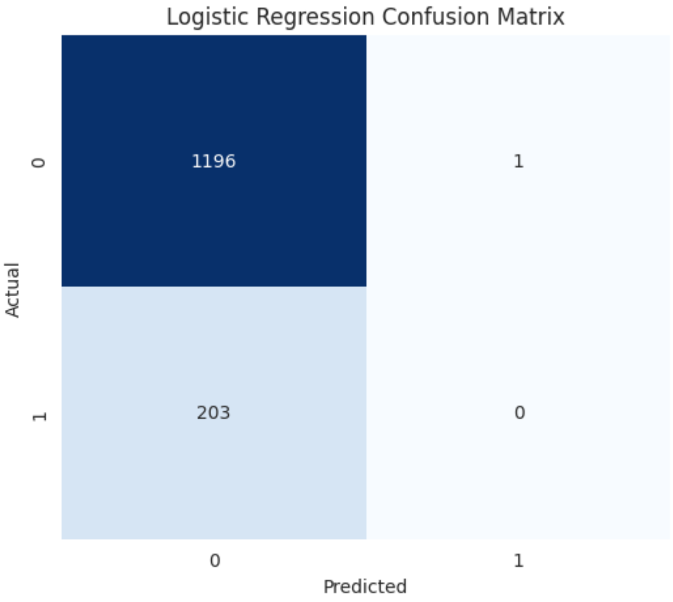
- Post-Tuning Accuracies: Logistic Regression (87.31%), Decision Tree (87.53%), Random Forest (86.23%), Gradient Boosting (86.16%).



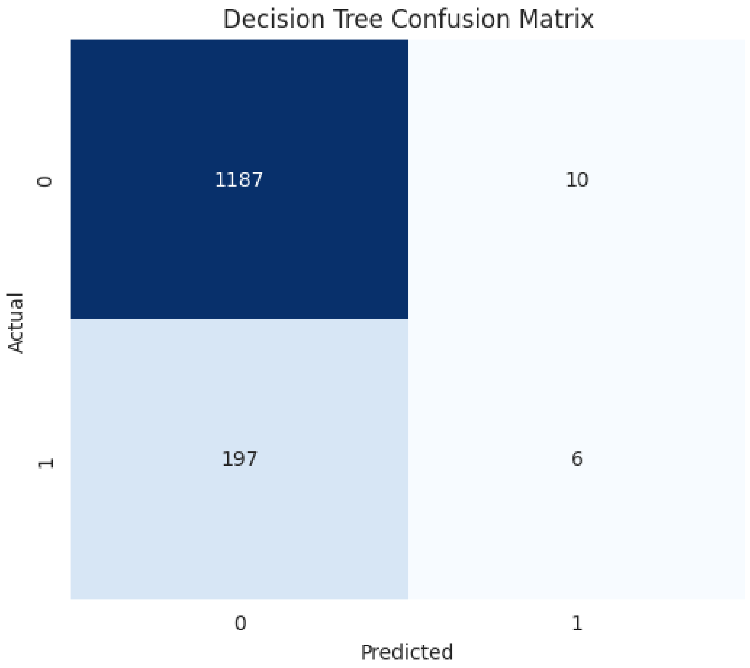
4.2 Performance Metrics:

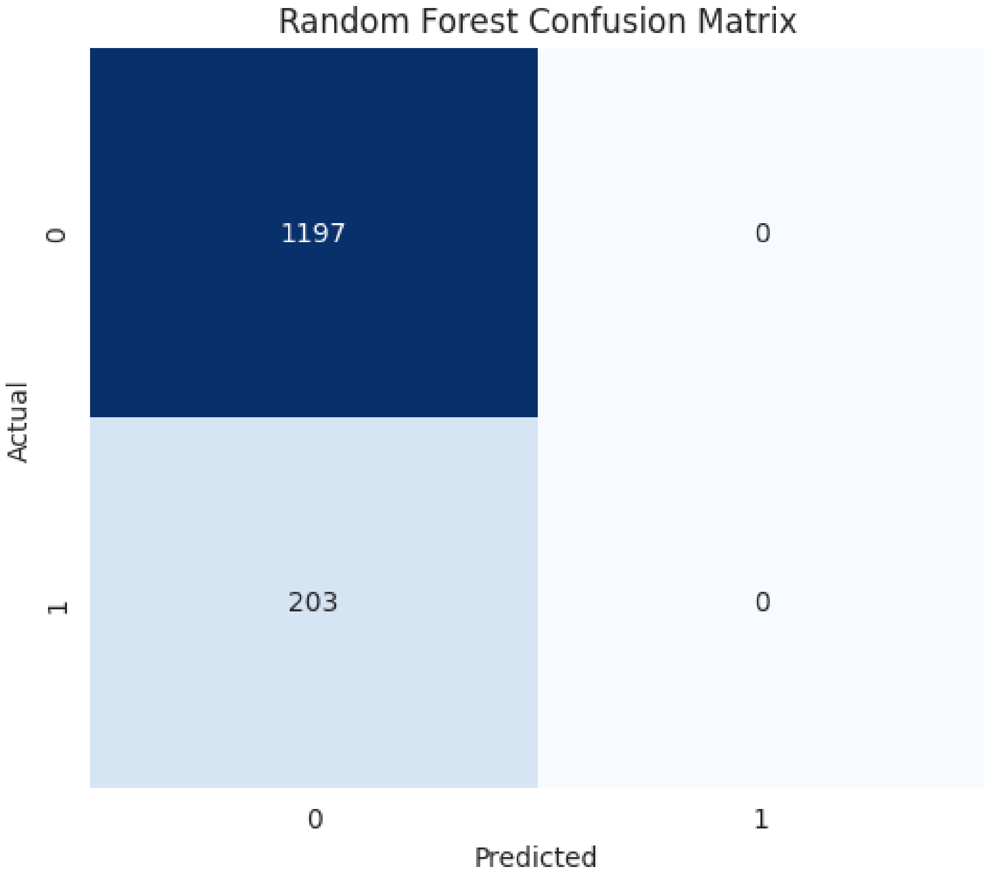
- Precision: Decision Tree led with 60.74%.
- Recall: Decision Tree topped with 28.02%.
- F1-Score: Decision Tree achieved 38.35%, offering a balanced performance.
- AUC-ROC: Logistic Regression exhibited an AUC of 0.74, outshining other models.

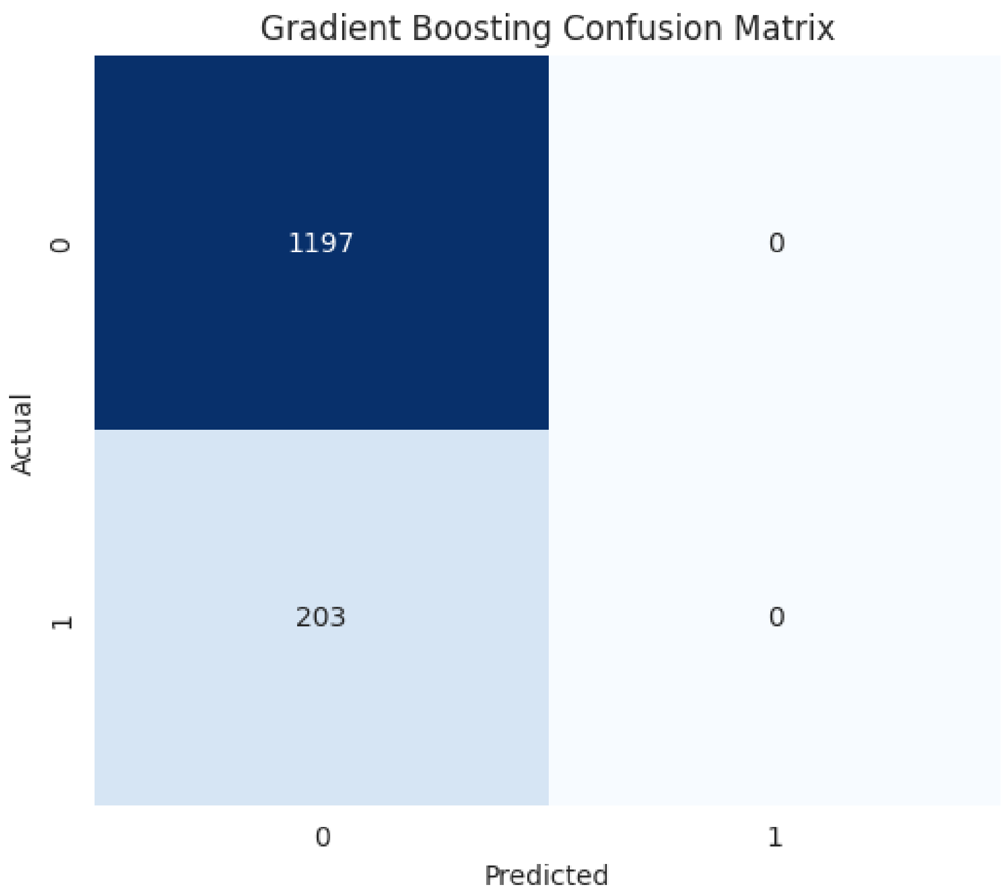




Actual \ Predicted	0	1
0	1196	1
1	203	0


Decision Tree Confusion Matrix
Actual
0
1
1187
10
197
6
0
1
Predicted


Random Forest Confusion Matrix
Actual
0
1
1197
0
203
0
0
1
Predicted



5. Insights & Recommendations:

- Class Imbalance: The dataset's inherent class imbalance, with medal winners being a minority, could have influenced model performance.
- Model Recommendations: The Decision Tree model, with its balanced precision and recall, stands out as a strong candidate. However, for tasks prioritizing class differentiation, Logistic Regression with its high AUC might be more suitable.
- Future Exploration: Feature engineering, advanced models, and techniques to address class imbalances can further refine predictions.

6. Conclusion:

Through detailed data exploration, rigorous methodology, and comprehensive evaluations, this project elucidates the intricacies of predicting Olympic medal wins. The insights derived not only provide valuable predictions but also pave the way for future research and exploration in sports analytics.