INCLUDE: doesn;t include injuries, physical toll of competition on mult. Events., team cohesion

Olympians who go with a team can also compete in individual events.

1. Abstract

The field of sports analytics, propelled by advancements in machine learning and data science, has revolutionized decision-making processes across various athletic disciplines. This research paper explores the comprehensive process of data scraping, cleaning, exploratory data analysis (EDA), and modeling, with the ultimate goal of selecting the best Men's and Women's Team and Individual USA Olympic Artistic Gymnasts. The notion of "best" is approached with flexibility, although, for the purposes of this paper, "best" is operationalized as maximizing the total medal count in artistic Gymnastics for the United States at the 2024 Olympic Games.  By embracing the power of data-driven decision-making, this research contributes to the evolving landscape of sports analytics and presents a comprehensive approach to assembling elite gymnastics teams for the United States. Ultimately, the objective is to equip Team USA with the talent, diversity, and strategies that will maximize their medal count and enhance the nation's standing in Olympic Artistic Gymnastics competitions.

Maximize Medal Count

1. Introduction

This research addresses the imperative task of assembling the optimal Men's and Women's Team USA Olympic Artistic Gymnastics teams for Paris 2024, with the primary objective of maximizing success. The challenge centers on identifying a group of five athletes for each team, leveraging analytics models to compare and forecast medal counts across the eight events for men and six events for women. The overarching goal is to strategically position Team USA for excellence in the forthcoming Olympic Games.

When determining the composition of the optimal Men's and Women's Team USA Olympic Artistic Gymnastics teams, a pivotal consideration in determining the “best”,  is the medal count. Establishing a baseline for individual event models involved implementing a rudimentary benchmark, employing the ZeroR algorithm. ZeroR, a simplistic baseline algorithm frequently used in machine learning for benchmarking, predicts the most frequent class or outcome within a dataset, regardless of its features or variables. The models under scrutiny in this study encompass a comprehensive list, including RandomForestClassifier, AdaBoostClassifier, Support Vector Classifier (SVC), KNeighborsClassifier, DecisionTreeClassifier, Gaussian Naive Bayes (GaussianNB), and Multi-Layer Perceptron (MLP) Classifier. Subsequently, a systematic grid search approach was employed to optimize the parameters of the best-performing models. Feature importance was systematically explored in the experimentation process to assess its potential impact on performance enhancements. Notably, the approach differed slightly when selecting the Men's and Women's gymnastics teams.

For the Women's teams, a range of algorithms, such as Constraint Satisfaction Problem (CSP), Hyper Heuristics, Stochastic Control, Tabu Search, Variable Neighborhood Search (VNS), and Brute Force algorithms, were considered.

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 in specific events. Analysis revealed several gymnasts likely to medal in individual events, some of whom may not align with the predicted Team USA lineup. Consequently, a team selection algorithm was applied to other countries that performed well in team qualification events for the Olympics, providing a guideline for estimating scores of top-performing nations. This knowledge aids in selecting gymnasts for Team USA, with the objective of optimizing individual medal potential while securing team success.

For the Men's teams, algorithms such as Tabu Search, Hyper-Heuristic, and Variable Neighborhood Search (VNS) were considered. The analytical process for the Men's team commenced similarly to the Women's team. Analysis indicates that the USA Men's team is likely not to medal but secure a 4th place finish in the team competition. With the focus shifted from building a robust team, the emphasis is now on identifying individuals with the potential to medal. Evaluation identified only two men with the potential to medal in individual events, both of whom were part of the team selected through the team modeling process. This approach still allows for the formation of a strong Team USA with potential medal prospects.

1. Data Flow

A diagram of a data flow

Description automatically generated

1. Data Collection and Preprocessing:
   1. Data Sources:

This research paper draws its data from a diverse array of sources, encompassing historical performance records and competition results. The primary sources of data include Thegymter.net and Wikipedia.org. The data spans from 2013 to 2023, primarily gathered from international competitions, world cups, world championships, the Olympics as well as data provided by the UConn Sports Analytics Symposium 2024. This comprehensive dataset includes individual scores for all apparatus and individual all-around scores. It also incorporates pertinent information such as gymnasts' rankings in various events and rounds, athletes' names, scores on each apparatus, execution scores, difficulty scores, penalties, and their final overall scores.

1. Data Preprocessing Steps:

Data preprocessing is a critical stage in the research, encompassing various essential procedures to ensure the integrity and utility of the dataset. The workflow begins in the 'Scraping' folder, where code is stored for web scraping, along with the resulting CSV files containing raw data. Subsequently, the data undergoes a series of transformations as it moves through different stages:

1. Cleaning: In the 'Cleaning' folder, the scraped CSV files are meticulously cleaned. This process involves tasks such as filling in or removing missing values to mitigate data inconsistencies. Additionally, it includes operations to standardize column names and convert nation names to their respective National Olympic Committee (NOC) codes, where necessary.
2. Data Integration: The preprocessed data then proceeds to the 'Combine Data' folder, where information from various competitions is merged. Data is consolidated based on key parameters such as the year, gender, and specific apparatus, providing a holistic view of the gymnastic performance dataset.
3. Feature Selection: An important aspect of data preprocessing is the selection of relevant features that contribute to model accuracy and effectiveness. Feature selection ensures that the most informative variables are considered in the subsequent analysis.
4. 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.
5. Methodology
   1. Selection of Machine Learning Algorithms

A diagram of a team

Description automatically generated

1. Individual Selection - Men's Individual Events

**Individual All-Around**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Classifier** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **Support** | **F2-Score** |
| Random Forest | 1 | 1 | 1 | 1 | 101 | 1 |
| AdaBoost | 0.9703 | 0.74 | 0.66 | 0.69 | 101 | 0.36 |
| SVM | 0.9703 | 0.49 | 0.5 | 0.49 | 101 | 0 |
| K-Nearest Neighbors | 0.9703 | 0.49 | 0.5 | 0.49 | 101 | 0 |
| Decision Tree | 0.9802 | 0.83 | 0.83 | 0.83 | 101 | 0.67 |
| Naive Bayes | 0.9703 | 0.75 | 0.98 | 0.83 | 101 | 0.83 |
| Neural Network | 0.9703 | 0.49 | 0.5 | 0.49 | 101 | 0 |

**Floor Event**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Classifier** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **Support** | **F2-Score** |
| Random Forest | 0.9815 | 0.98 | 1 | 0.99 | 108 | 0.38 |
| AdaBoost | 0.9815 | 0.98 | 1 | 0.99 | 108 | 0.38 |
| SVM | 0.9722 | 0.49 | 0.5 | 0.49 | 108 | 0 |
| K-Nearest Neighbors | 0.9722 | 0.49 | 0.5 | 0.49 | 108 | 0 |
| Decision Tree | 0.9722 | 0.74 | 0.66 | 0.69 | 108 | 0.36 |
| Naive Bayes | 0.9444 | 0.67 | 0.97 | 0.74 | 108 | 0.71 |
| Neural Network | 0.9722 | 0.49 | 0.5 | 0.49 | 108 | 0 |

**High Bar**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Classifier** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **Support** | **F2-Score** |
| Random Forest | 0.9802 | 0.98 | 1 | 0.99 | 101 | 0 |
| AdaBoost | 0.9802 | 0.74 | 0.74 | 0.74 | 101 | 0.5 |
| SVM | 0.9802 | 0.49 | 0.5 | 0.49 | 101 | 0 |
| K-Nearest Neighbors | 0.9802 | 0.49 | 0.5 | 0.49 | 101 | 0 |
| Decision Tree | 0.9901 | 0.99 | 0.75 | 0.83 | 101 | 0.56 |
| Naive Bayes | 0.8911 | 0.54 | 0.7 | 0.55 | 101 | 0.26 |
| Neural Network | 0.9703 | 0.49 | 0.49 | 0.49 | 101 | 0 |

**Parallel Bars**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Classifier** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **Support** | **F2-Score** |
| Logistic Regression | 0.9627 | 0.97 | 0.99 | 0.98 | 161 | 0 |
| XGBoost | 0.9752 | 0.98 | 0.99 | 0.97 | 161 | 0.4348 |
| Random Forest | 0.9752 | 0.97 | 1 | 0.98 | 161 | 0.2381 |
| AdaBoost | 0.9876 | 0.99 | 1 | 0.99 | 161 | 0.6522 |
| SVM | 0.9689 | 0.97 | 1 | 0.98 | 161 | 0 |
| K-Nearest Neighbors | 0.9689 | 0.97 | 1 | 0.98 | 161 | 0 |
| Decision Tree | 0.9938 | 1 | 0.99 | 1 | 161 | 0.9615 |
| Naive Bayes | 0.7702 | 1 | 0.76 | 0.87 | 161 | 0.4032 |
| Neural Network | 0.9689 | 0.97 | 1 | 0.98 | 161 | 0 |

**Pommel Horse**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Classifier** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **Support** | **F2-Score** |
| Random Forest | 0.9714 | 0.97 | 1 | 0.9857 | 70 | 0 |
| AdaBoost | 0.9857 | 0.99 | 1 | 0.99 | 70 | 0.5556 |
| SVM | 0.9714 | 0.97 | 1 | 0.9857 | 70 | 0 |
| K-Nearest Neighbors | 0.9714 | 0.97 | 1 | 0.9857 | 70 | 0 |
| Decision Tree | 0.9857 | 0.99 | 1 | 0.99 | 70 | 0.5556 |
| Naive Bayes | 0.9286 | 1 | 0.93 | 0.96 | 70 | 0.6667 |
| Neural Network | 0.9714 | 0.97 | 1 | 0.9857 | 70 | 0 |

**Rings**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Classifier** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **Support** | **F2-Score** |
| Random Forest | 1 | 1 | 1 | 1 | 124 | 1 |
| AdaBoost | 0.9919 | 0.67 | 1 | 0.8 | 124 | 0.9091 |
| SVM | 0.9839 | 0 | 0 | 0 | 124 | 0 |
| K-Nearest Neighbors | 0.9839 | 0 | 0 | 0 | 124 | 0 |
| Decision Tree | 0.9919 | 0.67 | 1 | 0.8 | 124 | 0.9091 |
| Naive Bayes | 0.9677 | 0.33 | 1 | 0.5 | 124 | 0.7143 |
| Neural Network | 0.9839 | 0 | 0 | 0 | 124 | 0 |

**Vault**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Classifier** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **Support** | **F2-Score** |
| Random Forest | 0.9524 | 0.95 | 1 | 0.98 | 42 | 0 |
| AdaBoost | 0.9048 | 0.25 | 0.5 | 0.33 | 42 | 0.4167 |
| SVM | 0.9524 | 0.95 | 1 | 0.98 | 42 | 0 |
| K-Nearest Neighbors | 0.9524 | 0.95 | 1 | 0.98 | 42 | 0 |
| Decision Tree | 0.881 | 0.2 | 0.5 | 0.29 | 42 | 0.3846 |
| Naive Bayes | 0.6429 | 0.12 | 1 | 0.21 | 42 | 0.4 |
| Neural Network | 0.9524 | 0.95 | 1 | 0.98 | 42 | 0 |

In this research, a variety of machine learning models were employed to predict individual event outcomes. These models were meticulously evaluated through the utilization of critical performance metrics such as confusion matrices, precision, recall, f1-score, accuracy, macro avg, and    weighted avg. The models assessed in this study encompassed a comprehensive list, including RandomForestClassifier, AdaBoostClassifier, Support Vector Classifier (SVC), KNeighborsClassifier, DecisionTreeClassifier, Gaussian Naive Bayes (GaussianNB), and Multi-Layer Perceptron (MLP) Classifier. Notably, the AdaBoostClassifier and DecisionTreeClassifier exhibited superior performance, emerging as the top-performing models.

1. Individual Selection - Women’s Individual Events

**Individual All-around**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Classifier** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **Support** | **F2-Score** |
| Random Forest | 0.982 | 0.98 | 1 | 0.99 | 167 | 0 |
| AdaBoost | 0.988 | 0.99 | 0.67 | 0.75 | 167 | 0.38 |
| SVM | 0.982 | 0.98 | 1 | 0.99 | 167 | 0 |
| K-Nearest Neighbors | 0.982 | 0.98 | 1 | 0.99 | 167 | 0 |
| Decision Tree | 0.9581 | 0.58 | 0.65 | 0.6 | 167 | 0.28 |
| Naive Bayes | 0.9042 | 0.58 | 0.95 | 0.61 | 167 | 0.48 |
| Neural Network | 0.982 | 0.98 | 1 | 0.99 | 167 | 0 |

**Balance Beam**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Classifier | Accuracy | Precision | Recall | F1-Score | Support | F2-Score |
| Random Forest | 0.9718 | 0.97 | 1 | 0.99 | 71 | 0 |
| AdaBoost | 1 | 1 | 1 | 1 | 71 | 1 |
| SVM | 0.9718 | 0.97 | 1 | 0.99 | 71 | 0 |
| K-Nearest Neighbors | 0.9718 | 0.97 | 1 | 0.99 | 71 | 0 |
| Decision Tree | 0.9859 | 0.99 | 1 | 0.99 | 71 | 0.5556 |
| Naive Bayes | 0.9437 | 0.99 | 0.96 | 0.97 | 71 | 0.4167 |
| Neural Network | 0.9718 | 0.97 | 1 | 0.99 | 71 | 0 |

**Uneven Bars**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Classifier** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **Support** | **F2-Score** |
| Random Forest | 0.9074 | 0.91 | 1 | 0.95 | 54 | 0 |
| AdaBoost | 0.9259 | 0.96 | 0.6 | 0.65 | 54 | 0.24 |
| SVM | 0.9074 | 0.91 | 1 | 0.95 | 54 | 0 |
| K-Nearest Neighbors | 0.9074 | 0.91 | 1 | 0.95 | 54 | 0 |
| Decision Tree | 0.9444 | 0.85 | 0.79 | 0.82 | 54 | 0.62 |
| Naive Bayes | 0.8148 | 0.67 | 0.9 | 0.69 | 54 | 0.71 |
| Neural Network | 0.9074 | 0.91 | 1 | 0.95 | 54 | 0 |

**Vault**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Classifier** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **Support** | **F2-Score** |
| Random Forest | 1 | 1 | 1 | 1 | 29 | 1 |
| AdaBoost | 0.9655 | 0.83 | 0.98 | 0.89 | 29 | 0.91 |
| SVM | 0.931 | 0.47 | 0.5 | 0.48 | 29 | 0 |
| K-Nearest Neighbors | 0.9655 | 0.98 | 0.75 | 0.82 | 29 | 0.56 |
| Decision Tree | 0.9655 | 0.83 | 0.98 | 0.89 | 29 | 0.91 |
| Naive Bayes | 0.7241 | 0.6 | 0.85 | 0.58 | 29 | 0.56 |
| Neural Network | 0.9655 | 0.98 | 0.75 | 0.82 | 29 | 0.56 |

**Floor**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Classifier** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **Support** | **F2-Score** |
| Random Forest | 0.9655 | 0.5 | 0.5 | 0.5 | 58 | 0.5 |
| AdaBoost | 0.9483 | 0.3333 | 0.5 | 0.4 | 58 | 0.4545 |
| SVM | 0.9655 | 0 | 0 | 0 | 58 | 0 |
| K-Nearest Neighbors | 0.9828 | 1 | 0.5 | 0.6667 | 58 | 0.5556 |
| Decision Tree | 0.9483 | 0.3333 | 0.5 | 0.4 | 58 | 0.4545 |
| Naive Bayes | 0.6379 | 0.0909 | 1 | 0.1667 | 58 | 0.3226 |
| Neural Network | 0.9483 | 0.3333 | 0.5 | 0.4 | 58 | 0.4545 |

**Women Most likely to Medal by Apparatus**

|  |  |
| --- | --- |
| **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 |

1. Team Selection Approach:

Two approaches:

1. The qualification round uses 4 up 3 counts, where 4 athletes will compete on each apparatus and the sum of the 3 best scores on each apparatus is used to see who will advance. Therefore the algorithms were used to pick 4 USA men's Gymnasts. With the goal of the algorithm being that a Gymnast cannot compete on the same apparatus more than once. And each apparatus needs to be competed on 3 times each, then get the largest sum of the top 3 out of 4 scores on each apparatus. The finals format uses a 3 up 3 count. So each athlete must compete on each apparatus and all of their scores will be used. Therefore I picked the gymnasts with the highest Individual all-around scores. If there is a gymnast who has a higher IAA score than the top 3 from the 4 selected from qualifications, then tha athlete will be added to the finals.

Tabu Search, Hyper-Heuristic, and Variable Neighborhood Search (VNS) algorithms were used with USA Men’s gymnastics team selection. With each algorithm 2 different data sets were used, each using data from 2022 and 2023: 1. Individual all-around scores from USA gymnasts. 2. Individual apparatus scores for USA gymnasts. If an individual had more than one entry in either of the data sets, their highest scores were used. 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.

With 4 of the 5 athletes selected from the Tabu Search, the last gymnast may be chosen from the top IAA scores from the last year. The top 3 IAA scores from 2023 are Yul Moldauer, Fred Richard, and Vitaliy Guimaraes. Therefore our 5 person team is Khoi Young, Asher Hong, Fred Richard, Yul Moldauer, and Vitaliy Guimaraes. Qualifying Gymnasts are Khoi Young, Asher Hong, Fred Richard, and Yul Moldauer. And gymnasts to compete in the Final being Fred Richard, Yul Moldauer, and Vitaliy Guimaraes.

Men's Team Selection:

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **IAA/Apparatus** | **Score** |
| Tabu Search | IAA | 260.579000 |
| Tabu Search | Apparatus | 253.195987 |
| Hyper-Heuristics | IAA | 254.940000 |
| Hyper-Heuristics | Apparatus | 250.588687 |
| Variable Neighborhood Search (VNS) | IAA | 257.342000 |
| Variable Neighborhood Search (VNS) | Apparatus | 251.115667 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Hyper-Heuristics** | **Tabu Search** | **Variable Neighborhood Search (VNS)** | **Sum** |
| khoi young | 1 | 2 | 1 | 4 |
| asher hong | 1 | 1 | 1 | 3 |
| brody malone | 1 | 1 | 1 | 3 |
| yul moldauer | 1 | 1 | 1 | 3 |
| fred richard | 0 | 2 | 1 | 3 |
| paul juda | 1 | 1 | 0 | 2 |
| fuzzy benas | 1 | 0 | 1 | 2 |
| ian lasic-ellis | 1 | 0 | 0 | 1 |
| donnell whittenburg | 0 | 0 | 1 | 1 |
| joshua karnes | 1 | 0 | 0 | 1 |
| shane wiskus | 0 | 0 | 1 | 1 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Athlete** | **FX** | **PH** | **SR** | **VT** | **PB** | **HB** | **AA** |
| yul moldauer | 14.400 | 13.957 | 14.116 | 14.250 | 15.543 | 13.500 | 85.342 |
| fred richard | 14.483 | 13.357 | 13.350 | 14.000 | 14.450 | 14.962 | 84.602 |
| vitaliy guimaraes | 14.250 | 14.350 | 13.600 | 14.450 | 13.900 | 13.400 | 83.950 |
| khoi young | 13.550 | 15.342 | 12.550 | 15.244 | 13.750 | 13.100 | 83.536 |
| asher hong | 14.029 | 12.949 | 14.566 | 16.680 | 14.700 | 13.650 | 83.029 |

The Women's Team has the addition of Brute force algorithms, which didn’t work well for the men's team selection. 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 Variable Neighborhood Search VNS algorithm.

**Women's Team selection:**

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **IAA/Apparatus** | **Total Score** |
| Tabu Search | IAA | 171.337250 |
| Tabu Search | Apparatus | 170.798117 |
| Variable Neighborhood Search (VNS) | IAA | 171.470750 |
| Variable Neighborhood Search (VNS) | Apparatus | 166.038475 |
| Hyper-Heuristics | IAA | 169.963000 |
| Hyper-Heuristics | Apparatus | 166.263283 |
| brute force | IAA | 169.963000 |
| brute force | Apparatus | 169.120117 |
| Stochastic control | IAA | 156.995500 |
| Stochastic control | Apparatus | 141.703750 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Constraint Satisfaction Problem (CSP)** | **Hyper - Heuristics** | **Stochastic control** | **Tabu Search** | **Variable Neighborhood Search (VNS)** | **brute force** | **Sum** |
| simone biles | 1 | 2 | 1 | 2 | 2 | 2 | 10 |
| shilese jones | 1 | 2 | 0 | 2 | 1 | 2 | 8 |
| jade carey | 1 | 1 | 1 | 2 | 1 | 2 | 8 |
| jordan chiles | 1 | 0 | 1 | 1 | 1 | 1 | 5 |
| konnor mcclain | 0 | 0 | 0 | 1 | 1 | 0 | 2 |
| addison fatta | 0 | 0 | 1 | 0 | 0 | 1 | 2 |
| ashlee sullivan | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| chloe cho | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| kayla dicello | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| skye blakely | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| kelise woolford | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| joscelyn roberson | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| leanne wong | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| christiane popovich | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| annalisa milton | 0 | 0 | 1 | 0 | 0 | 0 | 1 |

|  |  |  |
| --- | --- | --- |
| **Athlete** | **Nation** | **Score** |
| Simone Biles | United States | 59.3 |
| Shilese Jones | United States | 57.8 |
| Leanne Wong | United States | 55.75 |

The Team selection used two different data sets. The first was using scores from 20XX to 20XX  from USA women's Individual All-Around event scores, which is denoted in the datasets as IAA. The second dataset used was from 20XX TO 20XX using USA individual women's scores from all the apparatus. Each of the data sets was used with the TABU Search algorithm,  Hyper-Heuristics algorithm, Variable Neighborhood Search algorithm, Brute force Search algorithm,  and Mixed-Integer Linear Programming (MILP) algorithm. I also experimented using the Harmony Search algorithm and Simulated Annealing algorithm.

 The Harmony Search (HS) method is an emerging metaheuristic optimization algorithm, which has been employed to cope with numerous challenging tasks during the past decade1. Despite best efforts, HS was unable to narrow the team down to just 4 athletes.

Simulated Annealing (SA) is an effective and general form of optimization.  It is useful in finding global optima in the presence of large numbers of local optima.2. SA wouldn’t return consistent results. After adding a random seed to improve the reliability of the results, the quality of the results suffered and I decided not to be included in the research.

Because the final of the Team competition is 3 up 3 count, meaning that 3 athletes will compete on all events and all of the scores are used to determine their final score, it was decided to use the 3 athletes with the highest IAA scores.

1. The Qualification round uses the 4 up, 3 count mentioned earlier. I used algorithms to pick gymnasts for a 5 up 4 count, to include a good alternative if needed. The final round is a 3 up, 3 count and again chose the top 3 Individual All-around gymnasts from the 5 gymnasts.
2. Hyperparameter Tuning
   1. Grid search
3. Cross-Validation Strategies
   1. 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.
4. . Evaluation and Results
   1. Individual Events Selection
      1. Data for individual events were run through the following models and then checked for accuracy, precision and recall, F1-Score, and Confusion matrix.
         1. RandomForestClassifier
         2. AdaBoostClassifier
         3. SVC
         4. KNeighborsClassifier
         5. DecisionTreeClassifier
         6. GaussianNB
         7. MLPClassifier

1. Base model
2. Confusion Matrix: I analyze the confusion matrix to gain insights into false positives.
   1. Accuracy: Accuracy measures the proportion of correctly predicted outcomes among all predictions. It provides an overall assessment of model correctness.
   2. Precision and Recall: Precision measures the proportion of true positive predictions among all positive predictions, while recall measures the proportion of true positives among all actual positives. These metrics are particularly useful in assessing the model's ability to identify winners accurately.
   3. F1-Score: The F1-Score combines precision and recall to provide a balanced measure of a model's performance. It is especially valuable when dealing with imbalanced datasets.
   4. F2-Score:
3. Team Selection
   1. For the team selection the algorithms were used to pick 4 USA Gymnasts. Where the Gymnasts cannot compete on the same apparatus more than once, and each apparatus needs to be competed on 3 times each. Then get the largest sum of the top 3 out of 4 scores on each apparatus.
   2. The Team selection used two different data sets. The first was using scores from 20XX to 20XX  from USA women's Individual All-Around event scores, which is denoted in the datasets as IAA. The second dataset used was from 20XX TO 20XX using USA individual women's scores from all the apparatus. Each of the data sets was used with the TABU Search algorithm,  Hyper-Heuristics algorithm, Variable Neighborhood Search algorithm, Brute force Search algorithm,  and Mixed-Integer Linear Programming (MILP) algorithm. I also experimented using the Harmony Search algorithm and Simulated Annealing algorithm.
   3. The Harmony Search (HS) method is an emerging metaheuristic optimization algorithm, which has been employed to cope with numerous challenging tasks during the past decade1. Despite best efforts, HS was unable to narrow the team down to just 4 athletes.
   4. Simulated Annealing (SA) is an effective and general form of optimization.  It is useful in finding global optima in the presence of large numbers of local optima.2. SA wouldn’t return consistent results. After adding a random seed to improve the reliability of the results, the quality of the results suffered and was chosen not to be included in the research.
   5. Because the final of the Team competition consists of all scores from 3 different athletes competing on all events, it was decided to use the 3 athletes with the highest IAA scores.
4. Present the results of the classification models.
5. Discuss the model's accuracy and predictive power.
6. Evaluate the trade-offs between different models and feature selections.
7. Provide insights into which athletes are predicted to perform well in gymnastics events.
8. The "Evaluation and Results" section of an academic paper should provide a comprehensive presentation of the outcomes of your analysis and the assessment of your models. It should include the following components:

**Evaluation and Results:**

|  |  |  |
| --- | --- | --- |
| **MAG/WAG** | **Apparatus** | **Classifier** |
| MAG | Pommel Horse | AdaBoostClassifier |
| MAG | Parallel Bars | XGBClassifier |
| MAG | Vault | AdaBoost Classifier |
| MAG | Rings | AdaBoost Classifier |
| MAG | High Bar | AdaBoost Classifier |
| MAG | Floor | AdaBoostClassifier |
| MAG | Individual All-around | AdaBoost Classifier |
| WAG | Floor | AdaBoost Classifier |
| WAG | Vault | RandomForestClassifier |
| WAG | Uneven Bars | DecisionTreeClassifier |
| WAG | Balance Beam | AdaBoostClassifier |
| WAG | Individual All-around | AdaBoostClassifier |

**Best USA Women's Gymnasts by Event:**

* Individual All-around: Jordan Chiles.
* Vault: Simone Biles
* Uneven Bars: Shilese Jones and Zoe Miller.

* Floor: Kaliya Lincoln.
* Balance Beam: Skye Blakely.

These athletes stand out as the top performers in their respective events based on the analysis.

**USA Women's Gymnastics Team:**

| **Algorithm** | **IAA/Apparatus** | **Total Score** |
| --- | --- | --- |
| Tabu Search | IAA | 171.337250 |
| Tabu Search | Apparatus | 170.798117 |
| Variable Neighborhood Search (VNS) | IAA | 171.470750 |
| Variable Neighborhood Search (VNS) | Apparatus | 166.038475 |
| Hyper-Heuristics | IAA | 169.963000 |
| Hyper-Heuristics | Apparatus | 166.263283 |
| brute force | IAA | 169.963000 |
| brute force | Apparatus | 169.120117 |
| Stochastic control | IAA | 156.995500 |
| Stochastic control | Apparatus | 141.703750 |
| Constraint Satisfaction Problem (CSP) | IAA | 171.337250 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Constraint Satisfaction Problem (CSP)** | **Hyper- Heuristics** | **Stochastic control** | **Tabu Search** | **Variable Neighborhood Search (VNS)** | **brute force** | **Sum** |
| Simone biles | 1 | 2 | 1 | 2 | 2 | 2 | 10 |
| Shilese jones | 1 | 2 | 0 | 2 | 1 | 2 | 8 |
| Jade carey | 1 | 1 | 1 | 2 | 1 | 2 | 8 |
| Jordan chiles | 1 | 0 | 1 | 1 | 1 | 1 | 5 |
| Konnor Mcclain | 0 | 0 | 0 | 1 | 1 | 0 | 2 |
| Addison fatta | 0 | 0 | 1 | 0 | 0 | 1 | 2 |
| Ashlee sullivan | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| Chloe cho | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| Kayla dicello | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| Skye blakely | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| Kelise woolford | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| Joscelyn roberson | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| Leanne wong | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| Christiane popovich | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| Annalisa Milton | 0 | 0 | 1 | 0 | 0 | 0 | 1 |

**Men's Team**

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **IAA/Apparatus** | **Score** |
| Tabu Search | IAA | 260.579000 |
| Tabu Search | Apparatus | 253.195987 |
| Hyper-Heuristics | IAA | 254.940000 |
| Hyper-Heuristics | Apparatus | 250.588687 |
| Variable Neighborhood Search (VNS) | IAA | 257.342000 |
| Variable Neighborhood Search (VNS) | Apparatus | 251.115667 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Hyper-Heuristics** | **Tabu Search** | **Variable Neighborhood Search (VNS)** | **Sum** |
| khoi young | 1 | 2 | 1 | 4 |
| asher hong | 1 | 1 | 1 | 3 |
| brody malone | 1 | 1 | 1 | 3 |
| yul moldauer | 1 | 1 | 1 | 3 |
| fred richard | 0 | 2 | 1 | 3 |
| paul juda | 1 | 1 | 0 | 2 |
| fuzzy benas | 1 | 0 | 1 | 2 |
| ian lasic-ellis | 1 | 0 | 0 | 1 |
| donnell whittenburg | 0 | 0 | 1 | 1 |
| joshua karnes | 1 | 0 | 0 | 1 |
| shane wiskus | 0 | 0 | 1 | 1 |

The analysis yielded a refined selection for the USA Women's Gymnastics team, consisting of the following athletes in their respective events:

* Team: Simone Biles, Skye Blakely, Shilese Jones, Joscelyn Roberson, and Zoe Miller.
* Balance Beam: Simone Biles, Skye Blakely, and Shilese Jones.
* Uneven Bars: Skye Blakely, Shilese Jones, and Zoe Miller.
* Floor: Simone Biles, Shilese Jones, and Joscelyn Roberson.
* Vault: Simone Biles, Shilese Jones, and Joscelyn Roberson.

The expected team score was calculated to be approximately 169.375, just slightly below the winning score of 169.528 at the 2020 Olympics, with second place achieving 166.096. These selections represent the best composition of the USA Women's Gymnastics team based on the analysis.

**USA Men's Gymnastics Team:**

For the USA Men's Gymnastics team, the analysis led to the selection of the following athletes and their respective events:

* Team: Colt Walker, Khoi Young, Asher Hong, Curran Phillips, and Stephen Nedoroscik.
* Floor: Colt Walker, Asher Hong, and Khoi Young.
* High Bar: Khoi Young, Colt Walker, and Curran Phillips.
* Parallel Bars: Asher Hong, Colt Walker, and Curran Phillips.
* Pommel Horse: Asher Hong, Khoi Young, and Stephen Nedoroscik.
* Rings: Khoi Young, Colt Walker, and Asher Hong.
* Vault: Colt Walker, Curran Phillips, and Khoi Young.

The combined score for the team was calculated at approximately 251.023, resulting in a 4th place finish in the 2020 Olympics.

USA Men's Gymnasts Most Likely to Medal:

Based on the analysis, the following athletes are most likely to achieve medals in their respective events:

* Parallel Bars: Curran Phillips.
* Pommel Horse: Stephen Nedoroscik and Khoi Young.

These athletes exhibit strong potential for medal-winning performances in their respective events.

By incorporating these findings, you can provide a detailed and clear representation of the outcomes of your analysis, offering valuable insights into the selection of both the USA Women's and Men's Gymnastics teams for the Olympics. This section should help readers understand the rationale behind your selections and their expected performance.

Model Training and Evaluation

All of the following shows best with results of either Random Forest, AdaBoost, or Decision Tree.

* Women's uneven bar
* Women's BALANCE BEAM
* Women's INDIVIDUAL ALL-AROUND
* Women's UNEVEN BAR
* Women's VAULT
* Men's FLOOR
* Men's HIGH BAR
* clf = AdaBoostClassifier(n\_estimators=100, learning\_rate=0.41)

* Men's INDIVIDUAL ALL-AROUND
* Men's PARALLEL BARS
* Men's RINGS
* Men's VAULT
* Men's POMMEL HORSE