Data-Driven Team Selection Model for Paris Olympic Gymnastics

Abstract

This article presents a comprehensive overview of the UCSAS 2024 project, a pioneering initiative in data-driven decision-making for the selection of Olympic gymnastics teams. At the core of this project lies a sophisticated analytical model that seamlessly integrates a range of advanced statistical and computational techniques. These include Kernel Density Estimation (KDE), Gaussian Distribution, K-Nearest Neighbors (KNN), and a Genetic Algorithm, all meticulously combined to refine and optimize the team selection process.

The model's methodology is divided into three primary processes: data preparation, distribution estimation, and the selection of the best team. The initial phase involves extensive data gathering, cleaning, and structuring, laying the groundwork for accurate analysis. The second phase focuses on estimating distributions using both athletes' own historical data and that of their nearest neighbors, allowing for a more comprehensive evaluation of each athlete's potential.

The final phase is the core of the model, where the genetic algorithm is applied to select the optimal team composition. This phase includes initial filtering of athletes based on performance metrics, followed by rigorous simulations to determine the best team combinations. The model also emphasizes performance optimization and validation, ensuring reliability and efficiency in simulations.

Finally, the optimal team selected for the Paris Olympics would be as follows:

Women: Biles Simone, Lee Sunisa, Lincoln Kaliya, McClain Konnor, Skinner Mykayla.

Men: Hale Dallas, Mikulak Samuel, Richard Frederick Nathaniel, Wiskus Shane, Yoder Alec.

Key Words: Kernel Density Estimation (KDE), Gaussian Distribution, K-Nearest Neighbors (KNN), Genetic Algorithm, Parallel Processing

Content

1. Introduction 3
1.1. A Brief Introduction of UCSAS 2024 3
1.2. Analysis Approach 4
2. Process I: Data Preparation 6
2.1. Data Supplement 6
2.2. Data Cleaning 6
2.3. Date Construction 7
3. Process II: Distribution Estimation 8
3.1. KDE Analysis Using Self-Historical Data 8
3.2. Gaussian Distribution Analysis Using Self-Historical Data 8
3.3. KDE Analysis Utilizing Nearest Neighbor Historical Data 10
3.4. Integration of Three Methods for Score Estimation 10
4. Process III: Best Team Selection 12
4.1. Initial Filtering 12
4.2. Genetic Algorithm 13
4.3. Final Comparison 15
5. Feasibility Exploration 16
6. Code Efficiency Optimization 17
7. Summary 17

1. Introduction

1.1. A Brief Introduction of UCSAS 2024

The UCSAS 2024 USOPC Data Challenge aims to select the most effective Team USA Olympic Men's and Women's Artistic Gymnastics teams for the Paris 2024 Olympics. We are tasked with using a rich dataset from major competitions, including scores from the 2017-2023 seasons, to develop models that predict medal counts and optimize team success. Our goal is to analyze this data to identify potential athletes, focusing on constructing a team that maximizes medal prospects.
These are some samples of provided data:

Table 1: Samples of provided data in UCSAS (data_2017_2021.csv).

LastName	FirstName	Gender	Country	Date	Competition	Round	Location	Apparatus	Rank	D_Score	E_Score	Penalty	Score
ABDUL	HADI	w	MAS	SUN 25 JUI	Olympic Games	qual	Tokyo, Jap	BB	76	4.8	6.766		11.566
ABDUL	HADI	w	MAS	SUN 25 JUI	Olympic Games	qual	Tokyo, Jap	FX	64	4.6	7.633		12.233
ABDUL	HADI	w	MAS	SUN 25 JUI	Olympic Games	qual	Tokyo, Jap	UE	74	5.1	6.5		11.6
ABDUL	HADI	w	MAS	SUN 25 JUI	Olympic Games	qual	Tokyo, Jap	VT	72	4.6	8.566		13.166
ABDUL HADI	Farah Ann	w	MAS	25-Jul-21	Olympic Games	qual	Tokyo, Jap	BB	76	4.8	6.766		11.566
ABDUL HADI	Farah Ann	w	MAS	25-Jul-21	Olympic Games	qual	Tokyo, Jap	FX	64	4.6	7.633		12.233
ABDUL HADI	Farah Ann	w	MAS	25-Jul-21	Olympic Games	qual	Tokyo, Jap	UE	74	5.1	6.5		11.6
ADLERTEG	Jonna	w	SWE	25-Jul-21	Olympic Games	qual	Tokyo, Jap	UE	12	6.3	8.233		14.533
AKHAIMOVA	Liliia	w	ROC	SUN 25 JUI	Olympic Games	qual	Tokyo, Jap	BB	59	5.1	7.166		12.266
AKHAIMOVA	Liliia	w	ROC	25-Jul-21	Olympic Games	qual	Tokyo, Jap	BB	59	5.1	7.166		12.266
AKHAIMOVA	Liliia	w	ROC	SUN 25 JUI	Olympic Games	qual	Tokyo, Jap	FX	11	5.8	7.833		13.633
AKHAIMOVA	Liliia	w	ROC	25-Jul-21	Olympic Games	qual	Tokyo, Jap	FX	11	5.8	7.833		13.633
AKHAIMOVA	Liliia	w	ROC	SUN 25 JUI	Olympic Games	qual	Tokyo, Jap	UE	49	5.4	7.5		12.9
AKHAIMOVA	Liliia	w	ROC	25-Jul-21	Olympic Games	qual	Tokyo, Jap	UE	49	5.4	7.5		12.9
AKHAIMOVA	Liliia	w	ROC	SUN 25 JUI	Olympic Games	qual	Tokyo, Jap	VT	10	5.8	8.966		14.766
AKHAIMOVA	Liliia	w	ROC	25-Jul-21	Olympic Games	qual	Tokyo, Jap	VT1	7	5.8	8.966		14.766
AKHAIMOVA	Liliia	w	ROC	1-Aug-21	Olympic Games	final	Tokyo, Jap	VT1	6	5.8	8.866	0.1	14.666
AKHAIMOVA	Liliia	w	ROC	25-Jul-21	Olympic Games	qual	Tokyo, Jap	VT2	7	5.6	9.033		14.633
AKHAIMOVA	Liliia	w	ROC	1-Aug-21	Olympic Games	final	Tokyo, Jap	VT2	6	5.6	9.066		14.666
ALVARADO	Luciana	w	CRC	SUN 25 JUI	Olympic Games	qual	Tokyo, Jap	BB	37	5.3	7.666		12.966
ALVARADO	Luciana	w	CRC	25-Jul-21	Olympic Games	qual	Tokyo, Jap	BB	37	5.3	7.666		12.966
ALVARADO	Luciana	w	CRC	SUN 25 JUI	Olympic Games	qual	Tokyo, Jap	FX	66	4.7	7.466		12.166
ALVARADO	Luciana	w	CRC	25-Jul-21	Olympic Games	qual	Tokyo, Jap	FX	66	4.7	7.466		12.166
ALVARADO	Luciana	w	CRC	SUN 25 JUI	Olympic Games	qual	Tokyo, Jap	UE	55	4.9	7.841		12.741

Table 2: Samples of provided data in UCSAS (data_2022_2023.csv).

LastName	FirstName	Gender	Country	Date	Competitic	Round	Location	Apparatus	Rank	D_Score	E_Score	Penalty	Score
AAS	Fredrik	m	NOR	23-26 Feb	FIG Appara	qual	Cottbus, Germany	HB	39	4.6	6.7		11.3
AAS	Fredrik	m	NOR	23-26 Feb	FIG Appara	qual	Cottbus, Germany	PH	44	4.4	7.8		12.2
AAS	Fredrik Bjornevik	m	NOR	1-5 Aug 20	2023 FISU	qual	Chengdu, China	FX	54	4	8.566		12.566
AAS	Fredrik Bjornevik	m	NOR	1-5 Aug 20	2023 FISU	qual	Chengdu, China	HB	44	4.6	8.166		12.766
AAS	Fredrik Bjornevik	m	NOR	1-5 Aug 20	2023 FISU	qual	Chengdu, China	PB	54	4.4	8.066		12.466
AAS	Fredrik Bjornevik	m	NOR	1-5 Aug 20	2023 FISU	qual	Chengdu, China	PH	58	4.4	7.266		11.666
AAS	Fredrik Bjornevik	m	NOR	1-5 Aug 20	2023 FISU	qual	Chengdu, China	SR	73	3.5	8.366		11.866
AAS	Fredrik Bjornevik	m	NOR	1-5 Aug 20	2023 FISU	qual	Chengdu, China	VT1	69	4	8.966		12.966
ABAD	Nestor	m	ESP	29 Oct 202	2022 51st I	AAfinal	Liverpool, England	FX	13	5.5	8.3		13.8
ABAD	Nestor	m	ESP	29 Oct 202	2022 51st I	qual	Liverpool, England	FX	48	5.5	7.9		13.4
ABAD	Nestor	m	ESP	29 Oct 202	2022 51st I	AAfinal	Liverpool, England	HB	10	5.3	8.133		13.433
ABAD	Nestor	m	ESP	29 Oct 202	2022 51st I	qual	Liverpool, England	HB	48	5	8.166		13.166
ABAD	Nestor	m	ESP	29 Oct 202	2022 51st I	AAfinal	Liverpool, England	PB	12	5.8	8.4		14.2
ABAD	Nestor	m	ESP	29 Oct 202	2022 51st I	qual	Liverpool, England	PB	26	5.8	8.5		14.3
ABAD	Nestor	m	ESP	29 Oct 202	2022 51st I	AAfinal	Liverpool, England	PH	21	5.1	6.166		11.266
ABAD	Nestor	m	ESP	29 Oct 202	2022 51st I	qual	Liverpool, England	PH	55	5.2	7.566		12.766
ABAD	Nestor	m	ESP	29 Oct 202	2022 51st I	AAfinal	Liverpool, England	SR	12	5.3	7.933		13.233
ABAD	Nestor	m	ESP	29 Oct 202	2022 51st I	qual	Liverpool, England	SR	90	4.9	7.666		12.566
ABAD	Nestor	m	ESP	29 Oct 202	2022 51st I	AAfinal	Liverpool, England	VT	20	4.8	9.033	0.1	13.733
ABAD	Nestor	m	ESP	29 Oct 202	2022 51st I	qual	Liverpool, England	VT1	97	4.8	9	0.3	13.5
ABAD	Nestor	m	ESP	18-21 Aug	2022 Senic	qual	Munich, Germany	FX	86	5.4	7.133	0.1	12.433

1.2. Analysis Approach

Figure 1 below provides a detailed activity flow showing our analysis approach for this project. At a high level, our approach consists of constructing a probability density curve for each apparatus for each athlete and then inputting these distributions into a simulation algorithm to estimate the highest probability overall scores for the various possible men's and women's Olympic teams.
Thus, we faced two fundamental challenges in performing this analysis: how can we generate the most accurate estimates of the probability distributions and, given those distributions, how can we select the best combination of athletes in a computationally tractable manner.

Probability distribution estimation

The inputs to our probability distribution estimation methodology are all the scores from the previous competitions. A challenge is that some athletes have a significant number of scores while others have only one or two. The approach used is to generate and average together three different density curves for each athlete/apparatus as summarized below and depicted in the "Score Distribution Estimation" box in Figure 1:

- *Individual Kernel Density Estimates (KDE)*. Kernel Density Estimation is a non-parametric technique used to estimate the density function of a dataset by placing a continuous kernel function at each data point and summing these functions, resulting in a smoothed, continuous version of a histogram.
- *Averaging KDEs of nearest neighbors*. Because the number of data points for many athlete/apparatus pairs is too small to estimate all but the simplest distributions, a distribution is estimated by averaging the KDEs of each athlete/apparatus' x nearest neighbors. The nearest neighbor determinations are generating using the dataset shown in Table 4.
- *Gaussian distribution*. Finally, a parametric Gaussian distribution curve was generated for each athlete/apparatus pair by estimating a mean and variance using the approach described in section 3.2.

Team selection

After generating estimated probability distribution curves for each athlete/apparatus pair (by averaging the three distributions described above), the next step is to select the combination of male and female athletes that would maximize the expected value of the equation $3 * NumGoldMedals + 2 * NumSilverMedals + NumBronzeMedals$ where the medal counts are the number of medals the team would have received in the Tokyo Olympics (with adjustments for rule changes between the Tokyo and Paris Olympics).

To generate and score every possible team combination of US athletes would be computationally intractable, particularly with our non-parametric distribution curves, so simulations are performed and a genetic algorithm (GA) is used to explore the large space of team combinations and identify the team of athletes that is most likely to achieve the highest overall medal score. This algorithm is described in some detail in Section 4 of this report.

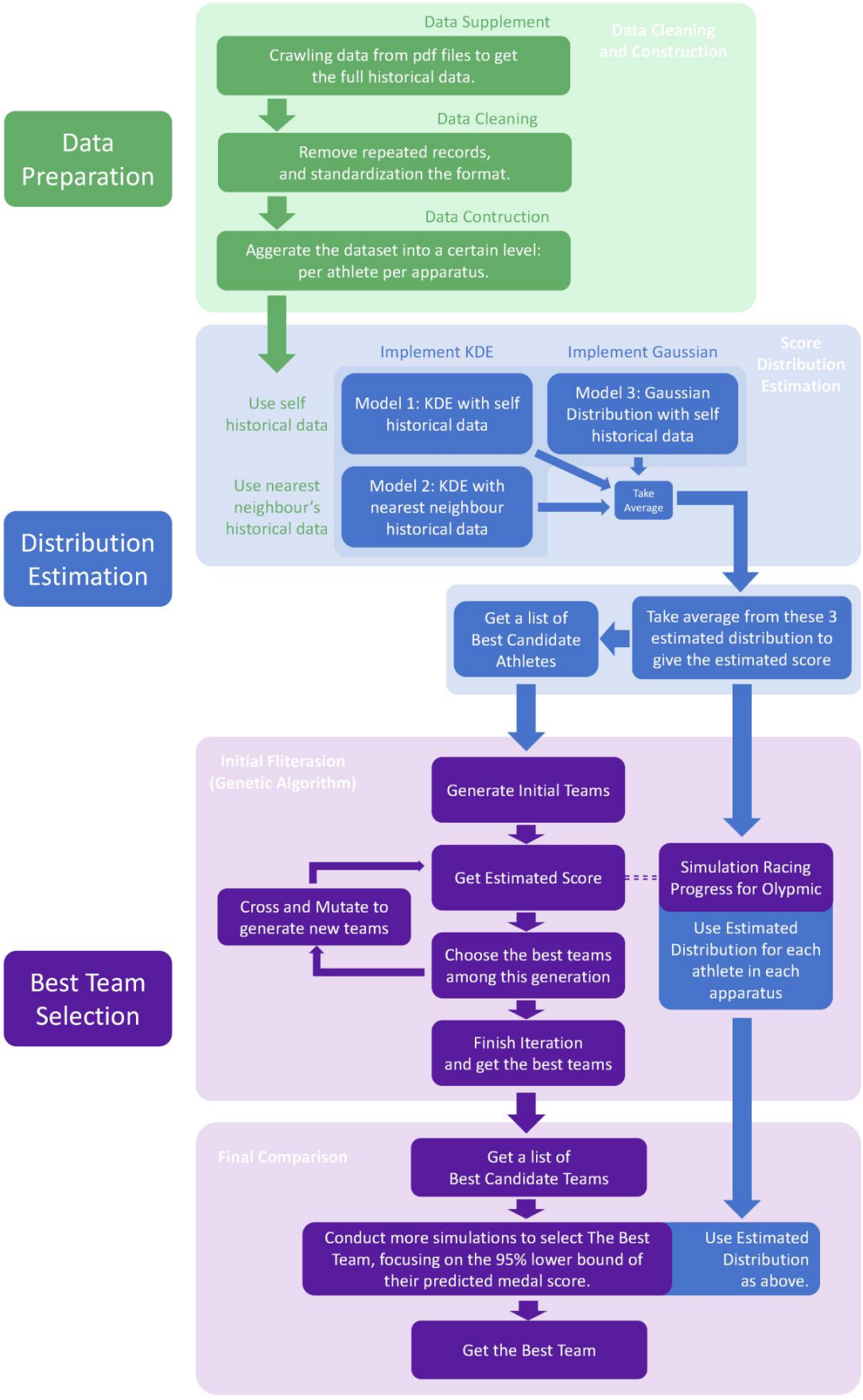
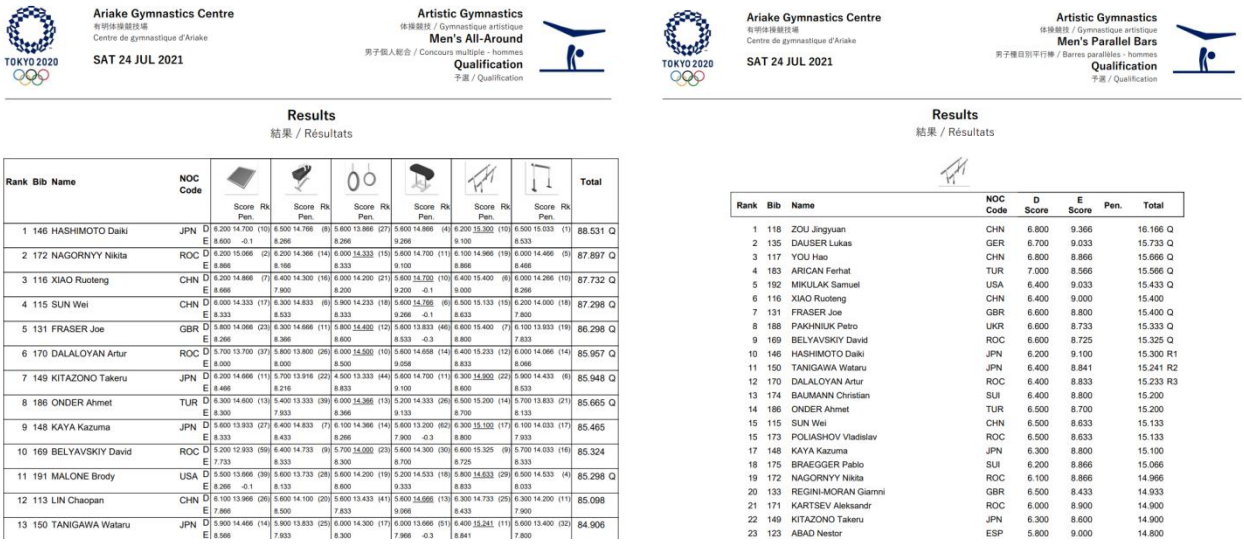


Figure 1: General flow chart.

2. Process I: Data Preparation

2.1. Data Supplement

To complement our model, we needed comprehensive data for the Tokyo Olympics. We sourced additional data for male athletes from 'https://gymnasticsresults.com/results/2021/olympics/'. By developing a Python-based data crawler, we extracted essential records of qualifications and finals across all apparatuses and the all-around match for male categories.
This process added 963 records to our original dataset.



Ariake Gymnastics Centre
SAT 24 JUL 2021
Artistic Gymnastics
Men's All-Around
Qualification
Results

Rank	Bib	Name	NOC Code	Total
1	146	HASHIMOTO Daiki	JPN	88.531 Q
2	172	NAGORNYY Nikita	ROC	87.897 Q
3	116	XIAO Ruoteng	CHN	87.732 Q
4	115	SUN Wei	CHN	87.298 Q
5	131	FRASER Joe	GBR	86.298 Q
6	170	DALALOYAN Artur	ROC	85.957 Q
7	149	KITAZONO Takeru	JPN	85.948 Q
8	186	ONDER Ahmet	TUR	85.665 Q
9	148	KAYA Kazuma	JPN	85.465
10	169	BELYAVSKIY David	ROC	85.324
11	191	MALONE Brody	USA	85.298 Q
12	113	LIN Chaopan	CHN	85.098
13	150	TANIGAWA Wataru	JPN	84.906

Ariake Gymnastics Centre
SAT 24 JUL 2021
Artistic Gymnastics
Men's Parallel Bars
Qualification
Results

Rank	Bib	Name	NOC Code	D Score	E Score	Pen.	Total
1	118	ZOU Jingyuan	CHN	6.800	9.366		16.166 Q
2	135	DAUSER Lukas	GER	6.700	9.033		15.733 Q
3	117	YOU Hao	CHN	6.800	8.866		15.666 Q
4	183	ARICAN Ferhat	TUR	7.000	8.566		15.566 Q
5	192	MIKULAK Samuel	USA	6.400	9.033		15.433 Q
6	116	XIAO Ruoteng	CHN	6.400	9.000		15.400
7	131	FRASER Joe	GBR	6.600	8.800		15.400 Q
8	188	PAKHNIUK Petro	UKR	6.600	8.733		15.333 Q
9	169	BELYAVSKIY David	ROC	6.600	8.725		15.325 Q
10	146	HASHIMOTO Daiki	JPN	6.200	9.100		15.300 R1
11	150	TANIGAWA Wataru	JPN	6.400	8.841		15.241 R2
12	170	DALALOYAN Artur	ROC	6.400	8.833		15.233 R3
13	174	BAUMANN Christian	SUI	6.400	8.800		15.200
14	186	ONDER Ahmet	TUR	6.500	8.700		15.200
15	115	SUN Wei	CHN	6.500	8.633		15.133
15	173	POLIASHOV Vladislav	ROC	6.500	8.633		15.133
17	148	KAYA Kazuma	JPN	6.300	8.800		15.100
18	175	BRAEGGER Pablo	SUI	6.200	8.866		15.066
19	172	NAGORNYY Nikita	ROC	6.100	8.866		14.966
20	133	REGINI-MORAN Giarnni	GBR	6.500	8.433		14.933
21	171	KARTSEV Aleksandr	ROC	6.000	8.900		14.900
22	149	KITAZONO Takeru	JPN	6.300	8.600		14.900
23	123	ABAD Nestor	ESP	5.800	9.000		14.800

Figure 2: The PDF file containing additional information about the Tokyo Olympics.

2.2. Data Cleaning

In the process of data cleaning, we have identified several instances of duplicated names. These could be variations or slightly different spellings of the same athlete's name, which our code has detected.

```
Matched: SMITH Kiplin | SMITH Kiplin Morrish | 20019
Matched: SMITH Lachlan | SMITH Lachlan Robert | 20023
Matched: SOUD Ahmad | SOUD Ahmad Abu Al | 20131
Matched: SOULIMAN Yazan | SOULIMAN Yazan Al | 20143
Matched: STICKLER Poppy | STICKLER Poppy Grace | 20432
Matched: SUBRI Muhammad | SUBRI Muhammad Syakir Aiman | 20510
Matched: SUE-DOMÍNGUEZ Sebastián | SUE-DOMÍNGUEZ Sebastián Andrés | 20516
Matched: SURTEES Jamie | SURTEES Jamie Leigh | 20715
Matched: TAMSEL Dominic | TAMSEL Dominic Daniel | 20871
Matched: TANG Chia | TANG Chia Hung | 20881
Matched: TAY Wei | TAY Wei An Terry | 21121
Matched: TEILLERS Wout | TEILLERS Wout Johan Alexander | 21131
```

Figure 3: Examples of confused athlete name pairs in the dataset.

Following the initial data collection, we performed cleaning procedures on the dataset. This included removing duplicate records, consolidating slightly different names that referred to the same athlete, and standardizing records that varied in format. Now, we obtained a clean and consistent dataset.

Table 3: A preview of the cleansed dataset.

New_LN_FN	LastName	FirstName	Gender	Country	Date	Competitic	Round	Location	Apparatus	Rank	D_Score	E_Score	Penalty	Score
AAS Fredrik Bjornevik	AAS	Fredrik	m	NOR	2/26/2023	FIG Appara	qual	Cottbus, Germany	HB	39	4.6	6.7	0	11.3
AAS Fredrik Bjornevik	AAS	Fredrik	m	NOR	2/26/2023	FIG Appara	qual	Cottbus, Germany	PH	44	4.4	7.8	0	12.2
AAS Fredrik Bjornevik	AAS	Fredrik Bjo	m	NOR	8/5/2001	2023 FISU	qual	Chengdu, China	FX	54	4	8.566	0	12.566
AAS Fredrik Bjornevik	AAS	Fredrik Bjo	m	NOR	8/5/2001	2023 FISU	qual	Chengdu, China	HB	44	4.6	8.166	0	12.766
AAS Fredrik Bjornevik	AAS	Fredrik Bjo	m	NOR	8/5/2001	2023 FISU	qual	Chengdu, China	PB	54	4.4	8.066	0	12.466
AAS Fredrik Bjornevik	AAS	Fredrik Bjo	m	NOR	8/5/2001	2023 FISU	qual	Chengdu, China	PH	58	4.4	7.266	0	11.666
AAS Fredrik Bjornevik	AAS	Fredrik Bjo	m	NOR	8/5/2001	2023 FISU	qual	Chengdu, China	SR	73	3.5	8.366	0	11.866
AAS Fredrik Bjornevik	AAS	Fredrik Bjo	m	NOR	8/5/2001	2023 FISU	qual	Chengdu, China	VT1	69	4	8.966	0	12.966
ABAD Nestor	ABAD	Nestor	m	ESP		2022 51st I	AAfinal	Liverpool, England	FX	13	5.5	8.3	0	13.8
ABAD Nestor	ABAD	Nestor	m	ESP		2022 51st I	qual	Liverpool, England	FX	48	5.5	7.9	0	13.4
ABAD Nestor	ABAD	Nestor	m	ESP		2022 51st I	AAfinal	Liverpool, England	HB	10	5.3	8.133	0	13.433
ABAD Nestor	ABAD	Nestor	m	ESP		2022 51st I	qual	Liverpool, England	HB	48	5	8.166	0	13.166
ABAD Nestor	ABAD	Nestor	m	ESP		2022 51st I	AAfinal	Liverpool, England	PB	12	5.8	8.4	0	14.2
ABAD Nestor	ABAD	Nestor	m	ESP		2022 51st I	qual	Liverpool, England	PB	26	5.8	8.5	0	14.3
ABAD Nestor	ABAD	Nestor	m	ESP		2022 51st I	AAfinal	Liverpool, England	PH	21	5.1	6.166	0	11.266
ABAD Nestor	ABAD	Nestor	m	ESP		2022 51st I	qual	Liverpool, England	PH	55	5.2	7.566	0	12.766
ABAD Nestor	ABAD	Nestor	m	ESP		2022 51st I	AAfinal	Liverpool, England	SR	12	5.3	7.933	0	13.233
ABAD Nestor	ABAD	Nestor	m	ESP		2022 51st I	qual	Liverpool, England	SR	90	4.9	7.666	0	12.566
ABAD Nestor	ABAD	Nestor	m	ESP		2022 51st I	AAfinal	Liverpool, England	VT	20	4.8	9.033	0.1	13.733
ABAD Nestor	ABAD	Nestor	m	ESP		2022 51st I	qual	Liverpool, England	VT1	97	4.8	9	0.3	13.5
ABAD Nestor	ABAD	Nestor	m	ESP	8/21/2018	2022 Senic	qual	Munich, Germany	FX	86	5.4	7.133	0.1	12.433
ABAD Nestor	ABAD	Nestor	m	ESP	8/21/2018	2022 Senic	qual	Munich, Germany	HB	10	5.3	8.5	0	13.8
ABAD Nestor	ABAD	Nestor	m	ESP	8/21/2018	2022 Senic	TeamFinal	Munich, Germany	HB	10	5.3	8.1	0	13.4

2.3. Date Construction

To delve deeper into our dataset, we have focused our analysis on a 'per athlete per apparatus' basis. This approach involves aggregating the original data to this specific level. Consequently, we now have the average and variance for Difficulty (D) Score, Execution (E) Score, and Penalty Score, along with one-hot encoding for each apparatus. With a clear grain, these refined data format facilitates its application in various algorithms.

Table 4: Per-Athlete-Apparatus Dataset (A sample)

Name	Country	Matches Participated	Average D_Score	D_Score Variance	Average E_Score	E_Score Variance	Average Penalty	Penalty Variance	Mean	Std	Apparatus_BB	Apparatus_FX	Apparatus_UE	Apparatus_VT
AKHAIMOVA LILIIA	ROC	5	5.72	0.0096	8.9794	0.00471824	0.02	0.0016	14.6994	0.055707	0	0	0	1
ANDRADE REBECA	BRA	5	5.86	0.0144	7.553	0.1979956	0	0	13.413	0.358715	1	0	0	0
ANDRADE REBECA	BRA	6	5.933333333	0.018888889	8.166333333	0.029614889	0.083333333	0.021388889	14.015833	0.253838	0	1	0	0
ANDRADE REBECA	BRA	5	6.06	0.0424	7.9664	0.29173344	0	0	14.0264	0.693344	0	0	1	0
ANDRADE REBECA	BRA	9	5.488888889	0.818765432	9.284888889	0.09366321	0.033333333	0.002222222	14.751556	1.175826	0	0	0	1
ASHIKAWA URARA	JPN	4	5.775	0.026875	7.933	0.0538445	0	0	13.70725	0.352186	1	0	0	0
ASHIKAWA URARA	JPN	2	4.2	0.09	7.2495	0.83997225	0	0	11.4495	1.2165	0	1	0	0
BILES SIMONE	USA	4	6.375	0.026875	7.9665	0.07971675	0	0	14.3415	0.340173	1	0	0	0
BILES SIMONE	USA	3	6.766666667	0.002222222	8.144333333	0.126264222	0.133333333	0.015555556	14.777667	0.517493	0	1	0	0
BILES SIMONE	USA	3	6.066666667	0.008888889	8.255333333	0.021123556	0	0	14.322	0.211587	0	0	1	0
BILES SIMONE	USA	5	5.96	0.0704	9.4364	0.05035744	0.28	0.0256	15.1764	0.322252	0	0	0	1
BLACK ELSABETH	CAN	5	5.86	0.1144	7.7862	0.02336096	0	0	13.645	0.29936	1	0	0	0
BLACK ELSABETH	CAN	3	5.1	0.02	7.855	0.395908667	0	0	12.955	0.487965	0	1	0	0
BLACK ELSABETH	CAN	3	5.7	0	7.855333333	0.285264222	0	0	13.555333	0.534102	0	0	1	0
BLACK ELSABETH	CAN	7	5.085714286	0.055510204	9.047285714	0.018707061	0	0	14.133	0.310756	0	0	0	1
BOYER MARINE	FRA	10	5.58	0.0196	7.7763	0.19478301	0	0	13.3563	0.45466	1	0	0	0
BOYER MARINE	FRA	9	4.933333333	0.015555556	7.929333333	0.017636	0.055555556	0.009135802	12.807111	0.25382	0	1	0	0
BOYER MARINE	FRA	3	4.633333333	0.055555556	7.266666667	0.695555556	0	0	11.9	1.067708	0	0	1	0
BOYER MARINE	FRA	3	4.333333333	0.035555556	9.088666667	0.003930889	0	0	13.422	0.226516	0	0	0	1
BRASSART MAELLYSE	BEL	7	5.242857143	0.04244898	7.266428571	0.622650816	0	0	12.509286	0.848195	1	0	0	0
BRASSART MAELLYSE	BEL	8	4.8625	0.01734375	7.697625	0.153676484	0.0125	0.00109375	12.547625	0.430308	0	1	0	0
BRASSART MAELLYSE	BEL	8	5.3875	0.06359375	7.851625	0.144833734	0	0	13.239125	0.401708	0	0	1	0
BRASSART MAELLYSE	BEL	9	3.866666667	2.008888889	7.955222222	7.955943284	0.022222222	0.001728395	11.799667	4.177955	0	0	0	1

3. Process II: Distribution Estimation

3.1. KDE Analysis Using Self-Historical Data

Kernel Density Estimation (KDE) is a statistical technique used to estimate the probability density function of a random variable. It's particularly useful when dealing with limited or discrete datasets.

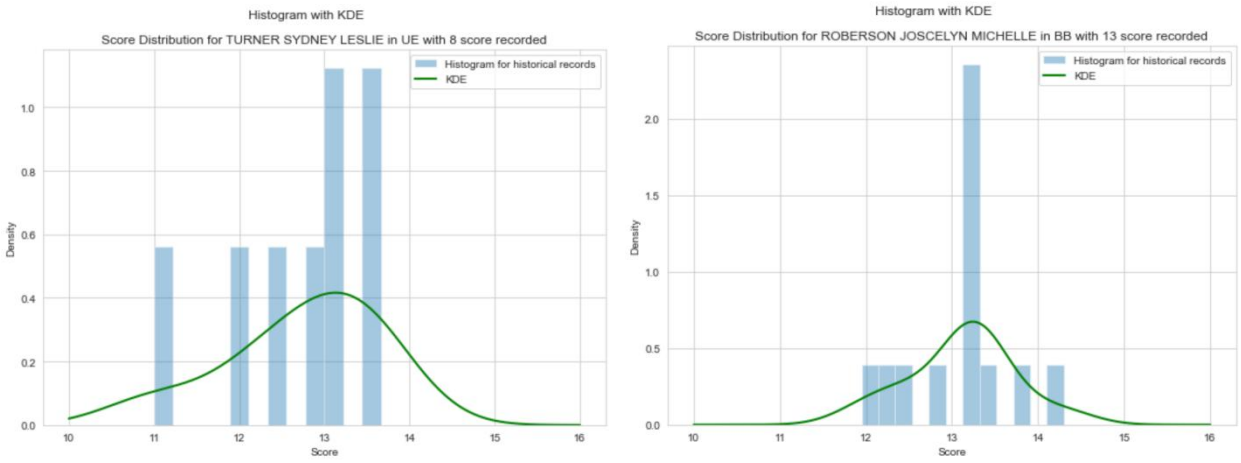


Figure 4: Two examples of KDE's fitness overview.

When implementing Kernel Density Estimation (KDE) in our specific context, the process is straightforward and requires only the historical scores of a particular athlete in a specific apparatus. This method intuitively illustrating the athlete's score distribution on that apparatus.

Challenges arise with singular or identical scores, as they can skew KDE fitting. To mitigate this, we manually augment data by adding and subtracting 0.25 from a single score.

3.2. Gaussian Distribution Analysis Using Self-Historical Data

The Gaussian or normal distribution, with its symmetrical bell curve, is defined by two key parameters: the mean and the variance.

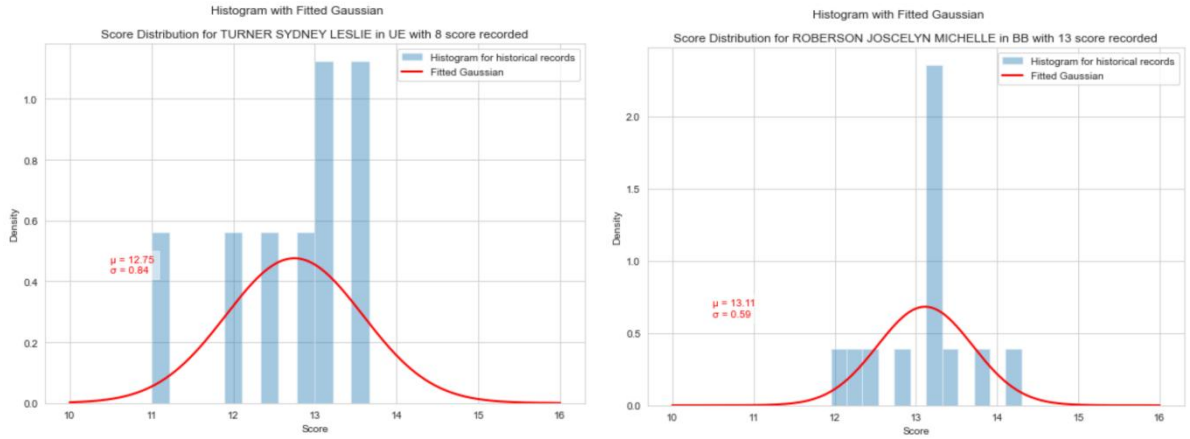


Figure 5: Two examples of Gaussian Distribution's fitness overview.

Direct Gaussian distribution estimation can only be calculated for athletes with at least two scores per apparatus, adhering to the statistical requirement for multiple data points to estimate Gaussian parameters reliably. A single score would not adequately capture variance, essential for evaluating data spread and consistency. Thus, for athletes with a single record, who therefore lack a Gaussian Distribution estimate, we choose to predict their mean and standard deviation rather than using the original data outright. For fairness, and to maintain consistency, we also apply these predicted values to other athletes' data.

Use Per-Athlete-Apparatus Dataset, with the mean and standard deviation as dependent variables, and the individual performance features of each athlete on each apparatus as independent variables, we employed three regression techniques: linear regression, random forest, and XGBoost. The model demonstrating the highest accuracy was selected as our optimal approach.

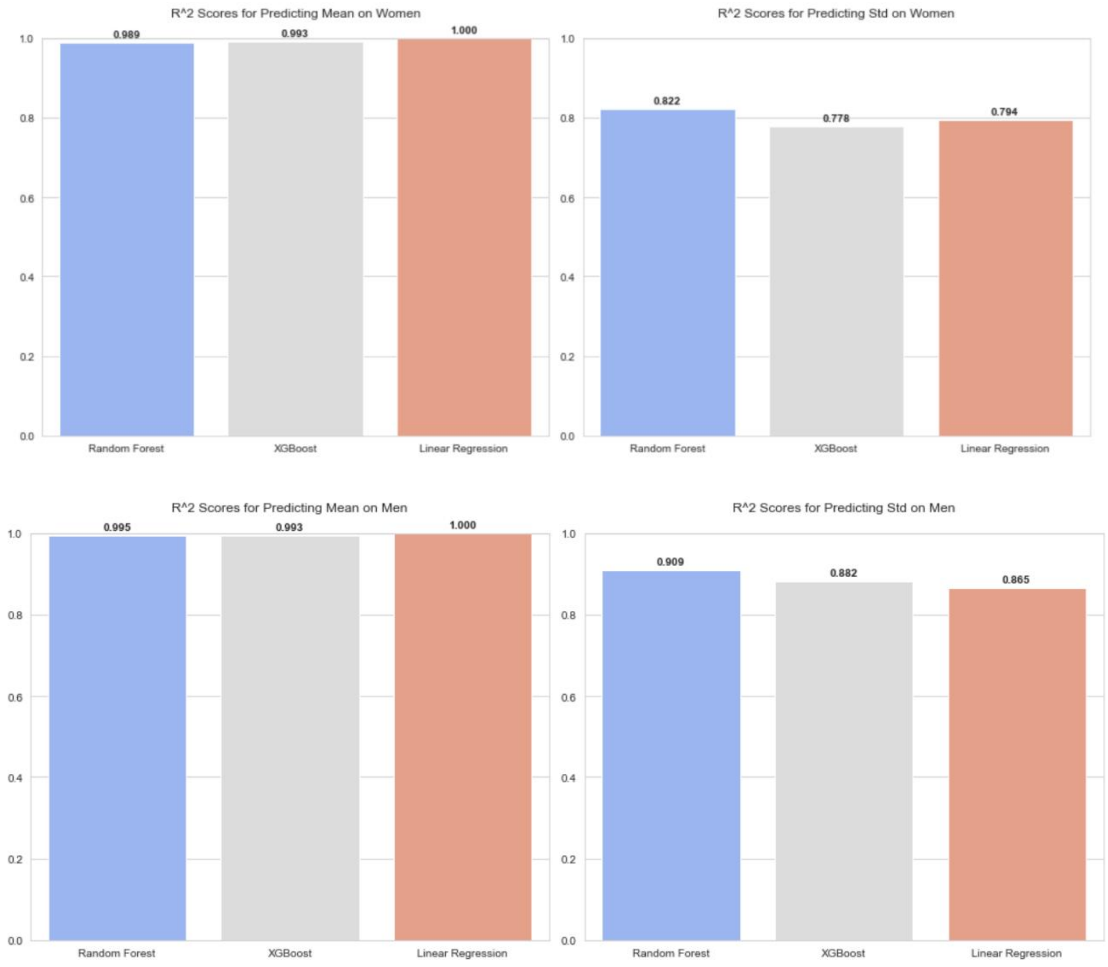


Figure 6: Accuracy of three models in predicting mean and standard deviation for Women and Men data.

We created gender-specific models to predict mean and variance of scores using athletes' features. Applying these models to all athletes in each apparatus, we established an expected Gaussian distribution for every athlete per apparatus.

3.3. KDE Analysis Utilizing Nearest Neighbor Historical Data

Given the small sample of historical results per athlete, which accurately indicates performance but lacks breadth, we've adopted a K-Nearest Neighbors (KNN)-inspired approach. .

In our implementation, we apply the KNN algorithm within each apparatus, using the variables from the Gaussian distribution model as the basis for comparison. We also normalize the data to address the issue of varying magnitudes.

Upon Per-Athlete-Apparatus Dataset, with KNN applied, we identify the three athletes with characteristics most similar to each target athlete in the dataset.

In predicting an athlete's score for an apparatus, we consider not only their own historical scores but also the score distributions of three similar athletes. This approach broadens our data usage, counteracting biases from limited historical data.

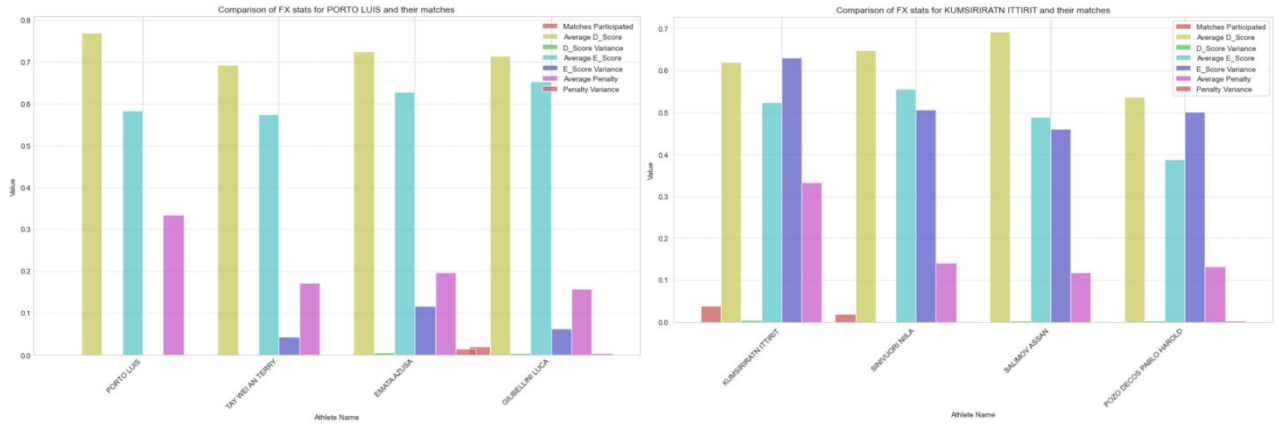


Figure 7: Two examples of the pairing effect in KNN algorithm applications.

3.4. Integration of Three Methods for Score Estimation

Moving forward, when generating a score for an athlete in a specific apparatus, our approach involves resampling using each method and subsequently averaging these scores to determine the final outcome.

By employing a multifaceted approach, we ensure that our predictions are not disproportionately influenced by any particular model or data set. This is particularly important in the context of athletes with limited historical data, as it allows for a more balanced and comprehensive assessment of their potential performance.

Furthermore, since this approach is uniformly applied to every athlete, it ensures equitable treatment across the board. This uniformity in our methodology underpins the reasonableness of comparing the generated scores between athletes. Each athlete's score is the result of a similarly rigorous and multifaceted analytical process, thereby providing a fair basis for comparison.

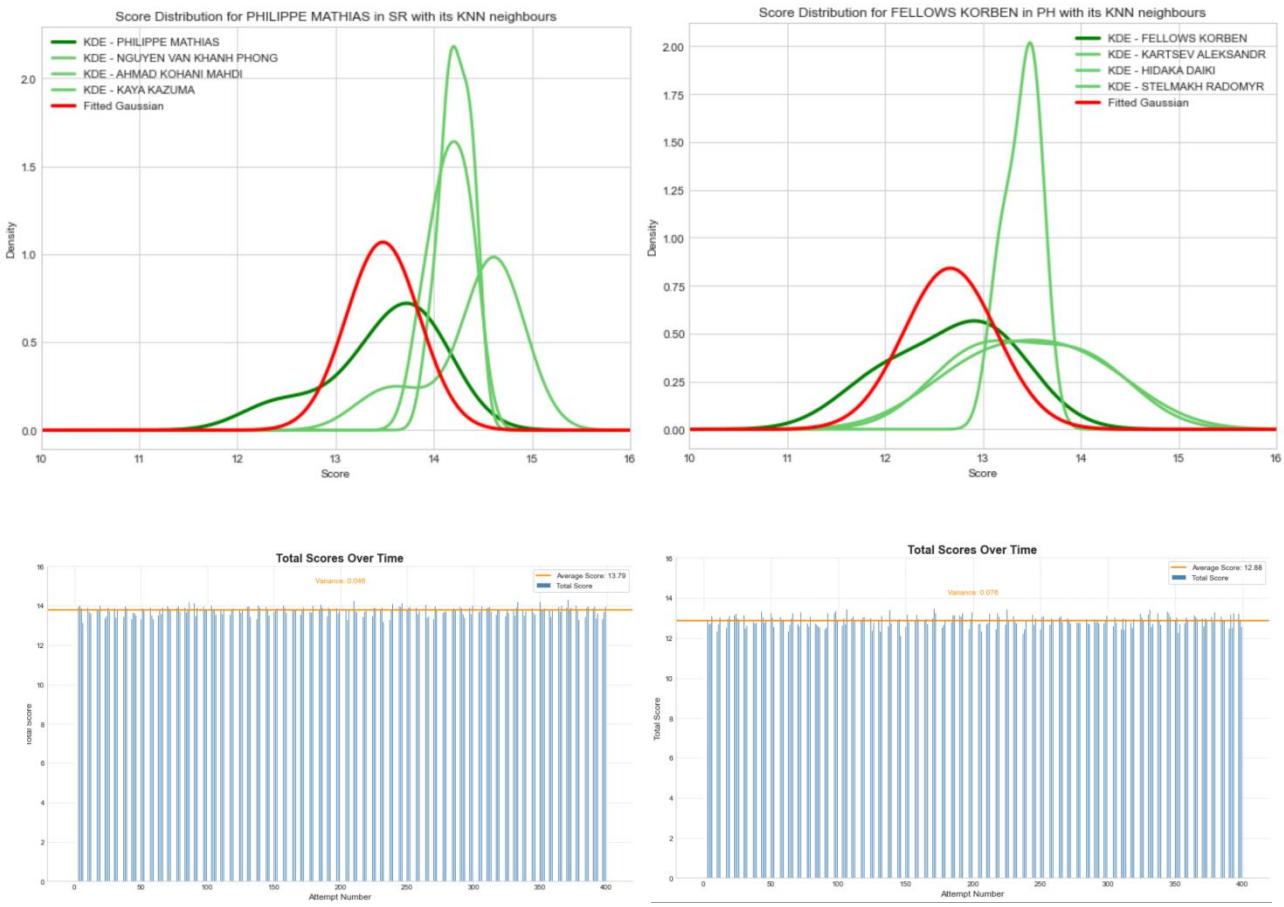


Figure 8: Two examples of estimated distributions for an athlete on an apparatus from three models, and the 400 scores sample generated by them

4. Process III: Best Team Selection

4.1. Initial Filtering

Leveraging the estimation method outlined earlier, we've been able to quantify the performance scores for each athlete in their respective apparatuses, as illustrated in the accompanying data. This evaluation reveals significant disparities in athlete proficiency. In the interest of efficiency, it would be advisable to conduct an initial filtering stage prior to simulating the full scope of the Olympic competition.

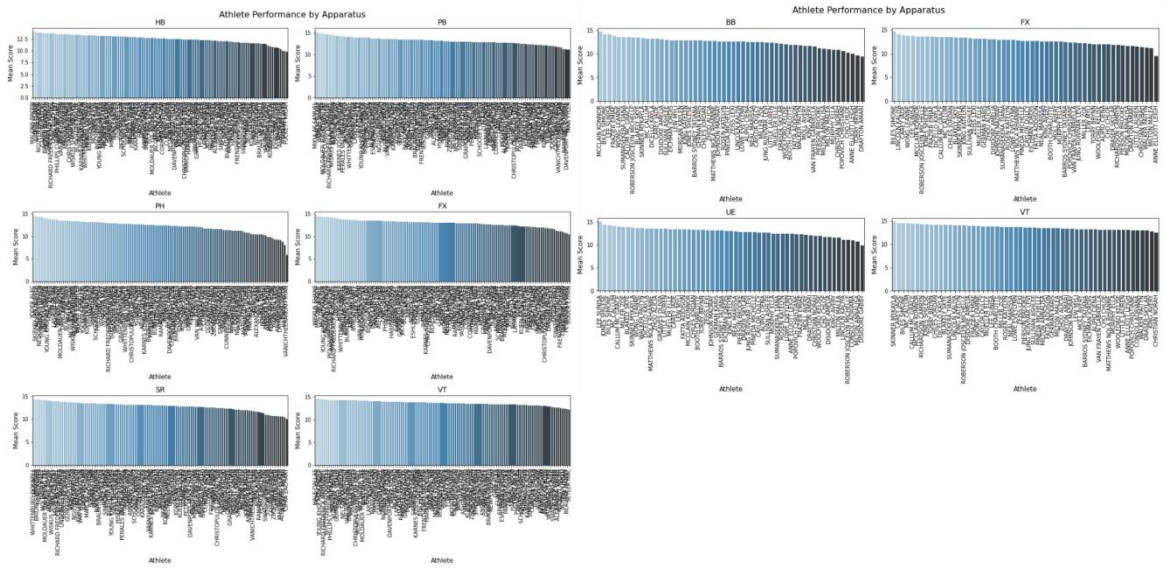
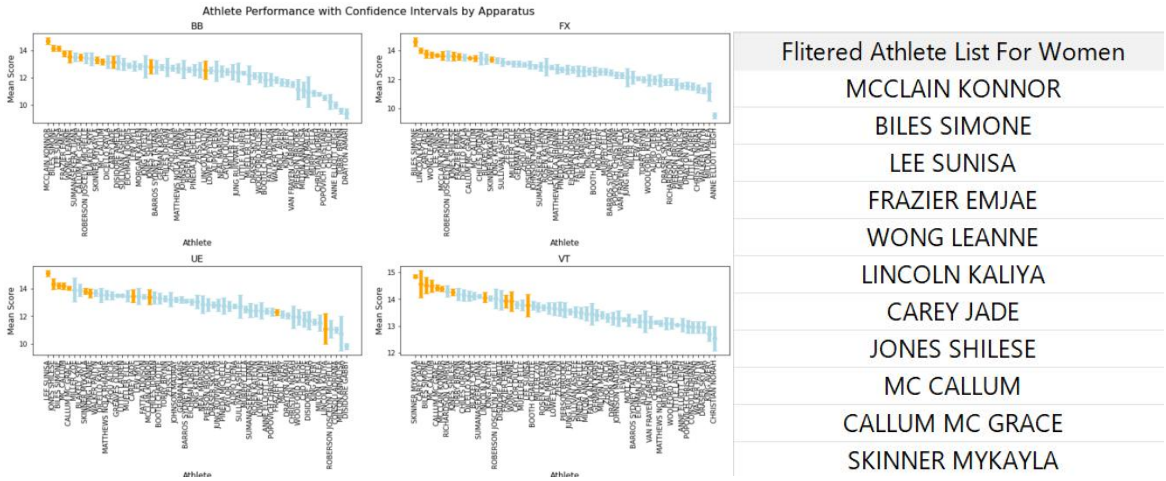


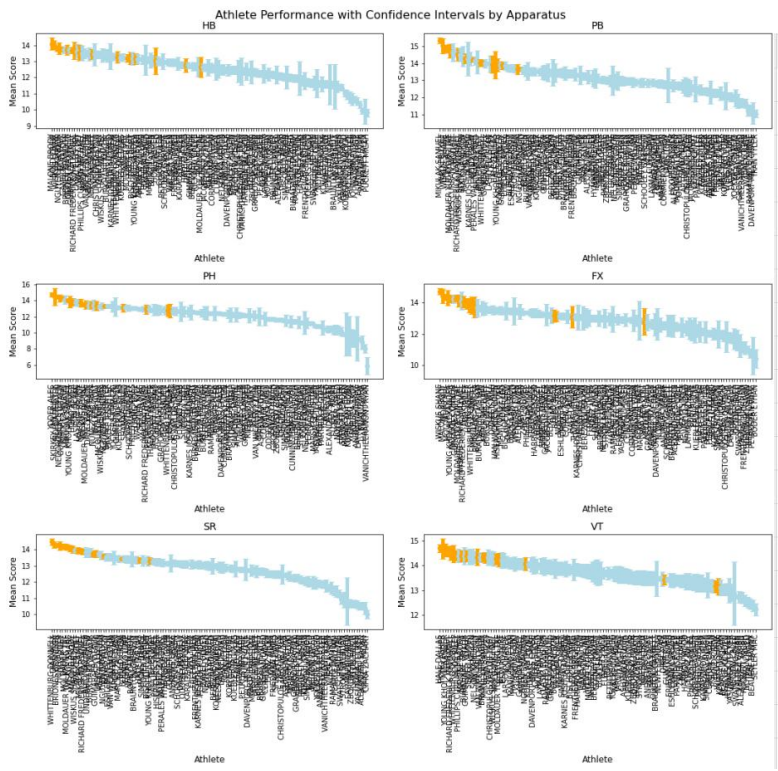
Figure 9: Estimated average score of each athlete in every apparatus.

We selected the top five athletes in each apparatus, creating 11 potential candidates for our Olympic women's team and 21 candidates for the men's team. Notably, athletes excelling in one apparatus often performed well in others, which benefits team construction.



Flitered Athlete List For Women
MCCLAIN KONNOR
BILES SIMONE
LEE SUNISA
FRAZIER EMJAE
WONG LEANNE
LINCOLN KALIYA
CAREY JADE
JONES SHILESE
MC CALLUM
CALLUM MC GRACE
SKINNER MYKAYLA

Figure 10: Estimated score range and selected athlete list for men.



Flitered Athletes List For Men
YODER ALEC
SKIRKEY IAN HUNTER
MOLDAUER YUL
NEDOROSCIK STEPHEN
GUIMARAES VITALIY
WISKUS SHANE
MALONE BRODY
JUDA PAUL
YOUNG KHOI ALEXANDER
BOLD CREW
NGUYEN BRANDON
MIKULAK SAMUEL
WALKER COLT
HONG ASHER
MOLDAUER YUL KYUNG TAE
WHITTENBURG DONNELL
DIAB ALEX
BRIONES BRANDON
LOOS RILEY
HALE DALLAS
RICHARD FREDERICK NATHANIEL

Figure 11: Estimated score range and selected athlete list for women.

4.2. Genetic Algorithm

The core of our team selection strategy for the Paris Olympics is a comprehensive simulation of the entire event. By closely examining the progression of the matches, it is not hard to develop a code that simulates all events in the Olympics, thus determining the outcomes for each competition.

However, a major unresolved issue is determining the teams from other countries and their participating athletes. Due to limited information about the upcoming Paris Olympics, we rely on data from the last Olympics in Tokyo, assuming a relatively stable competitor landscape.

Acknowledging rule changes from Tokyo to Paris Olympics, we tailored our strategy for teams over five athletes, choosing the best lineup for all events to maximize scores. For countries with less than five athletes, we considered adding members for optimal team composition. This led to fixed, optimized five-member teams from 11 countries. Hence, the whole simulation can be constructed.

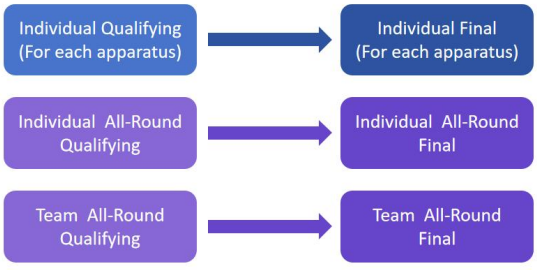


Figure 12: Matches to simulate for the Paris Olympics.

Given the numerous possibilities for selecting a five-member team from these candidates, we developed a genetic algorithm to identify the optimal team composition efficiently and in a traceable manner. The construction of our genetic algorithm involves the following key components:

- *Initialize_population.* Initially, we repeatedly invoke 'GetUSSample' function to generate 12 distinct initial teams from the pool of American athletes, pre-screened through the 'InitialFilter'.

- *Selection.* For selection, we currently employ total_score as the sole metric for evaluating our genetic algorithm. The score is calculated using the formula:

$$Team\_Score = Gold\ Medals * 3 + Sliver\ Medals * 2 + Bronze\ Medals * 1$$

  Prior to reproduction, we eliminate the bottom 50% of teams, allowing the top 50% to crossbreed. These teams are also subject to a 20% mutation rate, where random replacements are drawn from our pool of candidate athletes. Additionally, to ensure consistent generational improvement, we retain the four most outstanding teams from each generation within our genetic algorithm. To guarantee score stability, every team undergoes 1024 simulation matches, providing reliable and stable mean and variance data.

- *Crossover.* The crossover strategy involves concatenating the athletes from both parent teams into a list of 10 names and randomly drawing five times. Each draw removes the selected athlete from the list, including any duplicates.

- *Mutate.* Post-crossover, random mutations are applied in team selection. Teams are alphabetically sorted by athlete names and compared with next-generation teams. A team is added only if unique; if not, it undergoes further mutation.

Based on the plot provided, it is evident that after 16 generations of iteration, the average score of each generation tends to stabilize. Additionally, the composition of the teams highlights the significance of certain athletes, as evidenced by their frequent selection.

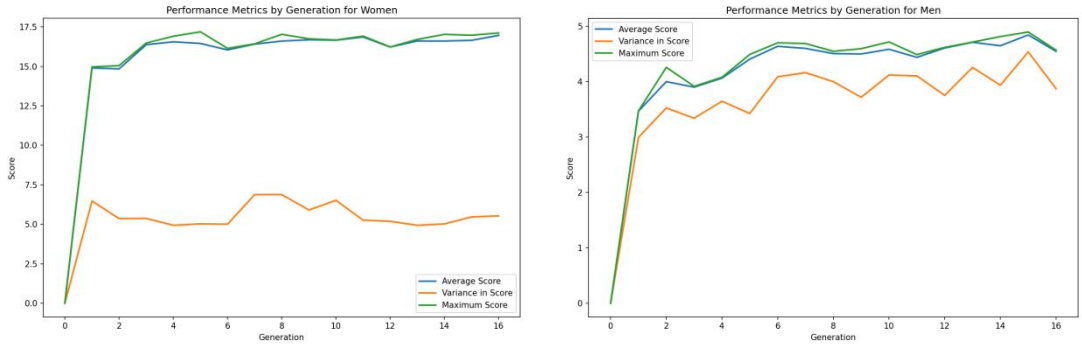


Figure 13: Genetic algorithm performance by generation for men and women.

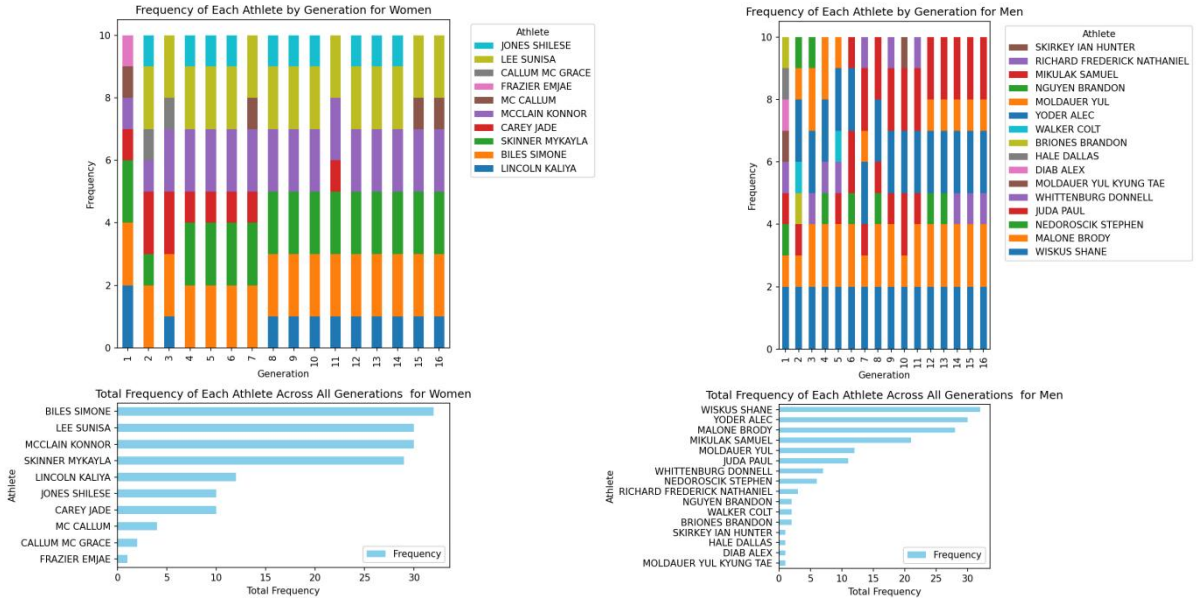


Figure 14: Athlete selection frequency by generation in genetic algorithm for men and women.

4.3. Final Comparison

After some separate iterations of the genetic algorithm, with 16 generations each, the top three teams from the best two iterations were selected.

These teams underwent 2048 simulations to assess their performance, demonstrating exceptional results. Considering these teams all have a good performance on score, for the final selection, both average score and variance were considered. Each team's 95% lower score bound was calculated based on their average score, variance, and number of simulations. Teams were then ranked by this metric, leading to the selection of the best-performing team based on their lower score bound.

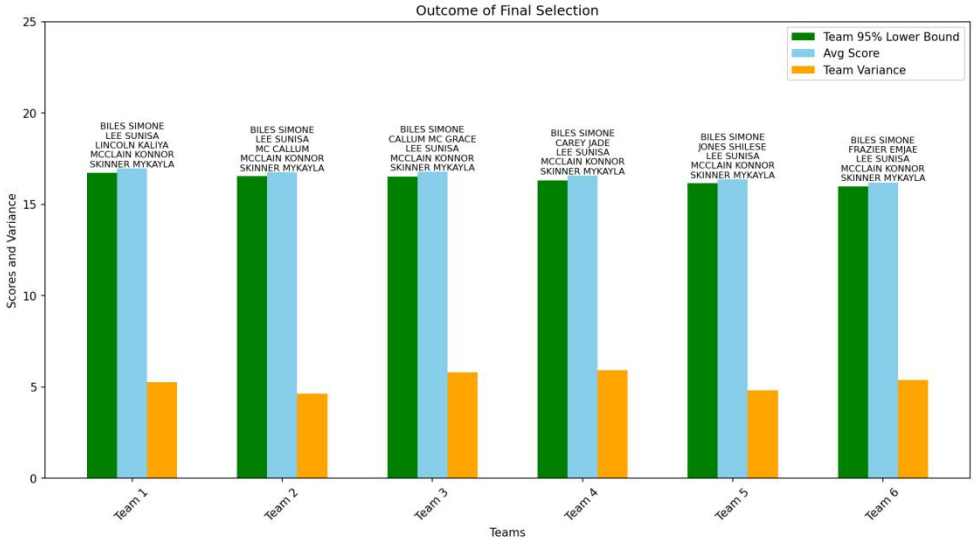


Figure 15: Final comparison with ranking of the top 6 candidate teams for women.

In the women's team selection, upon analyzing mean and variance, Team 1 stands out with an average score of 16.95 and a variance of 5.25, securing its position as our final top choice due to its outstanding performance. Team 2 follows as the second choice with a competitive average score of 16.75 and a lower variance of 4.65, indicating a consistent and stable performance. Team 3, chosen as the third option, has a slightly higher average score of 16.77 and a variance of 5.81, offering performance levels comparable to Team 2 but with greater score variability, while still upholding a high performance quality.

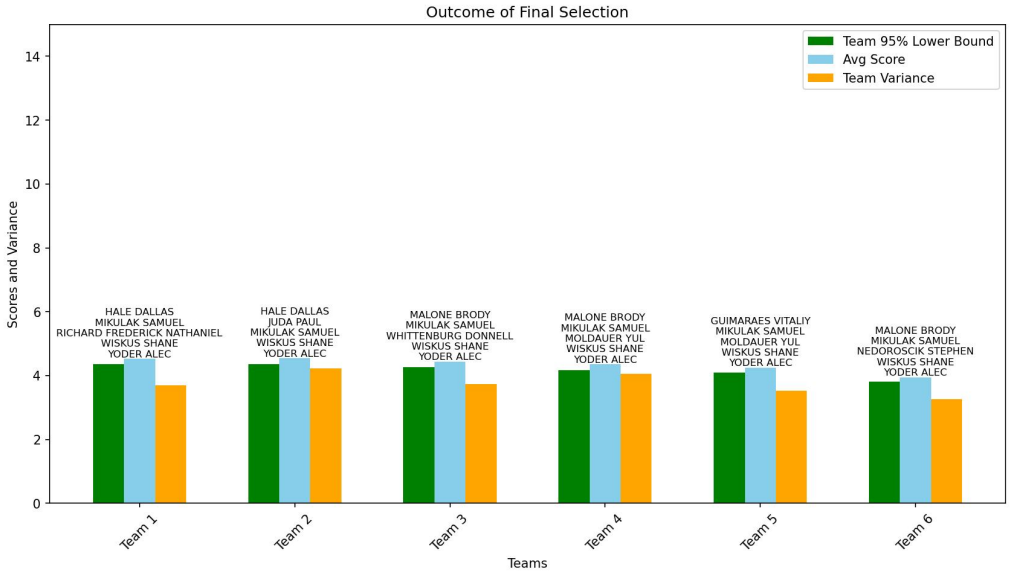


Figure 16: Final comparison with ranking of the top 6 candidate teams for men.

In the men's team selection, despite overall scores being modest in comparison to the women's, Team 1, with an average of 4.52 and a variance of 3.7, has been identified as our first choice due to its strong and relatively stable performance. Team 2, our second choice, presents a closely competing average score of 4.54 with a higher variance of 4.23, suggesting a comparable level of performance with a little more variability. Team 3, chosen as the third option, offers an average score of 4.43 and a variance of 3.73, showing a level of consistency and reliability akin to that of Team 1.

5. Feasibility Exploration

For our model to be accurate and feasible, we base it on three key assumptions:

- An athlete's performance can be partly predicted from their historical record and overall follows some distribution.
- Over a short period (2-3 years), an athlete's overall performance is relatively stable.
- The overall competitor level in the Olympics is relatively stable, regardless of the participating countries and athletes.

Our model for the 2024 Olympics is based on three fundamental assumptions. Firstly, the data used in our model spans the most recent two years. Despite variations in team compositions and participating countries from the Tokyo to the Paris Olympics, our second assumption enables us to simulate the Paris Olympics effectively, with expectations of medal outcomes mirroring actual counts.

Furthermore, we focus on the relative rankings of athletes rather than absolute accuracy in their scores. This approach ensures fairness and provides reliable rankings, despite potential inaccuracies in individual performance estimations.

In conclusion, our strategy employs a genetic algorithm for initial broad team selection, followed by identifying the best team based on the lower bound of their confidence interval. This method is designed to ensure that our selected team is likely to perform well statistically, thus providing a stable and feasible foundation for our model.

Hence, we establish the feasibility of our model.

6. Code Efficiency Optimization

Efficiency is crucial in the genetic algorithm for the Paris Olympics, as it involves hundreds of thousands of simulations to reliably score teams. To enhance the efficiency of genetic algorithm, we focused on two key areas of optimization:

- *Data Handling and Pre-Computation.* We optimized data operations by initially setting up structured dataframes, instead of expanding them incrementally. Operations were aggregated on lists, merging into dataframes only at the end. This reduced computational load. A 'Score_Stack' was also created to pre-calculate athletes' scores, minimizing the need for repeated distribution model calls during simulations.

- *Parallel Processing.* We utilized parallel processing to speed up execution, distributing simulations across multiple processes (4-16), depending on the total number of matches. This balanced the workload per process, greatly accelerating the overall simulation.

7. Summary

The whole project follows a structured approach, starting with data preparation, including data supplement and cleaning, and then distribution estimation using KDE and Gaussian models based on historical data. Simulation progress with a genetic algorithm is then applied for team selection, considering performance stability, overall Olympic competitor levels and the match schedule of the Paris Olympics. This comprehensive process leads to the final selection of the most effective teams for the Paris 2024 Olympics, demonstrating the model's practical application and feasibility.