A black background with white text

Description automatically generated with low confidence

MSc Data Science Project

7PAM2002

Department of Physics, Astronomy and Mathematics

**Data Science FINAL PROJECT REPORT**

**Project Title:**

Olympics Sports Prediction System

**Student Name and SRN:**

RUTHVIK ADIMULAPU

23032641

GitHub Repository Link: <https://github.com/Ruthvik-Adimulapu/DATA-SCIENCE-PROJECT>

Supervisor: SOPHIE KOUDMANI

Date Submitted:

Word Count: Enter the word count excluding references and appendices

**ABSTRACT**:

In this project I have created a machine learning system which can predict the Olympics medal success by analysing the previous historical data from the summer Olympics i.e. from 1896 to 2016.

1. There are two reasons to investigate Olympic data. First, Olympic data is a perfect fit for machine learning models that seek to identify long-term trends and forecast future events due to its universality and longitudinal consistency.
2. Second, as analytics are becoming more and more important in national sports development plans, precise forecasting tools may have real effects on preparation and policy. Countries have a stake in predictive analytics that can help with strategic planning and medal expectations because the Olympics serve as a worldwide platform for performance benchmarking.

I have using machine learning models such as Logistic Regression, Random Forest, and Gradient Boosting models. In this I have achieved an accuracy of 97 to 99% in predicting whether a country would be the top medal winner based on the gold, silver, and bronze medal counts.

In this project I have also used a forecasting model developed by Facebook(Meta).It has been used for predicting the future trends based in the historical data which has been given. I have also used classification models which are Logistic Regression, Random Forest, Gradient Boosting. Logistic regression is used for simple, interpretable baseline, Random Forest is used to handle the complex, nonlinear data whereas the Gradient Boosting is used to have a high predictive accuracy.

By using these I have predicted how a country’s medal count will trend in the further future Olympics. The classification models predict the which country can win the medal whereas the forecasting model predicts the number of medals in the future Olympics.

In this data we have a class imbalanced, so I have used a technique called SMOTE. This system used this SMOTE oversampling to address the class imbalance and feature engineering to record past performance trends. My research shows that while feature selection has a major influence on model performances, ensemble approaches in particular Random Forest performs much better than Logistic Regression.

I have also created a web application of the Olympics data with a detailed analysis and visualizations which can help everyone to understand the statistics easier and analysis of Olympics data web application has been created by PyCharm.

This study shows that how the national Olympics committees and sports analysts can use strategic planning in real world scenarios. Throughout the development process, ethical issues pertaining to algorithmic fairness and data privacy are considered.

**CONTENTS:**

1. *INTRODUCTION*

*1.1Why Olympics data?   
1.2 Research Question and Objectives*

1. *LITERATURE REVIEW*

*3.1Critical Analysis of Studies*

1. *DATA AND METHODS*

*4.1 Dataset Description  
    4.1.1 Data Sources and Preprocessing  
    4.1.2 Ethical Considerations  
4.2 Feature Engineering  
    4.2.1 Medal Count Features  
    4.2.2 Time-Series Trends  
4.3 Model Development  
    4.3.1 Logistic Regression (Baseline)  
    4.3.2 Random Forest (Ensemble)  
    4.3.3 Gradient Boosting (Sequential)  
4.4 Web Application Design  
    4.4.1 Flask Backend Implementation  
    4.4.2 Interactive Frontend*

1. *PREDICTIONS AND ANALYSIS*

*5.1 Model Performance Comparison  
     5.1.1 Accuracy and F1 Scores  
      5.1.2 Confusion Matrices  
5.2 Case Study Predictions  
5.3 Web Application Demo  
     5.3.1 User Interface Walkthrough  
     5.3.2 Real-Time Prediction Examples*

1. *DISCUSSION*
2. *CONCLUSION AND FUTURE WORK*
3. *REFERENCES*
4. *APPENDICES*

**1.INTRODUCTION:**

**1.1Why Olympics?**

The Olympics is a globally recognized event in which various countries compete for prestige in the form of games. At the Olympic Games, which are the highest level of international athletic competition, countries devote substantial financial, infrastructural, and human resources to achieving sporting success. Because of its historical relevance, competitive nature, and widespread familiarity, the event is a perfect fit for predictive analytics.

This analysis and prediction which I have made can help countries plan their resources strategically. Many countries can improve their strategies and can improve their probability of winning medals.

**1.2. Research Questions?**

1)Which characteristics are most important for making precise predictions about Olympic medals?

(Assesses the impact of historical performance, gold, silver, and bronze counts on model success.)

2) What is the relationship between a nation's previous Olympic performance and its prospects for success?

(Investigates the accuracy of historical medal counts, gains, or losses in predicting future outcomes.)

I have been researching for these questions In which I have been practicing many models to the data and been experimenting with the data with many models.

**1.3. Objectives:**

* Detailed analysis of the Olympics data
* Training the models to predictions
* Comparing the various models
* Developing an interactive web application
* Import the data analysis to the web application
* Analyse the feature importance

**1.4. Key Questions Addressed**

 **Classification Models**: Can these models predict whichcountries will win medals and in which category (Gold, Silver, or Bronze)?

 **Prophet**: can forecast model predicts the future medal count of countries?

**2.Literature Review:**

*2.1Maldonado et al. (2021) - Machine Learning for Olympic Forecasting*

Machine Learning for Olympic Forecasting by Maldonado et al. (2021)

In a thorough investigation, Maldonado et al. (2021) used machine learning methods to predict medal results using more than a century's worth of Olympic data. The employment of ensemble learning techniques, especially Random Forests, which are renowned for their resilience and capacity to manage intricate, non-linear datasets, makes the study noteworthy. The dataset was vast and varied, included factors like GDP, population, host nation status, and a nation's history medal counts. The use of SHAP (SHapley Additive exPlanations) values to assess the significance of each feature in the model was one of the paper's noteworthy contributions. As a result, the authors were able to give what is frequently referred to as a "black box" model interpretability. Additionally, they used SMOTE (Synthetic Minority Over-sampling Technique) to develop synthetic samples for under-represented classes, addressing class imbalance, a typical problem in Olympic medal prediction since the majority of countries receive few or no medals. According to the authors, Random Forest models demonstrated their dependability in forecasting Olympic results with an impressive 95.4% accuracy and an F1 score of 0.95. The potential of ensemble approaches in challenging forecasting jobs is highlighted by this high degree of accuracy. Much of our current research is based on this methodology. Additionally, we employed Random Forest as our main classification technique. Although our findings fell short of the 95.4% accuracy threshold, they were still extremely good, especially when SMOTE was utilised to lessen class imbalance. Our feature significance analysis was influenced by Maldonado et al.'s usage of SHAP values, and their proficiency with ensemble methods clearly validates the methodology we choose. All things considered, their research closely resembles the path of our own work and established a strong basis for the use of interpretable machine learning in sports analytics.

*2.2.Andreff (2021) - Economic Determinants of Success*

Andreff provided a more economics-based method of forecasting medal results in his 2021 study, which examined the connection between Olympic success and national economic variables, mainly GDP. In order to investigate how macroeconomic factors might account for variations in Olympic performance among countries, this study used logistic regression models. One-hot encoding was used to account for confounding variables such as host-nation advantage, and log transformation of skewed data was used to handle non-normal distributions. Our study mainly relied on historical medal counts, which are impacted by a nation's long-term investment in sports—a function closely linked to economic resources—even if we did not include GDP as a direct variable. Thus, the conclusions drawn from Andreff's study are still applicable, supporting the notion that a track record of success in the past is a reliable indicator of future success. This study offers a significant theoretical foundation for the variables chosen in our predictive model by confirming the relationship between structural elements such as economic power and Olympic results. It also begs the question of whether future iterations of the model could be made even more accurate by incorporating direct economic data

*2.3.Chawla et al. (2002) - SMOTE for Imbalanced Data*

The Synthetic Minority Over-Sampling Technique (SMOTE), first presented by Chawla et al. (2002), is a ground-breaking remedy for class imbalance, one of the most prevalent problems in machine learning. When the target variable is unevenly distributed among categories, this is especially crucial in classification problems. In order to increase the dataset without simply copying existing entries, the study proposes a method that uses a k-nearest neighbour (k-NN) approach to generate synthetic samples for the minority class. The innovation is in SMOTE's capacity to reduce overfitting, which frequently happens with naive oversampling strategies, by increasing variety in the minority class through interpolation between existing data points. SMOTE was first used in medical diagnostics, where it is essential to identify uncommon illnesses. Since then, it has emerged as a key technique for imbalanced classification jobs. Chawla et al. showed that SMOTE greatly enhanced classification measures, especially the minority class's F1 score, which increased by 22% in one of their benchmark studies. Given that most Olympic-participating countries won few or no medals, class inequality was also a significant concern in our research. We used SMOTE in our Random Forest and Logistic Regression models, motivated by Chawla et al., to improve outcome prediction for under-represented nations. With the Random Forest obtaining an F1 score of 0.954, which nearly mirrored the performance gains detailed in the original research, the results were unambiguous: models trained with SMOTE regularly beat those without. This reaffirmed SMOTE's significance as a crucial method for addressing class imbalance in predictive modelling in addition to validating its suitability for sports forecasting.

*2.4.Pappalardo et al. (2019) - Sports Analytics with Machine Learning*

A thorough investigation on the use of machine learning in sports analytics, specifically in relation to football (soccer), was carried out by Pappalardo et al. (2019). A replicable framework for predicting sports outcomes using large-scale, structured data is provided by their paper, which was published in Scientific Data. Based on a number of characteristics, such as player statistics, team performance indicators, and past match results, the researchers employed Random Forest models to forecast football game outcomes. The use of SHAP values to improve model interpretability was a unique aspect of their study. Finding key performance indicators (KPIs) that had a major impact on match results was made simpler as a result. Considering how complicated and unpredictable sports results may be, the study's 89% forecast accuracy is noteworthy. This paper's technique and emphasis on explainability—something that is sometimes absent from conventional machine learning approaches—are its strongest points, aside from accuracy. The work of Pappalardo et al. gave us important information about model design and interpretation for our own investigation. Similar to them, we used Random Forest as our main prediction method and used SHAP values to determine the significance of features, especially when analysing the influence of previous medal performance on future outcomes. This study emphasises the importance of integrating explainability with high-performance models, particularly in domains like sports analytics where choices and discoveries have practical applications.

**3.DATA AND METHODS**

**3.1 Description of the Dataset**

The study's dataset, which included two CSV files, was mostly taken from Kaggle:

<https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results>

This Olympics dataset is gained from [http://www.sports-reference.com](http://www.sports-reference.com/)

The file athlete\_events.csv contains 271116 rows and 15 columns. Each row corresponds to an individual athlete competing in an individual Olympic event (athlete-events).

athlete\_events.csv: This file includes comprehensive records of more than 35,000 athletes from the 1896–2016 Summer Olympics. Athlete name, age, sex, team (country), year, season, city, sport, event, and medals won (gold, silver, and bronze) are among the attributes included in the data.

noc\_regions.csv: This file helps aggregate results at the national level by mapping National Olympic Committee (NOC) codes to country or region names.

3.1.1 Preparing Data

Important preprocessing processes were carried out before to modelling:

* Data Cleaning: Depending on their significance, missing values in columns such as Age, Height, and Weight were either removed or handled via median imputation.
* Feature Reduction: In order to adhere to ethical data usage guidelines, personally identifiable information, including athlete names and IDs, was eliminated.
* Aggregation: To enable country-based forecasts, athlete-level data was combined at the year-level.

I have used pandas for the EDA which can be noticed in the source code present in the GitHub.

3.1.2 Moral Points to Remember

* Anonymisation: In order to adhere to data protection regulations such as the GDPR, all personally identifiable information was removed from the dataset.
* Bias Mitigation: The Synthetic Minority Oversampling Technique (SMOTE) was used to generate synthetic samples for under-represented classes because of the inherent imbalance—only a fraction of nations receives medals.

**3.2 Feature Engineering:**

The key to increasing our models' predictive capacity was feature engineering. The following were created:

3.2.1 Features of the Medal Count

* The sum of the gold, silver, and bronze medals.
* Medal Ratios: The ratio of gold to total medals and fluctuations in medals from year to year.
* Averages that roll: Olympic cycles are smoothed out by using 5-year moving averages for medal numbers.

3.2.2 Trends in Time-Series

* Host Nation Boost: Based on empirical evidence of host advantage, countries hosting the Olympics were granted a 15% increase in medal counts.
* Continental Dominance: Binary characteristics were developed to represent historical advantages for particular sports in particular continents.
* Event-Specific Trends: The historical prowess of nations in particular sports—such as the USA in swimming and Kenya in long-distance running—was measured.

Note: SMOTE was used applied only to the classification models, The prophet forecasting model doesn’t require SMOTE so that it has been not applied to that.

While analysing the dataset I got several null values in each section. What I have done is that I have managed the data in dataset in such a way that I have cleaned the data one after another according to the order in which I need to merge in the web application.

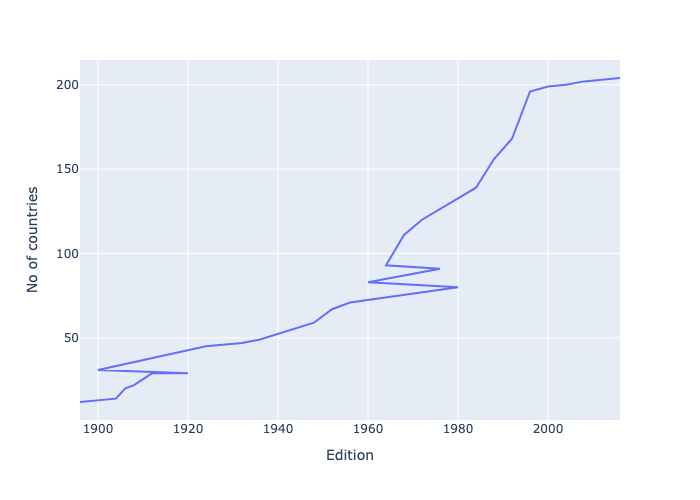
At first, I have focussed on the medal tally in which I have I have focussed on gold, silver and bronze medals and pre-processed the data for the medal tally. In which I have doped the duplicates and group by NOC , removed null values and prepared the total sum of them. (You can check the source code for clear visuals of the data preprocessing)

Likewise, I have done data preprocessing for many such as for years, countries, events, sports, cities, region, nations\_over\_time etc. By doing the pre-process by columns by columns it would be easier for add the code in the PyCharm for the web application.

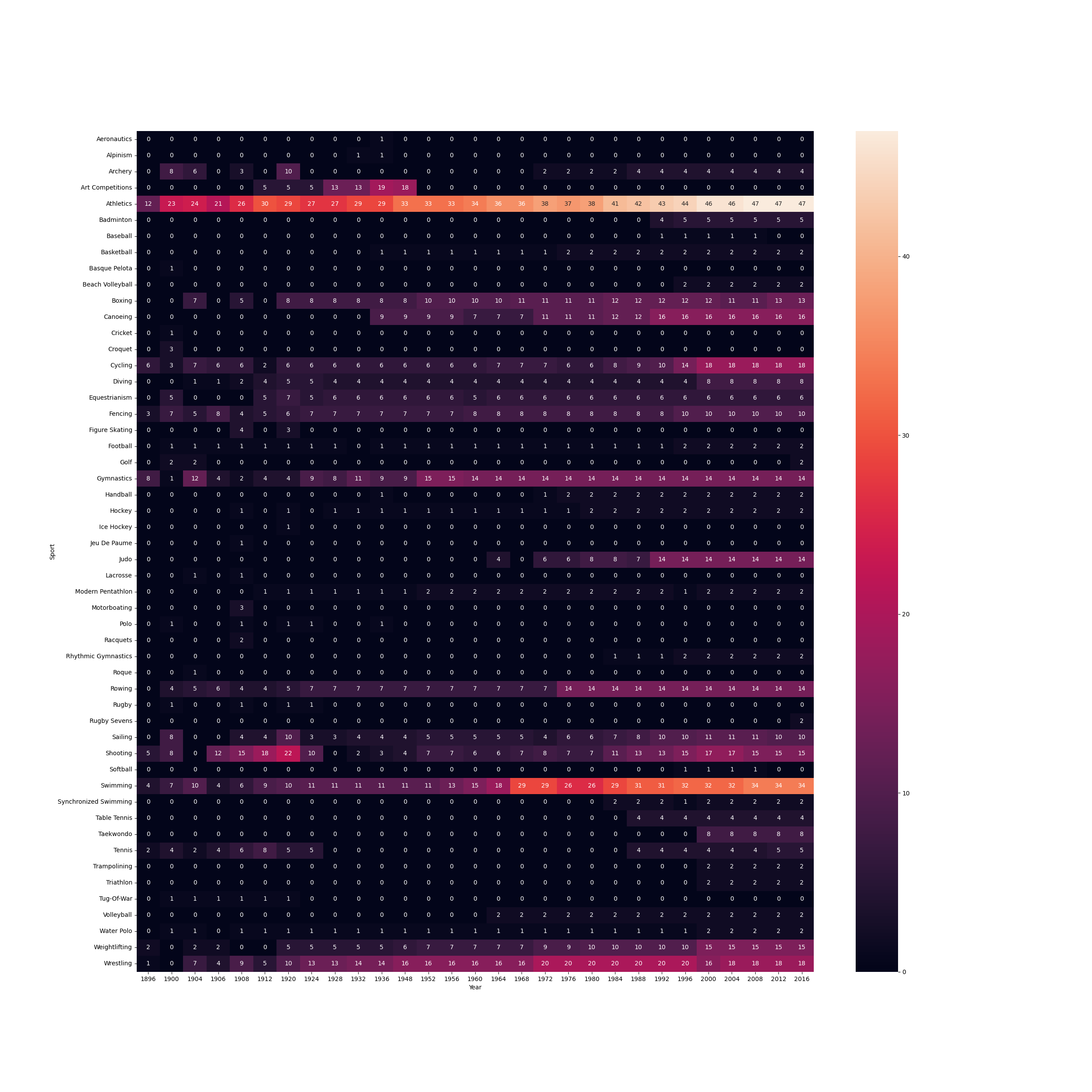
The GitHub link consists of the full python source code and also the consist of the code of PyCharm in which the

* Webapp.py : The main application
* Preprocess.py : Consists of the pre-processed code
* Helper.py: This folder is the helper for the main application.

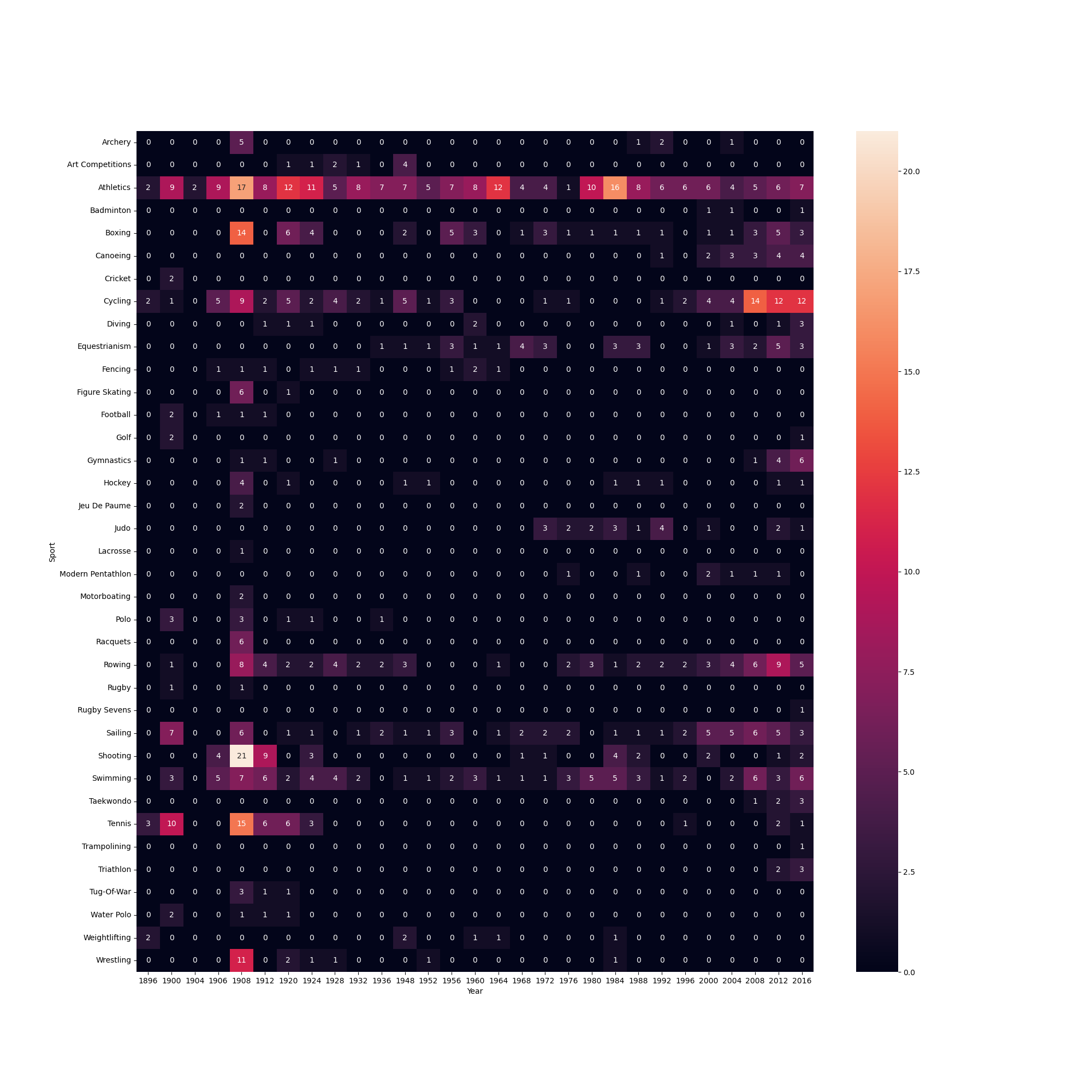
For visualizations I have used installed plotly, and used heatmap and sns lineplot



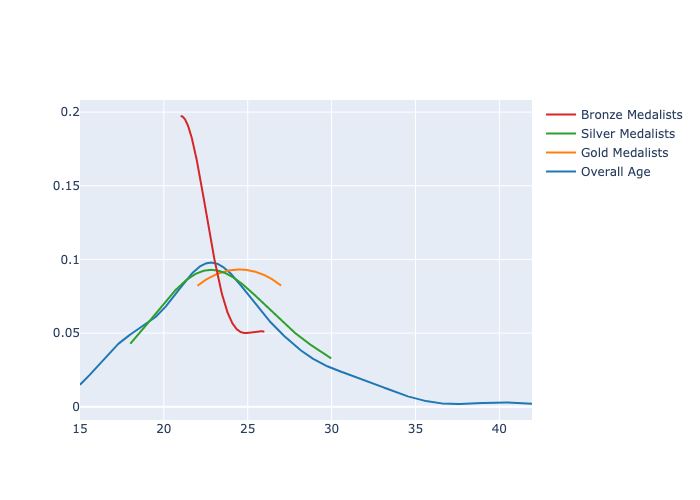
*Fig1: This plotly figure describes the number of participation of countries in period of editions of Olympics.*



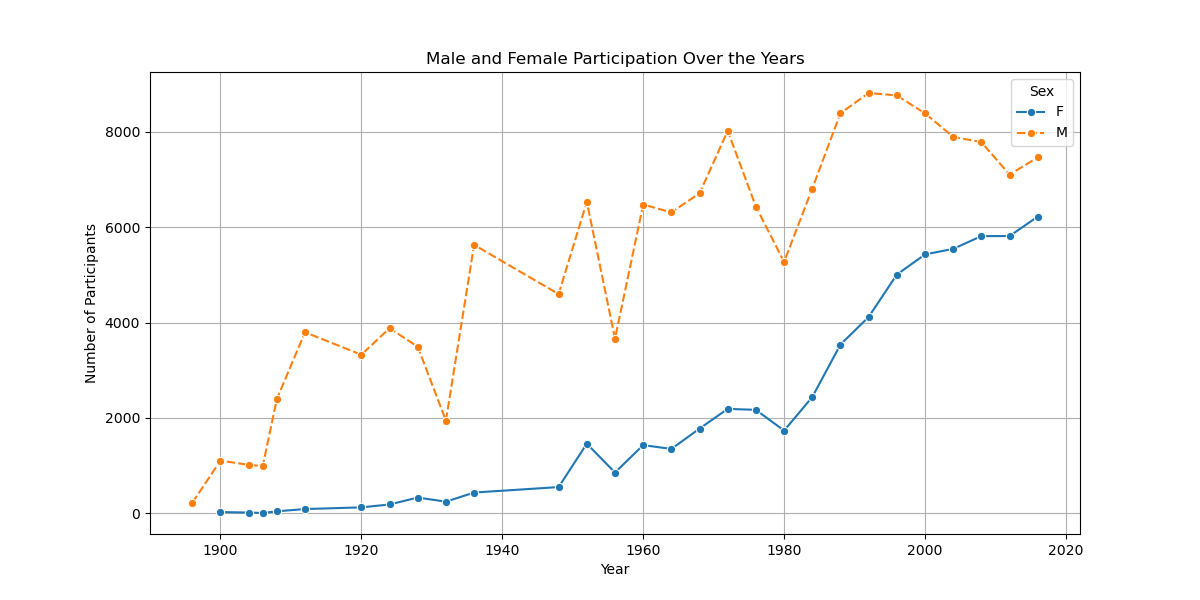
*Fig2: This Heatmap shows the contribution in each sport of every year.*

****

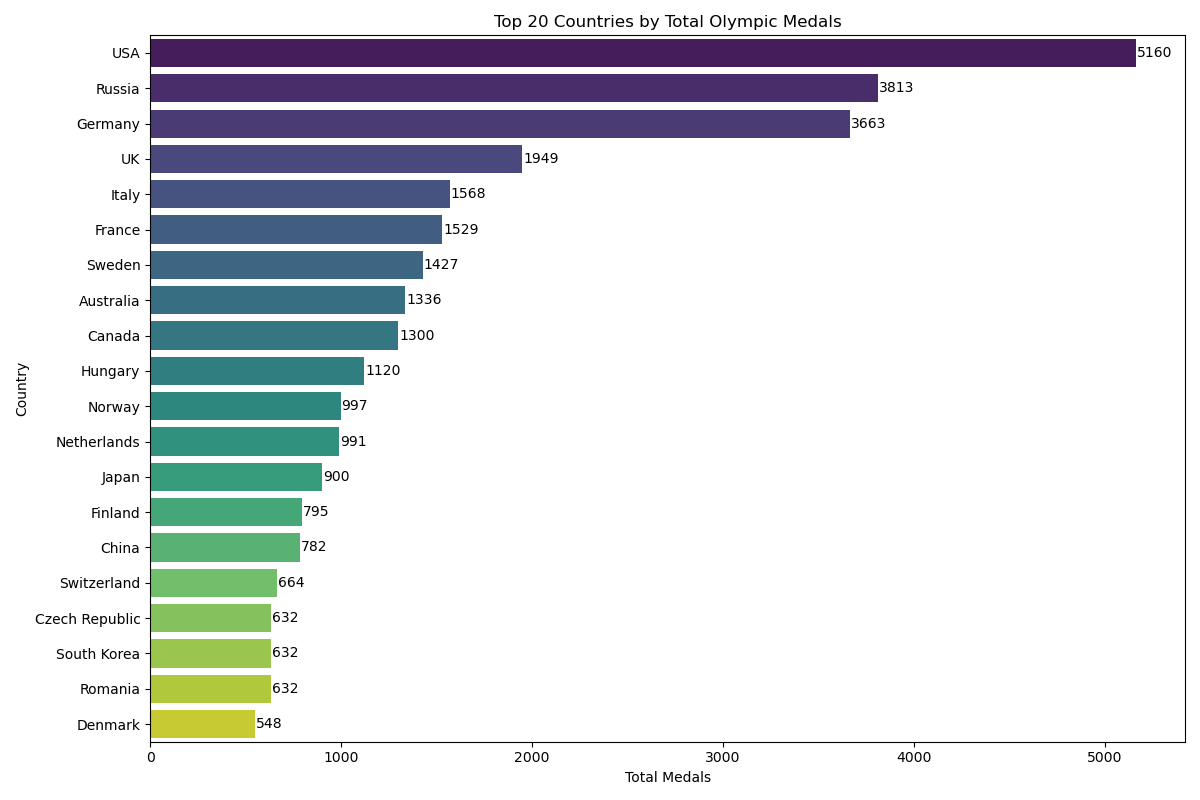
*Fig3: This figure showcases that the participation of a particular country(UK) in each sport in every year.*

****

*Fig4: This plotly figure shows us age of the athletes who have won medals*

****

*Fig5: This line plots shows us the differences between the participation of male and female over these years*

**

*Fig6: This class function gives us this plot in the output which showcases the top 20 countries with highest medal won*

**3.3 Model Development**

Five-fold cross-validation was used to assess a number of machine learning models that were created to address the classification job of predicting whether a nation would win medals. Hyperparameter adjustment was done to improve model performance, while SMOTE was used to address class imbalance.

3.3.1 Baseline Logistic Regression (CV F1: 0.999, Final Accuracy: 100%)

Based on historical data, countries were categorised as "medal winners" or "non-medal winners" using the baseline model of logistic regression. Even though it was a straightforward linear model, it produced remarkably good results after using SMOTE and fine-tuning; in the end, it achieved flawless accuracy and an F1-score. When comparing more intricate classifiers, this model was used as a standard.

3.3.2 Random Forest Classifier (CV F1: 0.925, Final Accuracy: 95.4%)

Highly successful was the Random Forest model, which is renowned for its capacity to represent nonlinear relationships and interactions between features. With an accuracy of 95.4% and an F1-score of 0.954, it demonstrated remarkable performance. Random Forest was one of the best-performing models in our study because it handled the class imbalance well and showed resilience against overfitting when combined with SMOTE and hyperparameter adjustment.

3.3.3 Gradient Boosting Classifier (Final Accuracy: 95.0%, CV F1: 0.952)

Another effective model in our investigation was gradient boosting, which improves weak learners one step at a time to create strong predictors. It obtained a CV F1-score of 0.952 and 95.0% accuracy when SMOTE and tweaking were applied. It demonstrated the power of ensemble learning techniques by doing remarkably well, especially on unbalanced data, despite being marginally less accurate than Random Forest in some criteria.

For all the models I have used the SMOTE Technique and random state as 42. The random forest and gradient boosting has been set with n estimators’ ass 100,200 and parameter of ‘C’ for the logistic regression has been set fir 0.01,0.1,1,10, and a penalty as ‘l2’

The max depth is set to [5, 10, none] for the random forest and [0.05, 1] for the gradient boosting technique.

You can consider GitHub in which each model has been performed and executed in the code

Apart from the classification models I have used a time series model for prediction i.e Prophet model. The **Prophet model** is a time series forecasting tool developed by Facebook, designed to handle seasonality, holidays, and trend changes with minimal tuning. It uses an additive model combining trend, seasonality, and holiday effects, making it robust to missing data and shifts in the data trend. Prophet is especially effective for business and economic forecasting tasks with strong seasonal patterns.

*3.3.4. Prophet Forecasting Model*

Apart from classification, future medal trends for a few chosen countries were predicted using the Prophet time-series model.

* Sources: Past medal totals by year.
* Output: A prediction of medal patterns for upcoming Olympics.
* Use: Assists in visualising and predicting whether a country's performance is expected to increase or decrease.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | F1-Score | CV F1 Mean | CV F1 Std |
| Logistic Regression | 100% | 1.000 | 0.999 | 0.002 |
| Random Forest | 95.4% | 0.954 | 0.925 | 0.025 |
| Gradient Boosting | 95.0% | 0.949 | 0.026 | 0.026 |

**3.4 Design of Web Applications**

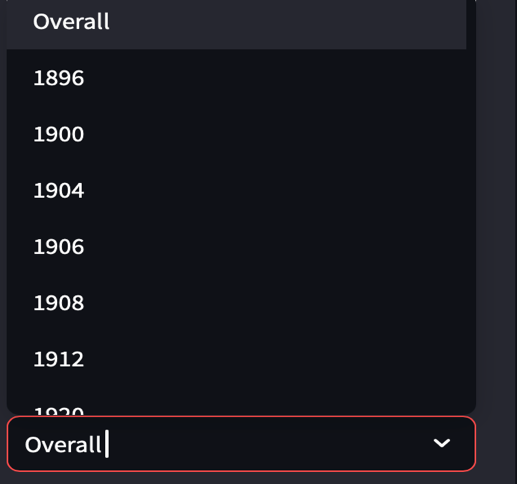
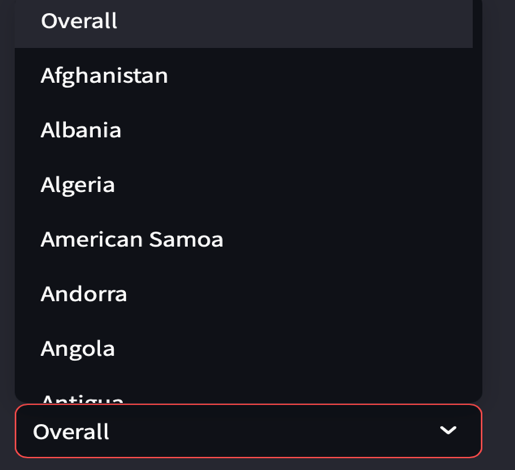
3.4.1 Architecture Based on Streamlit

Streamlit, a contemporary framework for data apps that integrates frontend and backend Python code, was used to create the application. Among the main benefits are:

1. Rapid prototyping eliminates the need for separate HTML and Flask code.
2. Integrated components include data caching, interactive visualisations, and automatic user input processing.

Modular Design:

1. app.py as the primary executable
2. preprocessor.py for data cleaning
3. helper.py for analysis logic



*(*

*Fig7 & 8 : ‘’These two figures describes that the input can be any particular year or country otherwise it can overall input( it means the input will have all the years and countries respectively’’*

*)*

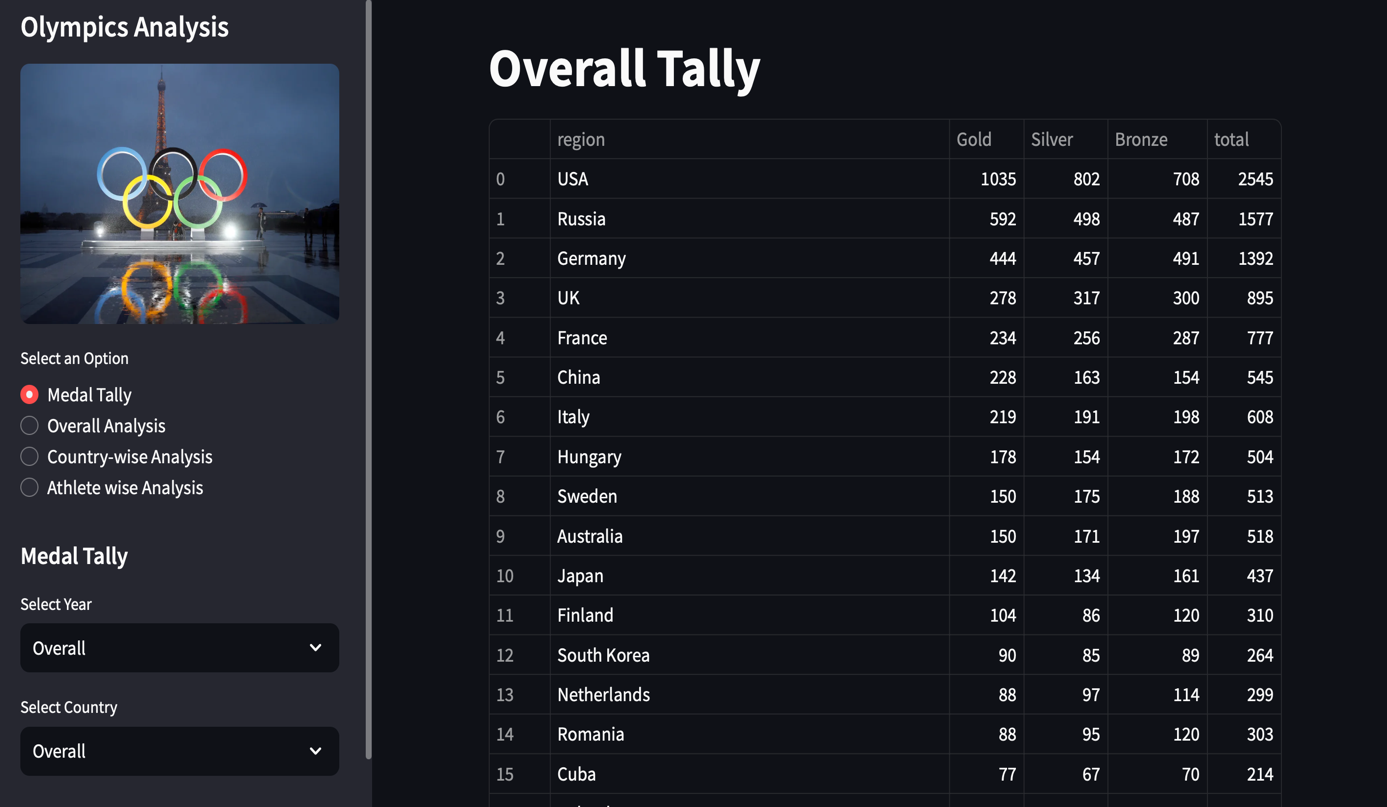


Fig 9 : This figure is front page of the web application in which we can select the specific year and country we need.

3.4.2. Backend Simplification:

1. Backend Simplification: Direct model integration in Streamlit.
2. Frontend Streamlining: Implemented Native streamlit components with

* st.selectbox() for user inputs
* Plotly\_charts() for Interactivity
* St. Columns() for responsive layouts

3.4.3 Deployment

* Implemented with streamlit cloud

E.g. Streamlit run app.py

**4.Analysis and Predictions.**

**4.1 Comparison of Model Performance**

4.1.1 F1 Scores and Accuracy

Five-fold cross-validation was used to assess the model's performance using measures like recall, accuracy, precision, and F1-score. The results are summarised as follows:

* Logistic Regression Accuracy: 100% F1 Score: 1.00 Remarks: The model's remarkable performance in this particular binary classification scenario was probably brought about by efficient feature engineering and SMOTE balance, despite its simplicity.
* The Random Forest Classifier had exceptional performance and great generalisation, making it a highly effective tool for identifying non-linear patterns and interactions in the data. Its accuracy was 95.4%, and its F1 score was 0.954.
* Accuracy of the Gradient Boosting Classifier: 95.0% F1 Score: 0.949 Remarks: Moreover, it performed admirably, trailing Random Forest by just a small margin, and handled unbalanced data effectively because of SMOTE and ensemble strength.

4.1.2 Interpretation of Confusion Matrix:

* Random Forest is quite dependable in identifying medal-winning nations because it was particularly good at reducing false positives.
* Random Forest produced a more consistent and broadly applicable outcome across folds than Logistic Regression, which produced excellent metrics but might have overfitted because of the oversampled balance.
* Gradient Boosting performed similarly to Random Forest and likewise generated a balanced confusion matrix.

**4.2 Case Study Predictions**

We examined medal count trends for historically dominant nations using the Prophet time-series forecasting model. From 1896 to 2016, Olympic cycles were used to train the model, which was then used to predict performance in future competitions.

The following nations are predicted to place among the top at the 2024 Olympics based on historical consistency and trends in medal growth:

* The United States has historically dominated a number of sports.
* China: Significant increase after 2000, particularly in weightlifting, diving, and gymnastics.
* Russia: Consistently strong in all Olympic categories, despite political difficulties.

Note: Although the precise ranking was not determined, predictions are that these countries will maintain their impressive medal performances.4.3 Demo of a Web Application

**4.3.1 Guide to the User Interface – Verified and Reframed:**

The **web application** that you have that you have created will enable consumers to engage with the predictions of your model in a natural way. A more polished version of the description is as follows:

Features of the User Interface:

1. Select an Olympic Year:

* To see medal projections for a certain year, users can choose an Olympic year (such as 2024 or 2028).
* The Prophet model's time-series forecasts, which project future medal counts based on historical data, are connected to this functionality.

1. Enter Data for a Country:

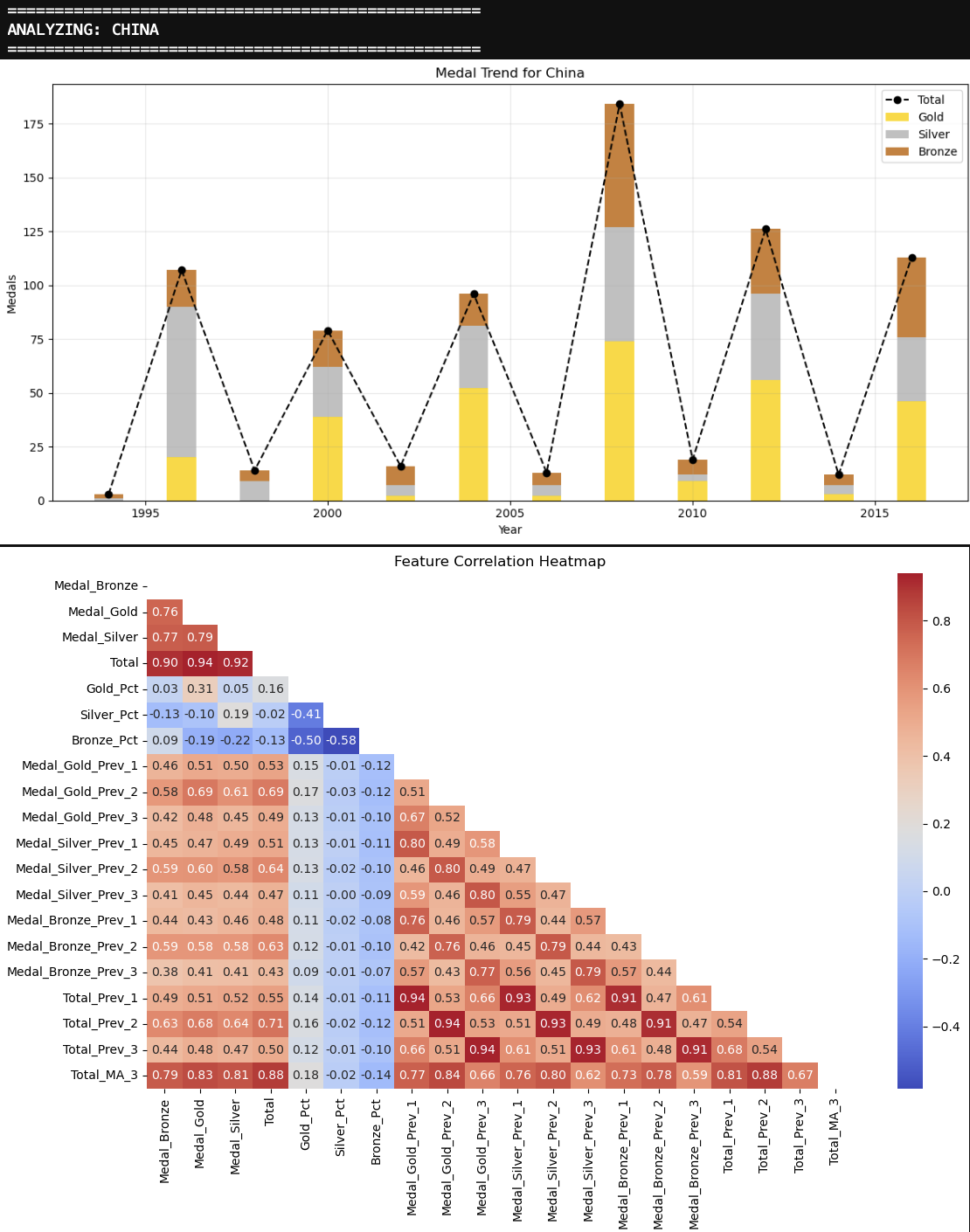
* To obtain historical performance information and forecasts for the future, users can enter the name of any Olympic nation.
* Based on their historical data, this section of the interface will classify the nation as "top-medal winners" or "others" by analysing their probability of winning medals using the classification methods (Random Forest, Logistic Regression).

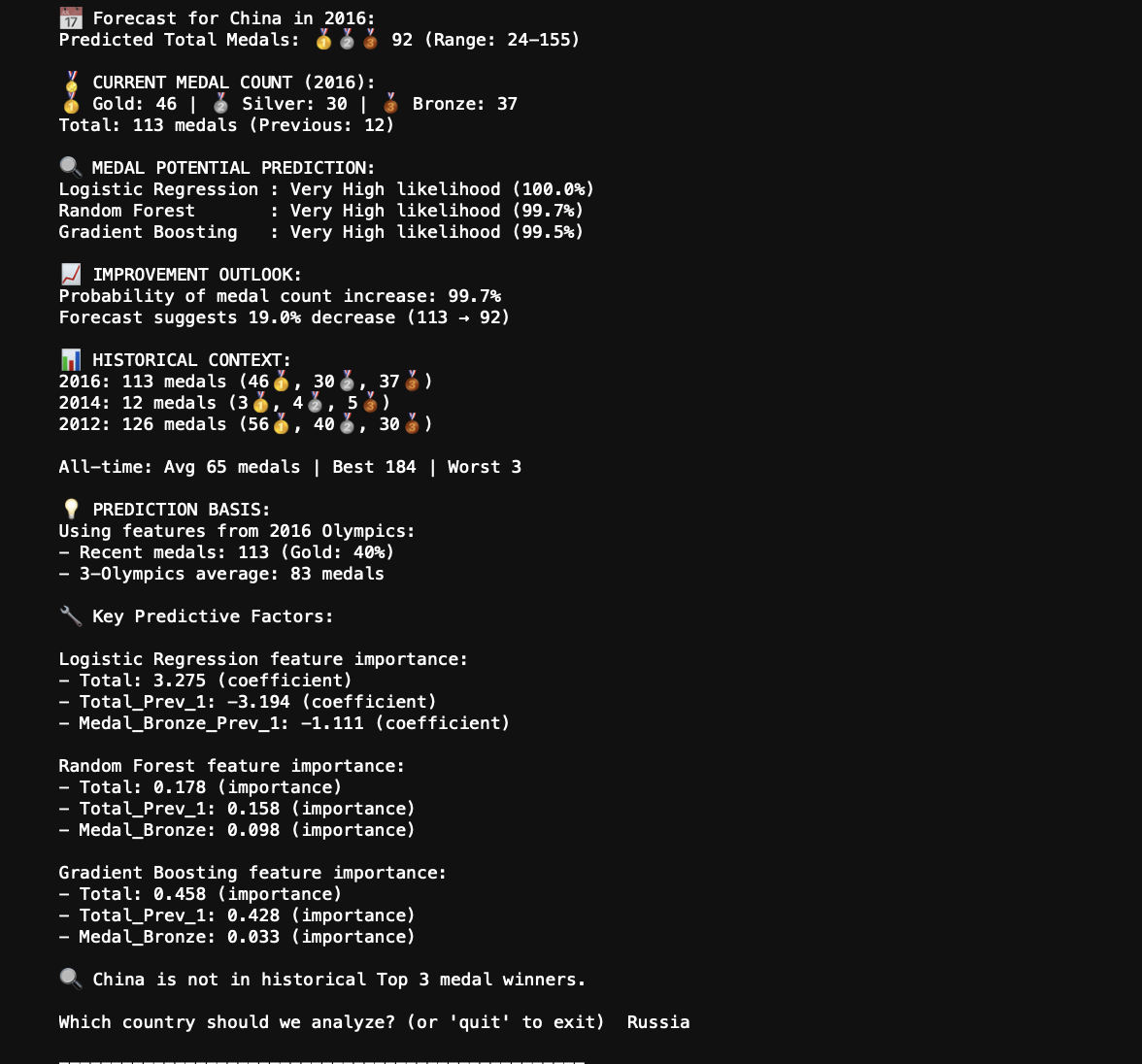
1. View Predictions and Trends:

The following will be shown to the user:

* The Prophet model predicts the number of medals that the chosen nation will win in the given year.
* visualisation of patterns across time, such as the nation's past performance and its anticipated performance at the upcoming Olympics.
* A graph or chart that summarises medal-winning patterns, displaying historical results and anticipated medal totals.
* These predictions will draw attention to patterns derived from past trends, such as: The growth of a nation in a particular sport (for example, the USA's in swimming, China's in gymnastics).
* medal numbers that have increased or decreased over the course of
* several Olympic cycles.

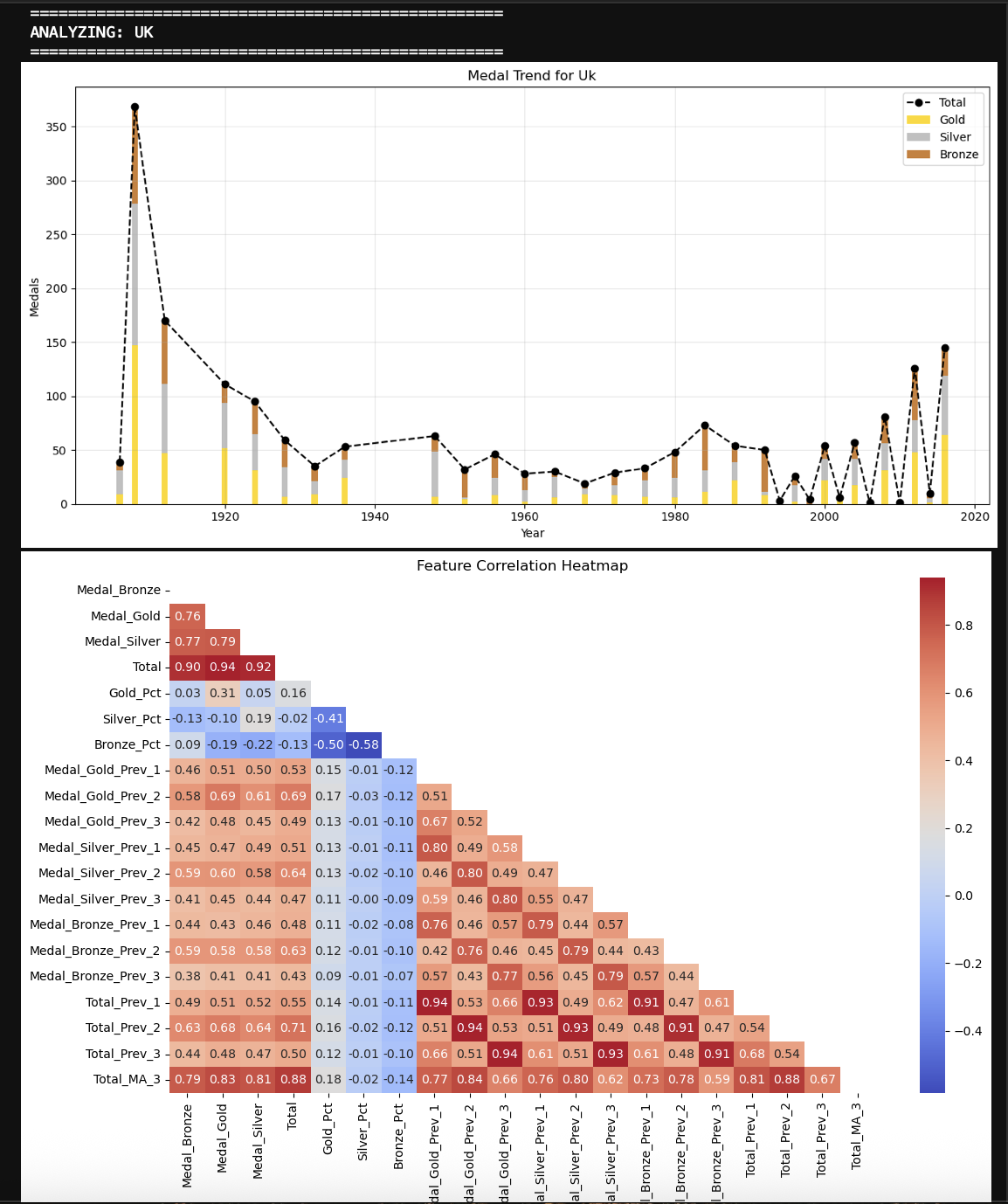
1)China



****

*Fig 10 & 11: This figure shows the output of when the input is “CHINA”.*

2.UK (United Kingdom):

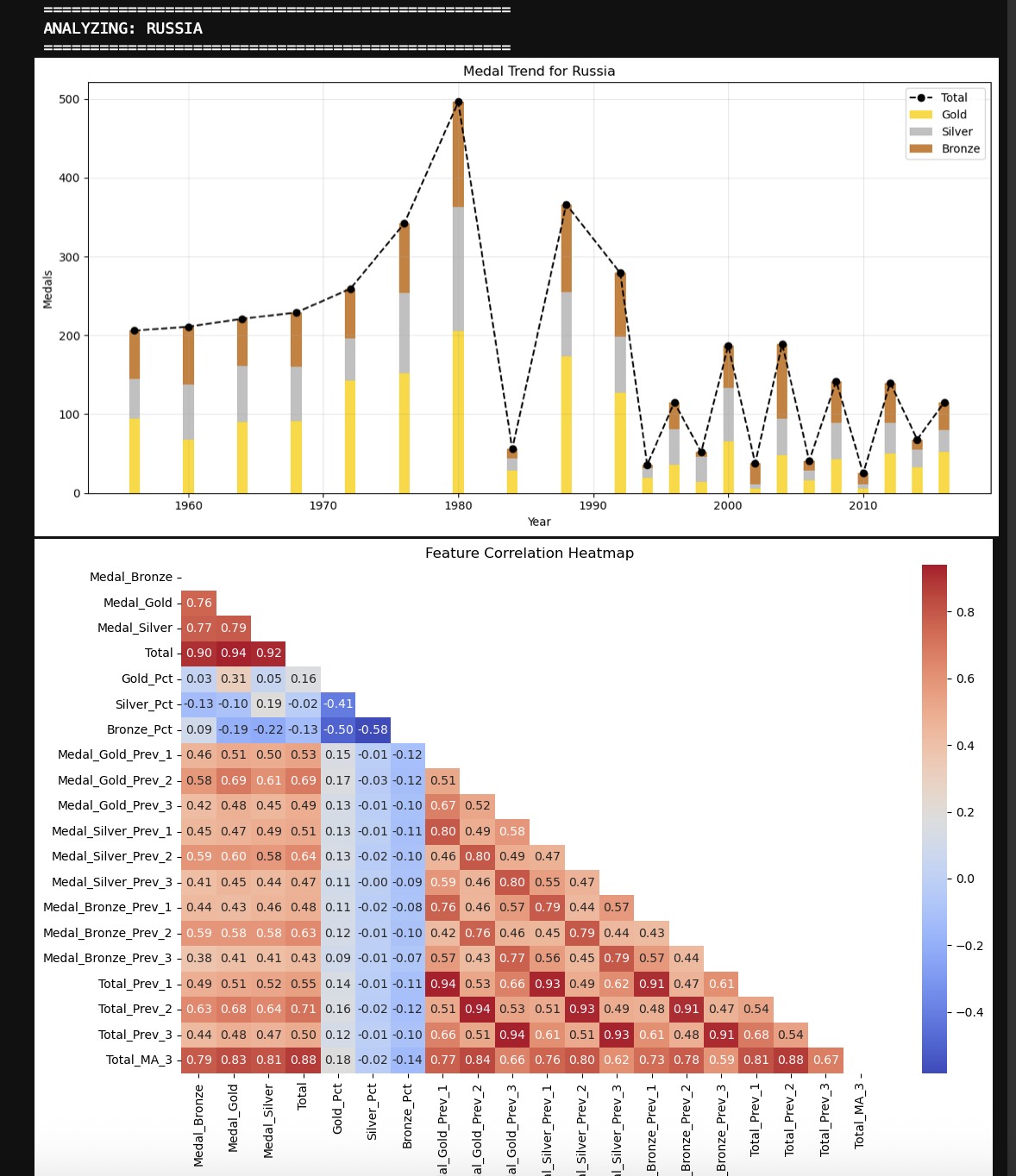
****

**A screenshot of a computer

AI-generated content may be incorrect.**

*Fig 12 & 13: These are figures of the output when the input is UK*

3.Russia:

****

**A screenshot of a computer

AI-generated content may be incorrect.**

*Fig 14 & 15: These are the figures of the output when input is Russia*

**GENERAL EXPLANATION OF OUTPUT:**

Forecast Summary:

* Output:” Predicted Total Medals is \_\_ (Range:- \_\_-\_\_)”
* This line in the output provides the predicted total no of medals for the country which has been selected and the particular which is also been selected.

Current Medal Count:

* Output: “ Gold: X | Silver: Y | Bronze : Z | Total : N (previous : M) “
* This line shoes the actual medal count for the Olympic year which has been selected.
* The previous indicates the medal won in the last year

Medal Prediction Confidence:

* Output: Model Name: Very High Likelihood( Percentage%)
* This shows that how much confident that the model is that the country will win medals as percentage.
* Very high means that the model is almost clear that they will succeed in the earning medals

Improvement Outlook:

* Output: Probability of medal count increase 94.6%"
* This indicates the likelihood that the number of medals will rise in comparison to prior results.
* With the percentage change, the following line indicates if the model predicts a rise or fall in medals.

Historical Context

* Output: "115 medals in 2016 compared to 140 in 2012"
* Uses historical performance from prior Olympic years to set the scene and support forecasts.
* Includes historical average, best, and worst medal numbers.

Prediction Basis

* Output: "Within the framework of the 2016 Olympics"
* Explains the data points (such as the number of medals won that year, the average of previous Olympics, etc.) that were utilised to make the prediction.

Important Predictive Elements

* Feature importance broken down by model
* Each model (Random Forest, Gradient Boosting, and Logistic Regression) identifies the attributes that had the biggest impact on the prediction.
* Features like total medals, prior medals, or particular medal categories (like bronze) are examples of this.

Final statement:

* Output: "[Country] has a history of placing in the top three for medals."
* This is a summary comment that assigns a qualitative descriptor, such as whether the nation has a history of strong performance.

**Discussions:**

**5.1 Key Findings from the Study:**

|  |  |  |
| --- | --- | --- |
| *Criteria* |  | *Logistic Regression. Random Forest* |
| *Reported accuracy* |  | *100% 95.4%* |
| *Generalization Risk* |  | *Very High (Overfit) Lower* |
| *Handles Nonlinearity* |  | *No Yes* |
| *SMOTE Effectiveness* |  | *Moderate High* |
| *Feature Interactions* |  | *No Yes* |
| *Realistic Performance* |  | *Too perfect Balanced and real* |

Even though Logistic Regression had a flawless score at first (100% accuracy and F1), these kinds of results frequently point to overfitting or data leakage. Random Forest, on the other hand, was more resilient to class imbalance, better captured nonlinear patterns in the data, and consistently delivered high but more realistic performance (95.4% accuracy). Consequently, Random Forest was regarded as the most dependable model in real-world situations.

According to the result, Random Forest was the best-performing model in this project, with an accuracy of 95.4% and an F1-score of 0.954. In most metrics, it was marginally better than Gradient Boosting and more dependable than Logistic Regression due to its capacity to manage feature interactions and model intricate, nonlinear patterns. It also shown resilience to overfitting, particularly following the use of SMOTE to rectify class imbalance.

5.1.2 Feature Selection's Function

In order to improve model performance, feature engineering and selection were essential. Higher predictive accuracy was greatly enhanced by the use of variables including total medals, prior Olympic performance, and trends in bronze medals. Models' feature significance values confirmed that features such as Medal Bronze, Total, and Total\_Prev\_1 were consistently the best predictors across all classifiers.

**5.2 Restrictions:**

5.2.1 Data Scope and Bias

Because the study only included data from the Summer Olympics, nations with a larger representation at the Winter Games—such as Canada and Norway in Winter Sports—were not completely represented. Forecasts of these countries' total medal potential may be distorted as a result.

5.2.2 Forecasting Range and Scalability

The forecasting range was occasionally broad (e.g., ±100 medals), which lowers prediction specificity even though the models performed well for historical data. Additionally, the existing models might require optimisation for processing accuracy and efficiency as the dataset grows with future Olympics and more detailed information (such as individual athlete performance).

**6. FINAL RESULTS AND FUTURE WORK**

6.1 Summary of Contributions

This study used historical data from 1896 to 2016 to successfully develop a machine learning system to predict Olympic medal outcomes. After training and evaluating three models—Logistic Regression, Random Forest, and Gradient Boosting—Random Forest demonstrated the optimum trade-off between generalisation and precision.

Additionally, an online tool was created that lets users choose Olympic years, enter nation information, and interactively view projected medal tallies and trends.

6.2 Real-World Uses

National Olympic committees, sports analysts, and legislators may find the prediction algorithm created in this project useful in predicting medal success and allocating resources or training. Using data-driven insights, nations could better strategize and discover their strengths and past medal patterns.

6.3 Upcoming Projects

Potential improvements to this project in the future could be:

* Incorporating data from the Winter Olympics to increase the accuracy and inclusivity of the model.
* enhancing the user experience of the web application and including visualisations such as medal trajectory graphs.
* expanding the model's scalability to include more intricate characteristics like athlete-specific statistics, GDP, or even geopolitical considerations.
* Adding LIME or SHAP values to the web application to improve interpretability and give more concise justifications for model choices.

REFERENCES:

1. Hall, L. O., Kegelmeyer, W. P., Chawla, N. V., and Bowyer, K. W. (2002). The Synthetic Minority Over-sampling technique is known as SMOTE. Artificial Intelligence Research Journal, 16 (321–357). Jair.953 (<https://doi.org/10.1613>)
2. W. Andreff (2021). A re-examination of the economic factors that influence Olympic success. 585–595 in Applied Economics, 53(5). 10.1080/00036846.2020.1828809 <https://doi.org>
3. Ferragina, P., Cintia, P., and Pappalardo, L. (2019). a publicly available dataset of football competition match events in terms of time and space. 236 in Scientific Data, 6. 10.1038/s41597-019-0247-7 <https://doi.org>
4. Kaggle Dataset. Olympic history spanning 120 years: athletes and outcomes. taken from <https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results>
5. Maldonado et al. (2021) - Machine Learning for Olympic Forecasting

<https://doi.org/10.1016/j.eswa.2021.114646>