Abstract

This research paper delves into the exhaustive process of data scraping, cleaning, exploratory data analysis (EDA), and modeling to determine the "best" Men's and Women's Teams and Individual USA Olympic Artistic Gymnasts. The concept of "best" is operationally defined as maximizing the total medal count in artistic gymnastics for the United States at the 2024 Olympic Games. The study employs a comprehensive approach, integrating machine learning models, combinatorial optimization algorithms, and advanced data analysis techniques. The Initial phases involve predicting individual apparatus outcomes using machine learning models, evaluated rigorously through metrics like accuracy, precision, recall, F1-Score, and F2-Score. Data, spanning 2013 to 2023 from a myriad of sources, undergoes critical preprocessing. Team selection involves diverse datasets for Men's and Women's teams, utilizing Tabu Search, Hyper-Heuristic, Variable Neighborhood Search, and Brute Force algorithms. Grounded in data analysis and algorithmic decision-making, the research strategically positions Team USA to maximize their medal count in the Paris 2024 Olympics, showcasing the integration of machine learning and optimization algorithms in gymnastics team selection. Binary classification techniques predict whether a gymnast will win a medal in individual events, utilizing historical data from 2013 to 2021 for Women's events and 2018 to 2020 for Men's events.

1. Introduction

This research addresses the imperative task of selecting the best Men's and Women's USA Olympic Artistic Gymnastics teams for the Paris 2024 Olympics. The primary objective is to strategically position Team USA to maximize their medal count in the 2024 Olympics. At the heart of this research is the challenge of identifying five exceptional athletes for each team, considering not only their ability to medal in individual events but also their capacity to medal in their team events. Leveraging advanced analytics models, the study aims to forecast and compare medal counts, providing a data-driven foundation for the strategic decision-making needed to excel in the 2024 Olympic Games.

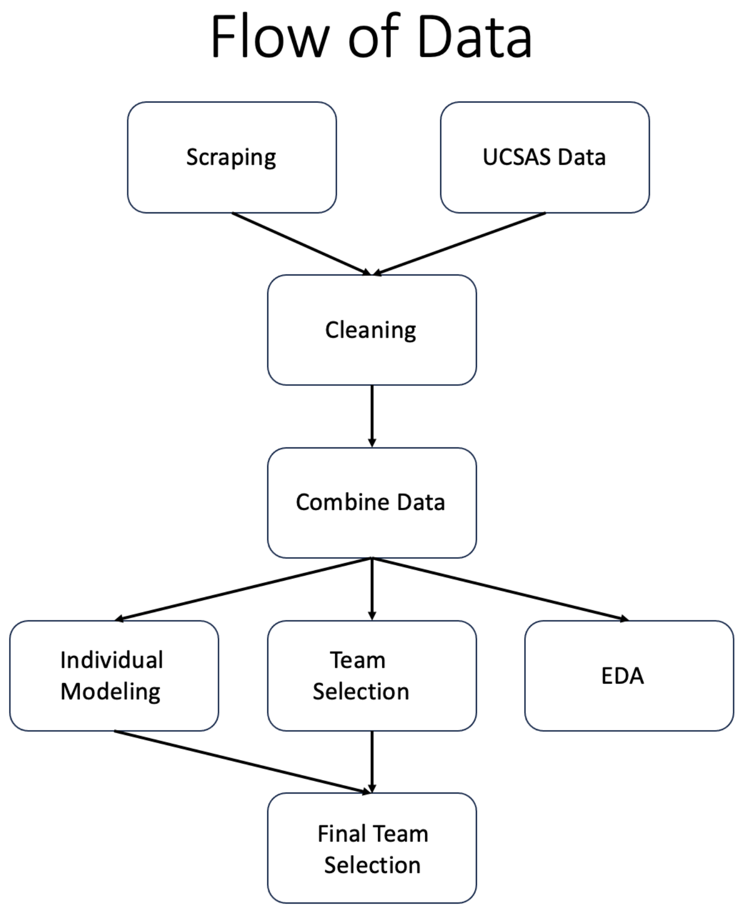
The determination of the "best" gymnast’s hinges on a nuanced evaluation of medal counts, necessitating the establishment of baseline models for individual apparatus outcomes. The research employs the ZeroR algorithm as a foundational benchmark and subsequently explores a variety of machine learning models, incorporating classifiers such as Random Forest Classifier, AdaBoost Classifier, Support Vector Classifier, K Neighbors Classifier, Decision Tree Classifier, Gaussian Naive Bayes, and Neural Networks. This approach allows for a comprehensive assessment of individual gymnasts' performance and the factors influencing their success.

Assembling the optimal Men's and Women's teams involves delving into the realm of optimization algorithms, including Hyper Heuristics, Stochastic Control, Tabu Search, Variable Neighborhood Search, and Brute Force. These algorithms play a pivotal role in navigating the complexities of team composition.

The foundation of this research rests on a diverse dataset sourced from reputable platforms such as Thegymter.net, Wikipedia.org, and the UConn

Sports Analytics Symposium 2024. Covering the period from 2015 to 2023, the dataset includes individual scores for all apparatus, all-around scores, rankings, execution scores, difficulty scores, penalties, and final overall scores. The data undergoes meticulous preprocessing, ensuring its integrity and uniformity for accurate analysis. In this pursuit of excellence, the integration of machine learning models, Optimization algorithms, and comprehensive data analysis emerges as a distinctive feature of this research. The goal is to provide Team USA with a strategic advantage, marrying the precision of analytics with the artistry of gymnastics in the quest for Olympic glory in Paris 2024.

2. Data Collection and Preprocessing



2.1. Data Sources

This research procures its data from a diverse range of reputable sources, emphasizing historical performance records and competitive outcomes. Predominantly, the key data origins encompass Thegymter.net, UCSAS 2024 USOPC Data Challenge, and Wikipedia.org. Spanning from 2013 to 2023, the dataset is predominantly compiled from various international competitions, including world cups, world championships, and the Olympics. This comprehensive dataset incorporates individual scores for all apparatus and overall individual performances. Furthermore, it includes vital details such as gymnasts' rankings across diverse apparatus and rounds, athletes' names and country, scores on each apparatus, execution scores, difficulty scores, penalties, and their conclusive overall scores. It is important to note that due to time constraints, the primary focus of data collection centers on the years 2023 and 2022, primarily derived from international competitions, resulting in a dataset that, while extensive, may not be as comprehensive as optimal.

2.2. Data preprocessing steps

The data preprocessing stage plays a pivotal role in this research, involving several essential procedures to ensure the dataset's integrity and usability. The workflow initiation occurs in the 'Scraping' folder, housing code for web scraping and resulting CSV files containing raw data. Subsequently, the data undergoes a sequence of transformations as it progresses through distinct stages:

2.2.1. Cleaning

Within the 'Cleaning' folder, meticulous procedures are applied to the scraped CSV files. This encompasses tasks such as filling in or removing missing values to alleviate data inconsistencies. Additionally, operations are conducted to standardize column names and convert nation names to their corresponding National Olympic Committee (NOC) codes where necessary.

2.2.2. Data integration

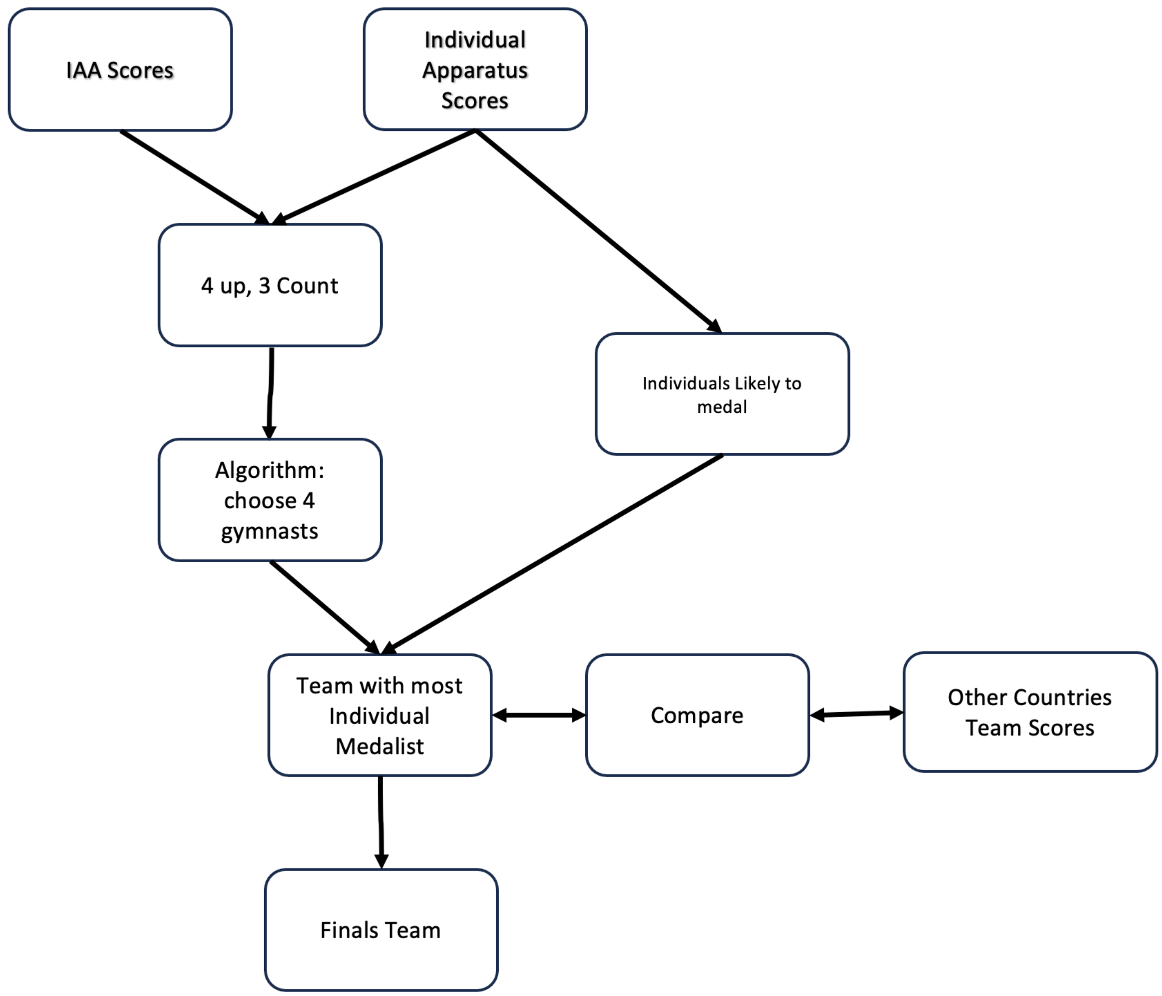
The preprocessed data advances to the 'Combine Data' folder, where information from diverse competitions is consolidated. The data is consolidated based on key parameters such as the year, gender, and specific apparatus, providing a holistic view of the gymnastic performance dataset.

2.2.3. Imbalance Data

Acknowledging the highly imbalanced nature of the datasets used in this analysis, where the minority class is notably underrepresented, poses challenges. Models trained on imbalanced data may exhibit bias towards the majority class, emphasizing the need for approaches resilient to such imbalances. Consequently, robust models tailored for imbalanced datasets will be employed to ensure fair and accurate representation across classes.

These data preprocessing steps are crucial in transforming raw data into a clean, structured, and uniform dataset, setting the stage for accurate and insightful analysis in subsequent stages of the research.

3. Methodology



3.1. Individual

3.1.1. Models

In this research, a variety of machine learning models were employed to predict individual apparatus outcomes. The models assessed in this study encompassed a comprehensive list, including:

- Random Forest Classifier

  “A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting." (“sklearn.ensemble.RandomForestClassifier — scikit-learn 1.3.2 documentation”)

- AdaBoost Classifier

  “An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.” (“All Machine Learning Algorithms Explained - thecleverprogrammer”)

- Support Vector Classifier (SVC)

  “Support vector machines (SVMs) are particular linear classifiers which are based on the margin maximization principle. They perform structural risk minimization, which improves the complexity of the classifier with the aim of achieving excellent generalization performance. The SVM accomplishes the classification task by constructing, in a higher dimensional space, the hyperplane that optimally separates the data into two categories.” (Djuris 2023)

- K-Neighbors Classifier

  “The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.” (“What is the k-nearest neighbors algorithm?”, n.d.)

- Decision Tree Classifier

  “Decision tree classifiers are used successfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision-making knowledge from the supplied data. Decision tree can be generated from training sets.” (Gunasekaran 2001, #)

- Gaussian Naive Bayes (GaussianNB)

  “The thought behind naive Bayes classification is to try to classify the data by maximizing $P(O|Ci)P(Ci)$ using Bayes theorem of posterior probability (where O is the Object or tuple in a dataset and “i” is an index of the class)” (Shobha 2023)

- Neural Networks

  “Neural networks (NN) are used as the machine learning model, which simulates the working mechanism of a biological neuron.” (“Neural Network” 2023)

These models were meticulously evaluated through the utilization of critical performance metrics discussed below.

3.1.2. Hyperparameter tuning

Grid Search: In the pursuit of optimizing the hyperparameters governing the predictive models, a comprehensive and systematic approach was employed known as Grid Search. This technique is instrumental in exploring a predefined hyperparameter space to identify the most optimal combination for enhancing model performance.

3.1.3. Cross-validation strategies

Holdout validation and k-folds cross validation were both utilized in the individual apparatus modeling.

- Holdout Validation: In holdout validation, the dataset is divided into two parts: the training set and the validation set (sometimes referred to as the test set). "The model is trained on the training set and then evaluated on the validation set." (“Model Evaluation and Validation: Describing the various ... - Medium”)
- K-folds cross validation: “divides all the samples in groups of samples, called folds of equal sizes (if possible). The prediction function is learned using folds, and the fold left out is used for the test.” (“3.1. Cross-validation: evaluating estimator performance — scikit-learn 1.3.2 documentation”, n.d.)

3.1.4. Base model

“ZeroR is the simplest classification method which relies on the target and ignores all predictors. ZeroR classifier simply predicts the majority category (class). Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods”.(“ZeroR - Data Mining Map”)

3.1.5. Confusion matrix

A confusion matrix was utilized to gain insights into the models' performance, which measures accuracy, precision, F1-Score, and F2-Score.

Accuracy measures the proportion of correctly predicted outcomes among all predictions. It provides an overall assessment of model correctness.
Precision measures the proportion of true positive predictions among all positive predictions, while recall measures the proportion of true positives among all actual positives. (“1 Self-supervised Learning for Pre-Training 3D Point Clouds: A

Survey") These metrics are particularly useful in assessing the model's ability to identify winners accurately.

The F1-Score combines precision and recall providing a balanced measure of a model's performance. ("Robust Weapon Detection in Dark Environments using ... - ScienceDirect") It is especially valuable when dealing with imbalanced datasets.

"The F2 score is the weighted harmonic mean of the precision and recall (given a threshold value). Unlike the F1 score, which gives equal weight to precision and recall, the F2 score gives more weight to recall than to precision. More weight should be given to recall for cases where False Negatives are considered worse than False Positives." ("Scorers — Using Driverless AI 1.10.6.2 documentation") When possible, it's better to have False Positives (FP): Instances where gymnasts did not medal but are predicted to medal than to have False Negatives (FN): Instances where gymnasts medaled but are predicted not to medal. Which would result in gymnasts being labeled as medaling, despite not medaling, rather than having gymnasts who medal being labeled as not medaling.

These comprehensive evaluation metrics provide a robust analysis of the individual machine learning models' performance in predicting gymnastics outcomes.

3.2. Team

This research employed two datasets for team selection algorithms, focusing on individual all-around scores and individual apparatus scores for USA gymnasts. To ensure fairness, in cases where an individual had multiple entries in either dataset, only their highest scores were considered.

3.2.1. Qualification Round:

The qualification round uses 4 up 3 counts, where four athletes will compete on each apparatus and the sum of the three best scores on each apparatus is used to see which team will advance. Several algorithms were used with the same goal in mind; to get the three highest scores from four gymnasts, where each apparatus must be used three times and gymnasts cannot compete on any apparatus more than once.

3.2.1.1. Men's team selection

For the men's team selection, Tabu Search, Hyper-Heuristic, and Variable Neighborhood Search (VNS) algorithms were utilized. Each algorithm employed two distinct datasets, incorporating data from 2022 and

2023, one with IAA scores and the other with individual apparatus scores, to enhance the selection process.

3.2.1.2. Women's team selection

The Team selection used two different data sets. The first was using scores from 2022 to 2023 from USA Women's Individual All-Around, which is denoted in the datasets as IAA. The second dataset used was from 2023 using USA individual women's scores from all individual apparatus.

The Women's Team selection integrated Tabu Search, Hyper-Heuristic, Variable Neighborhood Search (VNS), and Brute Force algorithms. Notably, the Brute Force algorithm, while a comprehensive computational approach, did not yield optimal results for men's team selection due to heightened complexity resulting from an increased number of features. This highlights the nuanced considerations required in algorithm selection based on the specific characteristics of the data.

3.2.2. Finals:

The finals follow a 3 up 3 count, with three gymnasts competing on each apparatus. All their scores contribute to the final team score. To determine the finalists, the gymnasts with the highest Individual All-Around (IAA) scores are selected. If there is a gymnast who has a higher IAA score than the top 3 from the 4 selected from qualifications, then the athlete will be added to the finals.

4. Evaluation and Results

4.1. individual Apparatus Evaluation and Results

The study utilizes historical data for Men's and Women's gymnastics, with Women's models trained on data from 2013-2021 of 2,153 different data points and tested on the 2016 and 2020 Olympics, and because of time restraints, the Men's models trained on data from 2018-2020 of 3193 data points and tested on the 2020 Olympics.

Apparatus	Men Data Size	Women Data Size
Vault	126	145
Floor	538	438
Balance Beam	-	587
Uneven Bars	-	446
IAA	512	537
High Bar	502	-

Pommel Horse	535	-
Rings	495	-
Parallel Bar	485	-
Total	3193	2153

4.2. Model Performance Metrics

For each apparatus, the data was run through the various models, and their accuracy, precision, F1-scores, and F2-scores were compared. Once the best model was selected, judged by its accuracy and F2-score, it was optimized using grid search for hyper-parameter tuning. From there the data was scrutinized with feature importance and scaling. The models were then run again to see if there were any improvements.

The findings revealed that the Random Forest classifier consistently displayed high accuracy across multiple gymnastics events, suggesting its effectiveness in correctly selecting the top gymnasts for Team USA. The AdaBoost classifier also demonstrated strong performance in terms of accuracy, consistently achieving high scores in multiple events. This indicates that the AdaBoost classifier is reliable in accurately selecting top gymnasts.

In addition to accuracy, I also evaluated the classifiers based on the F2-score metric. The F2-score considers both precision and recall, with a higher weight given to recall. It measures the overall effectiveness of the classifier in selecting top gymnasts while minimizing false negatives.

The results showed that the Random Forest classifier consistently achieved high F2-scores across multiple gymnastics events, suggesting its effectiveness in accurately selecting the top performers. Similarly, the AdaBoost classifier exhibited strong performance in terms of F2-scores, indicating its reliability in identifying top gymnasts.

In summary, the Random Forest and AdaBoost classifiers consistently demonstrated high accuracy and F2-scores in the selection process for Team USA gymnasts. These findings highlight the effectiveness of these classifiers in accurately identifying the top gymnasts while minimizing false negatives. The results provide valuable insights for improving the selection process and ensuring the inclusion of the most deserving gymnasts in Team USA.

4.3. Individuals likely to medal

The following are the resulting gymnasts likely to medal for team USA from the above models.

4.3.1. Women

Event	Gymnast

Balance Beam	Simone Biles
Balance Beam	Joscelyn Roberson
Vault	Simone Biles
Uneven Bars	Shilese Jones
Uneven Bars	Zoe Miller
IAA	Jordan Chiles
Floor	Simone Biles
Floor	Ashlee Sullivan

4.3.2. Men

Apparatus	Gymnast
Parallel Bars	Yul Moldauer
Parallel Bars	Curran Phillips

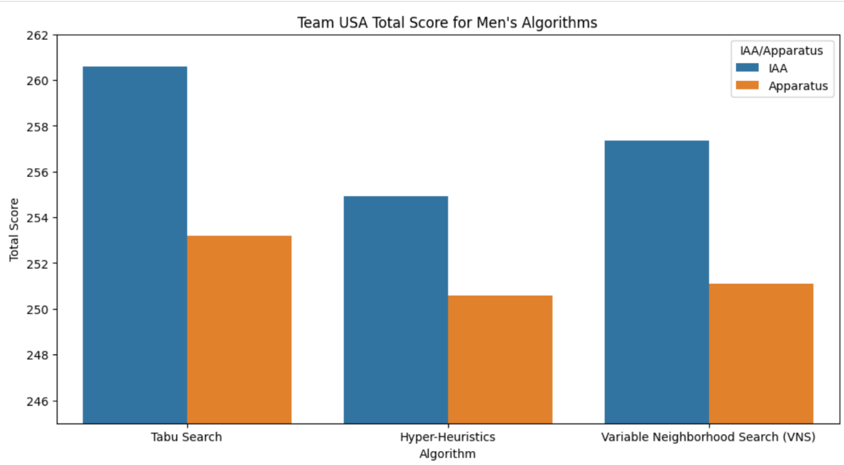
4.4. Team Evaluation and Results

Originally the earlier mentioned algorithms were implemented with data sets containing all American women or men from the data scraped from international events. This resulted in optimized teams for the men and women.

Running the individual events modeling, resulted in gymnasts not included in the previously selected teams. Requiring different approaches for the men's and women's teams.

4.4.1. Men's Team Evaluation and Results

The highest-scoring algorithms for the men's team selection were Tabu Search and Variable Neighborhood Search (VNS) algorithms, with the IAA data sets scoring higher on all models. The highest-scoring algorithm was the Tabu Search using IAA Scores.



The first phase of the research involves assessing the probability of Team USA achieving a medal in the team competition and discerning the gymnasts who would contribute to the team's success. Subsequent stages of the analysis entail identifying individual gymnasts with the highest likelihood of medaling in specific apparatus events. The examination uncovered several gymnasts expected to excel in individual apparatus events, even if their alignment with the initially predicted Team USA lineup was not perfect. As a result, the team selection algorithms were extended to other countries that demonstrated strong performance in the team qualification rounds for the 2024 Olympics. This extension provided a valuable framework for estimating the scores of top-performing nations, thereby aiding in the selection of gymnasts for Team USA. The overarching goal was to optimize the individual medal potential of gymnasts while ensuring the overall success of the team.

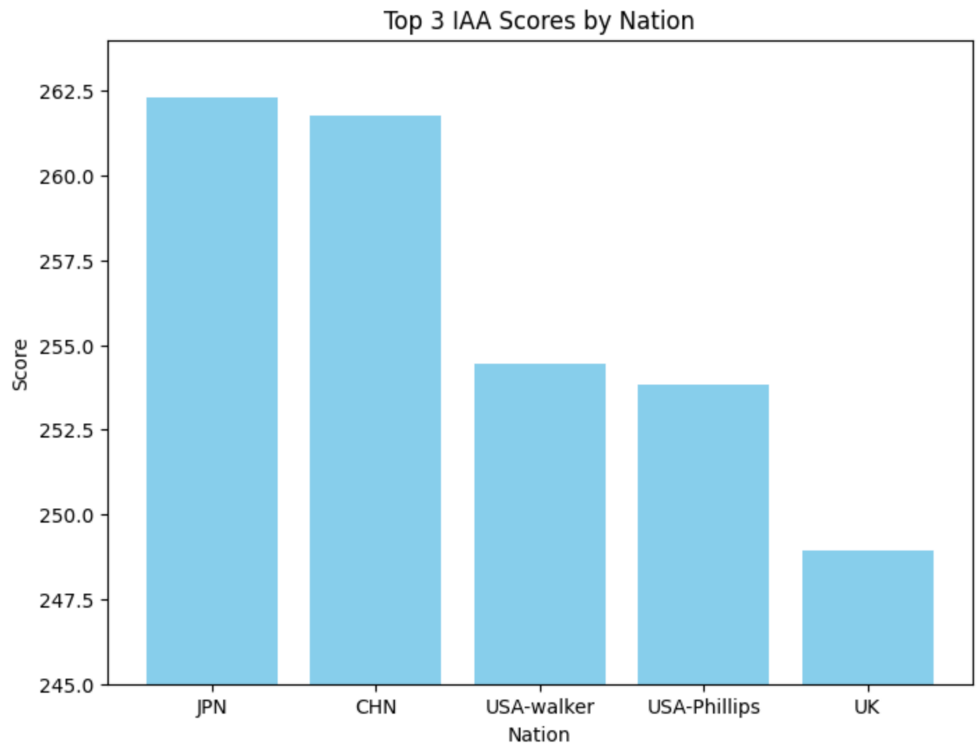
For the men's team selection, I began by choosing 4 athletes to compete in the qualification round. Which included Young, Hong, Richard, and Moldauer competing in the qualifying round. The combined score for the team was calculated at approximately 260.579.

Now, to include the two gymnasts with potential to medal individual events: Yul Moldauer and Curran Phillips. Since Moldauer was already selected to compete on the team, we added Phillips to round out our 5-person team.

At this point I analyzed the difference between the potential scores in the final round of the above team, and a team that didn't include Phillips, but rather Colt Walker. Walker has the top IAA score for any American this year with a score of 85.00. A team with Walker could potentially score 254.450 in the final round. While a team with Phillips, would have Moldauer, Richard, and Hong competing in the final round, with a potential score of 253.85. From here these scores were compared to the other top Olympic-qualified countries:

- Japan
- China
- United Kingdom
- Switzerland
- Germany

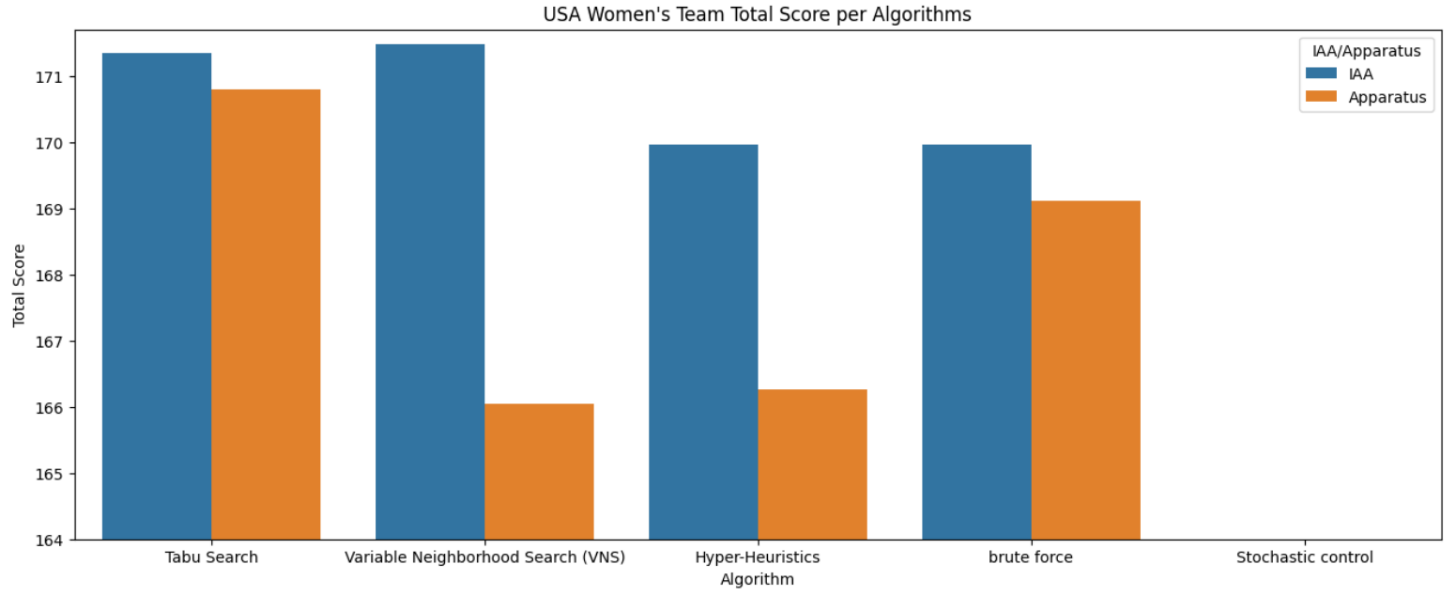
(Germany and Switzerland scores were below 245, and no it the following chart)



Above we can see that the team scores with Phillips still allows team USA to potentially place 3rd.

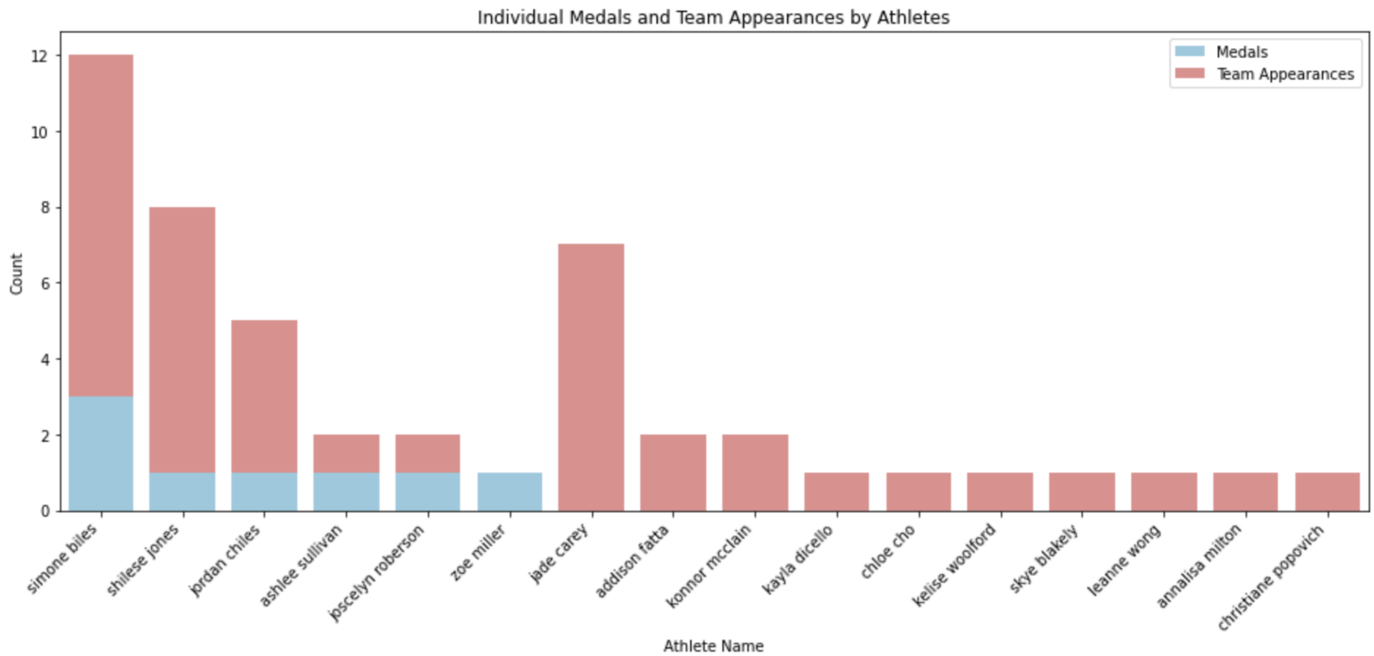
4.4.2. Women's Team Evaluation and Results

The comparative evaluation of diverse algorithms across the IAA and Apparatus data sets provided nuanced insights into their respective performances in finding the highest scoring team USA. Tabu Search demonstrated consistent competitiveness across both IAA and Apparatus scenarios, exhibiting marginal score variations. Variable Neighborhood Search (VNS) maintained stability in scores on the IAA data set but experienced a decline in performance when applied to the Apparatus data. Hyper-Heuristics showcased adaptability, achieving commendable scores across both data sets. Despite its computational intensity, the brute force algorithm delivered competitive total scores on both IAA and Apparatus. The highest score Algorithms were Tabu Search and Variable Neighborhood Search (VNS) algorithms, with the data sets using IAA scores performing better. The highest-scoring Algorithm was the VNS algorithm.



The initial step involves determining the likelihood of Team USA medaling in the team competition and identifying the gymnasts for the team. Subsequent stages include identifying individual gymnasts most likely to medal on specific apparatus. Analysis revealed several gymnasts likely to medal on individual apparatus, some of whom may not align with the predicted Team USA lineup. Consequently, the team selection algorithms were applied to other countries that performed well in team qualification for the 2024 Olympics, providing a guideline for estimating the scores of top-performing nations. This knowledge aided in selecting gymnasts for Team USA and optimizing individual medal potential while securing team success.

The original optimized women's team was analyzed to medal. The combination of the likelihood of a team medal and the possibility of 6 different women potentially medaling resulted in further data exploration and modeling. Below is the number of times each woman was chosen for team USA and each woman predicted to medal in an individual event.

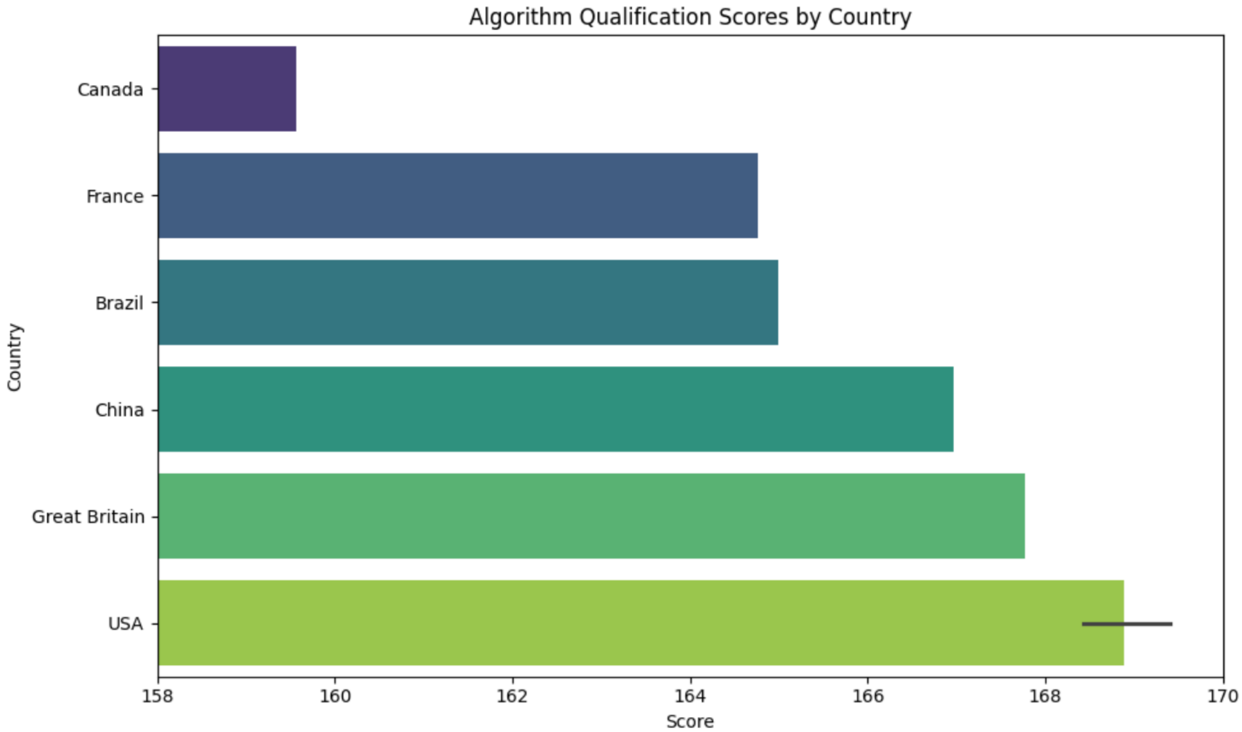


First, I went back and organized data of all female gymnasts from the top Olympic-qualified countries.

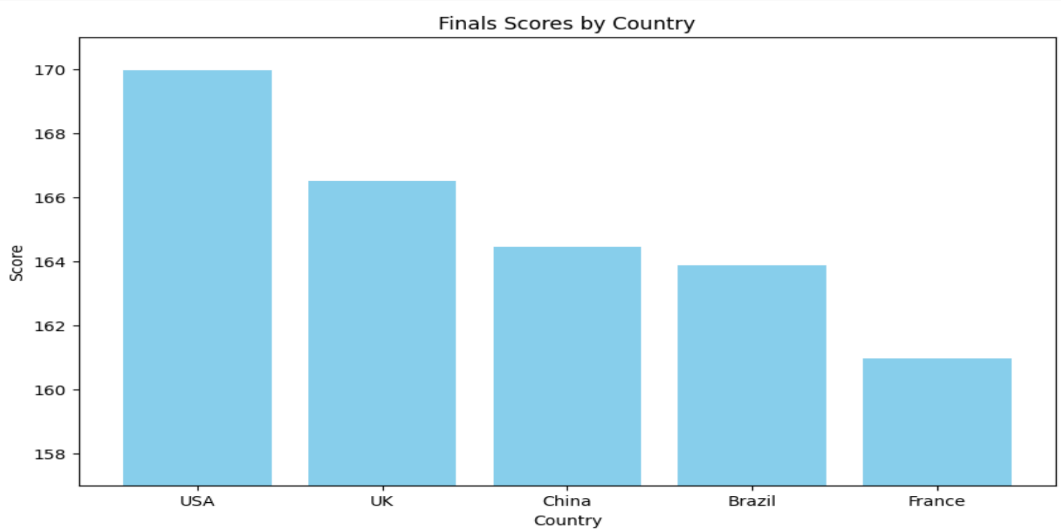
- Great Britain
- China
- Brazil
- France
- Canada

The gymnasts from these countries were run through the algorithms mentioned earlier. This resulted in approximate scores to expect from these countries for the qualifying rounds. The top 3 IAA scores from each of the top countries were also compared to team USA for the Finals.

Since maximizing medals is the goal, the women's team should include Simone Biles, who is predicted to win 3 individual medals, followed by 4 of the 5 remaining women predicted to potentially medal. I had the algorithms choose four gymnasts from the six potential medalists, for qualifying. These team scores were compared to the scores from the other top countries. Two athletes were chosen each time from the algorithms, Simone Biles and Shilese Jones, with Jordan Chiles being chosen seven times. In the graph below we can see team USA should easily pass the qualifying round regardless of which two of the remaining three gymnasts are chosen. Therefore, Team USA should consist of Simone Biles, Shilese Jones, Jordan Chiles, these three should also compete in the final round, and any two of Ashlee Sullivan, Joscelyn Roberson, or Zoe Miller.



The Women's Team is predicted to have a phenomenal 2024 Olympics with a potential gold medal in the team competition. Below is a graph showing their predicted score in the finals compared to other top Nations.



5. **Conclusion**

5.1. Findings

The exhaustive exploration of data scraping, cleaning, exploratory data analysis (EDA), and modeling has unveiled significant insights into the performance dynamics of both Men's and Women's USA Olympic Artistic Gymnastics teams for the upcoming 2024 Paris Olympics. My operational definition of "best," emphasizing the maximization of the total medal count, has guided a meticulous investigation across individual apparatus outcomes, team compositions, and strategic positioning for Team USA.

In terms of individual events, the diverse machine learning models, encompassing Random Forest Classifier, AdaBoost Classifier, Support Vector Classifier, among others, showcased nuanced performances across various gymnastic disciplines. Results underscored the effectiveness of the Random Forest and AdaBoost classifiers, consistently displaying high accuracy and F2-scores, thereby affirming their robustness in the intricate process of selecting top-performing gymnasts. These models demonstrated both reliability in identifying winners and adaptability to handle inherent imbalances within gymnastics datasets.

On the men's side, the optimization algorithms, including Tabu Search, Hyper-Heuristic, Variable Neighborhood Search, and Brute Force, contributed valuable insights into the complexities of team composition. Results indicated that Tabu Search and Variable Neighborhood Search (VNS) algorithms, particularly when utilizing Individual All-Around (IAA) scores, emerged as the highest-scoring algorithms. These findings provide a comprehensive understanding of the intricate decision-making processes involved in the selection of gymnasts for the Men's Team USA.

Meanwhile, the evaluation of the Women's Team revealed a potential gymnastics powerhouse for the 2024 Olympics, projecting an impressive total of eight medals. This includes both individual accolades and team achievements. The amalgamation of individual gymnast predictions and team dynamics highlighted the multidimensional nature of gymnastics performance, showcasing the prowess of athletes like Simone Biles, Shilese Jones, and Jordan Chiles, among others.

5.2. Limitations

It is imperative to acknowledge the inherent constraints and limitations that accompany the study. This section aims to transparently delineate the boundaries and potential sources of bias within the analysis, providing a nuanced perspective on the scope and applicability of the findings. While the research contributes valuable insights to the field of gymnastics performance prediction, it is crucial to recognize the contextual limitations that may impact the generalizability and comprehensive understanding of our results. The following outlines the primary limitations inherent in the study.

This study does not consider the impact of injuries, or the physical toll associated with participating in numerous events within a condensed time frame. The omission of these factors may limit the comprehensive understanding of athletes' capabilities, as injuries and physical fatigue can significantly influence performance outcomes.

The complexity of team sports, such as gymnastics, extends beyond individual performances. This analysis does not account for the dynamics of team cohesion, which plays a crucial role in team-based events. The collaborative and synchronized nature of gymnastics competitions underscores the importance of team dynamics, an aspect not addressed in this study.

The study employed historical data for Men's and Women's gymnastics, with Women's models trained on a substantial dataset spanning from 2013 to 2021 and Men's models trained on a dataset from 2018 to 2020. The inclusion of data from the 2016 and 2020 Olympics provided a comprehensive testing ground. However, the discrepancy, resulting from time constraints, in the duration of data collection for Men's and Women's models may introduce bias, and a more balanced approach could enhance the study's robustness.

5.3. Future research

Conducting a more fine-grained analysis of individual features that contribute to gymnastic performance could refine the models. Investigating specific skills, techniques, or routine components and their impact on outcomes could lead to more targeted and actionable insights for athletes and coaches. Having a Data set not only with the athlete's name, scores, penalties, and dates, but also which skills were performed and at what points in their performance, as well as combination of skills and in which order.

6. References

1. https://saedsayad.com/zeror.htm
2. https://docs.h2o.ai/driverless-ai/1-10-lts/docs/userguide/scorers.html
3. https://www.nature.com/articles/nmeth.4346
4. https://dev.heuristiclab.com/trac.fcgi/wiki/Features

7. Data Sources