Grenze International Journal of Engineering and Technology, June Issue

Track Mitra: A Smart Heats Allocation System for Equitable Sprint Competitions

Akshit K. Mahaur[1], Ruhi Patankar[2], Sarika Bobde[3], Era Aggarwal[4], Tushar Mittal[5], Ishan A. Oze[6] and Yash Lad[7]
[1-7]MIT World Peace University, Pune, India
Email: 1032201253@mitwpu.edu.in, {sarika.bobde, ruhi.patankar, 1032200936, 1032200956, ishan.oze,1032201231}@mitwpu.edu.in

***Abstract*— Track and field is a discipline that involves intense competition and hinges predominantly on statistical measures to evaluate performance and rank athletes. Despite the widespread application of rudimentary performance metrics to assess athletes, there has been scant utilization of advanced analytical techniques in the discipline. The current study endeavors to conceive and implement a suite of statistical and machine learning models that address key issues prevalent in track and field, especially concerning outdoor running events. The primary objective is to gain comprehensive insights into the discipline from an advanced analytics standpoint and augment ranking systems and race strategies. To ensure practicality and accessibility of the models for use at track meets, only readily available results information, such as athletes' prior result times, is used throughout the analyses. The model aims to mathematically cluster the athletes to assist a coach in determining which individuals to choose from a pool of athletes to comprise the fastest team and order, which will extend a guiding hand to the coach in the intricate task of selecting the crème de la crème from a pool of athletes, assembling the swiftest team, and meticulously arranging their sequence. In conclusion, this research work presents innovative implementations of advanced analytics in the field of track and field, offering valuable insights into the potential uses of statistical and machine learning models in enhancing performance and decision-making in the sport. The models developed herein have the potential to improve the precision and impartiality of ranking systems, optimize race strategies, and aid coaches in making well-informed choices regarding the selection and order of team members.**

***Index Terms*— Track and Field, Unsupervised Learning, Data Analysis**

I. INTRODUCTION

A. The Role of Analytics in Sports

Analytics plays a pivotal role in the realm of sports, revolutionizing how teams, athletes, coaches, and organizations approach their strategies and decision-making processes. Through the utilization of data analysis and machine learning, sports analytics provides valuable insights and actionable information that can have a profound impact on performance, training, and overall outcomes in various sports disciplines.

The field of sports analytics involves the collection, processing, and analysis of extensive data sets, encompassing player statistics, match results, game footage, and sensor-generated data. By employing advanced statistical techniques and predictive modeling, analytics professionals can unveil patterns, trends, and correlations within the data, enabling a deeper understanding of player performance, team dynamics, and game strategies.

The application of analytics in sports encompasses diverse domains. Performance analytics focuses on quantifying

Grenze ID: 01.GIJET.10.2.72

and assessing individual and team performance, identifying strengths, weaknesses, and areas for improvement. This entails analyzing player statistics, tracking player movements, and evaluating physical and physiological data to optimize training programs, prevent injuries, and facilitate rehabilitation.
Game strategy and decision-making analytics aid teams and coaches in making informed choices based on data-driven insights. This involves analyzing opponents' performance, studying game scenarios, and developing strategies to maximize the probability of success. Analytics also contribute to evaluating the influence of variables such as weather conditions, player substitutions, or tactical adjustments on game outcomes.

B. Track & Field – An overview

Track and field, a sport with an enduring legacy, continues to draw participants even in modern times. The core focus of these contests revolved around three fundamental categories: running, jumping, and throwing. Athletes from each of these disciplines convene in a track and field meet, where a range of specific events are contested. This standardized approach lends track and field its reputation as a numerically oriented sport, where performance is intrinsically linked to quantifiable outcomes.

C. Model

The rationale behind using the k-means clustering approach for this study is its ease of use, effectiveness, and interpretability. Well-known and simple to use, K-means is a clustering technique that works well for dividing data into different groups according to similarity. The objective of track and field analytics is to group athletes according to their racing patterns and performance traits, which makes k-means a suitable option.
The computational efficiency of k-means is one of its main benefits, which makes it a good choice for handling huge datasets that are frequently found in track and field results. Because of the algorithm's scalability and iterative nature, track meets can effectively use it without incurring significant computational costs. Additionally, as the results are simple to understand and apply in the context of team selection and race strategy, the interpretability of k-means clusters is in line with the objective of giving coaches and organizers practical insights.
The practical justifications for selecting k-means include its simplicity in terms of implementation, its low requirement for parameter tuning, and its capacity to manage the high-dimensionality of track and field performance data. K-means offers a workable method that supports the objective of making sophisticated analytics accessible and relevant in a broad, real-world track and field scenario, given the emphasis on leveraging easily accessible results data.

D. Model

This study's experimental approach focused on analyzing publicly available track and field event results data using k-means clustering methods, with an emphasis on outdoor running competitions. Improving ranking systems, optimizing race strategy, and supporting coaches in team selection were the main goals. The experimental design and its conformity to the specified objectives are delineated in the subsequent steps:

E. Data Collection

Collected and made publicly available the results data from multiple track and field meets, guaranteeing a varied dataset covering a variety of outdoor running competitions.

F. Data Preprocessing

Standardized and cleaned the data from the results to guarantee consistency and accuracy in the analysis. retrieved pertinent features to feed the k-means model, such as athletes' historical result times and performance indicators.

G. Model Selection

The k-means clustering algorithm was selected because it satisfies the actual needs of track and field analytics and is easy to understand, efficient, and straightforward.

H. Model Training

Used the supplied dataset to run the k-means algorithm, optimizing cluster assignments iteratively based on performance similarities. investigated a range of cluster numbers (k values) in order to determine the ideal setup for capturing significant patterns in athlete performances.

I. Cluster Analysis

Analyzed the ensuing clusters to learn more about the various athlete performance categories. Studied each cluster's features, finding similarities and contrasts in racing tactics and past results.

J. Performance Evaluation

Assessed how well the clustering model performed in accomplishing the specified goals, taking into account advances in ranking accuracy, race strategy optimization, and coach-useful application.

K. Validation and Iteration

Verified the outcomes using cross-validation or by contrasting them with recognized track and field benchmarks. ensured practical relevance by iteratively improving the model and experimental design in response to comments and new information.

L. Documentation and Reporting

Recorded every stage of the experimental procedure, including the model configuration, evaluation metrics, and data preprocessing procedures. findings were presented in a thorough report with a focus on how the k-means clustering models might improve track and field decision-making.

Through adherence to this experimental design, the study effectively employed k-means clustering to accomplish the specified goals, offering insightful information about athlete performance clusters that can guide ranking schemes, competition tactics, and coach decision-making in outdoor running competitions.

II. Related Work

Our review of existing literature did not yield research papers that directly corresponded to the specific parameters of our use case. Nevertheless, we identified publications that demonstrated notable proximity to the objectives of our study.

In the study conducted by Lin, Dong, and Li (2022), the authors employed a methodology to analyze big data behavior in sports track and field through machine learning models. Their approach involved collecting extensive datasets of relevant sports data, prioritizing key features, selecting suitable machine learning models, and evaluating the models to identify potential benefits. The primary goal was to leverage machine learning techniques for a comprehensive analysis of big data in sports, aiming to extract meaningful insights from the collected data.

The findings of their research revealed intriguing patterns. Specifically, the increase in the number of hidden neurons was correlated with a decrease in the NRMSE (Normalized Root Mean Squared Error) value. Conversely, an increase in the dropout rate led to an elevation in the NRMSE value. These results provided valuable guidance for selecting optimal parameters when employing machine learning models to analyze sports track and field data.

Despite these contributions, the study acknowledged certain drawbacks, including the limited availability of data and technical complexities inherent in handling extensive datasets. The identified research gap emphasized the importance of understanding the multifaceted factors influencing the behavior and performance of athletes in sports track and field.

In comparison to this foundational work, our research aimed to delve into a more nuanced aspect of the athlete's performance by scrutinizing the critical points of contention in the systems. While their work primarily focused on machine learning model performance, our study sought to narrow down and provide a deeper understanding of the factors contributing to athlete behavior and performance. This nuanced difference was pivotal, offering a unique perspective and contributing to the broader discourse on sports analytics and performance evaluation.

III. Motivation

In the realm of track and field, the focus on advanced analytics to supplement basic statistics has been relatively limited. Although automated timing systems efficiently record athletes' final times and lap splits, meet organizers often forgo the additional expense and effort required to employ specialized cameras for continuous race analysis. Consequently, the primary dataset collected and analyzed in the majority of track and field meets consists of the automatic results generated by each athlete upon race completion. Despite the abundance of results data available to track and field enthusiasts, its practical application in average meets has been primarily limited to basic statistics, with little exploration of its broader potential.

In relation to the sport's scale and popularity, there is a noticeable scarcity of research papers that integrate mathematical modeling into the field of track and field. However, access to personalized training data is generally restricted from the public domain due to privacy concerns, thereby impeding the implementation of such models in regular competitions.

In the domain of track and field research, a good majority of papers that rely solely on track and field results as their dataset tend to concentrate on the prediction of athletes' overall personal records or season records, rather than delving into the analysis of individual races. A personal record signifies an athlete's fastest time achieved in

a specific event, whereas a season record represents the athlete's best performance within a particular season of competition for a given event.

The existing body of literature in the realm of research surrounding track and field as a domain, demonstrates that the abundance of track and field results data holds immense potential for uncovering patterns and facilitating comprehensive comparisons among athletes in various stages: pre-race, during the race, and postrace. While previous studies leveraging advanced analytics in track and field have relied on data that is not accessible to the general public, such as proprietary training data or specialized cameras and sensors, this paper exclusively employs publicly available data. By utilizing commonly generated results data from nearly all track and field meets, the models proposed in this study can be readily applied by individuals without the need for sophisticated equipment or extensive background information on each athlete. Consequently, these models can be pragmatically implemented on a wide scale within the sport without requiring any additional meet-specific resources. The demand for models with practical applications serves as the driving force behind the development of various statistical and machine learning models to facilitate athlete comparisons in the realm of track and field. Given the numerous opportunities for comparison in running events within track and field, this paper will primarily focus on predictions and analysis pertaining to such events.

IV. Problem Domain

The domain at the heart of this problem lay in the convergence of sports analytics and data-driven decision-making within track and field. Specifically, the challenge resided in the manual processes of grouping athletes for sprint races, where traditional methods encountered inefficiencies and subjectivity, impeding the establishment of fair and competitive race heats. This domain explored the imperative for innovative solutions, incorporating advanced analytics and unsupervised learning techniques to automate and enhance the grouping of athletes. The focus was on revolutionizing the analysis of sports data, ensuring efficient and nuanced clustering reflective of complex relationships within the dataset. In this context, the pursuit of fairness, efficiency, and objectivity in sprint race competitions propelled the exploration of state-of-the-art solutions for automated heat formation.

V. Problem Definition

Even though track and field relies heavily on basic performance measures, advanced analytics are not widely integrated in the sport. Predictive analytics is especially underutilized in outdoor running events, because meet organizers mostly rely on automatic results for basic statistics. Due to privacy considerations that restrict access to individualized training data, there is a dearth of research incorporating mathematical modeling into track and field. Previous research has tended to focus on forecasting athletes' season or personal records instead of examining race-by-race data. The wealth of publicly accessible track and field results data represents an unrealized possibility for thorough comparisons between athletes at different phases of the event. To bridge the gap between readily available data and advanced analytics applications in the track and field domain, the task at hand is to develop and implement statistical and machine learning models exclusively using readily available results data to cluster athletes. These insights can improve ranking systems, optimize race strategies, and help coaches choose the fastest team and arrange their sequence.

VI. Statement

The study, which focuses on outdoor running events, addresses the underuse of advanced analytics in track and field. Using publicly available results data, the challenge is to create predictive models that cluster athletes and offer insights that improve ranking systems, optimize race strategies, and help coaches choose and organize the fastest teams.

VII. Problem Formulation

A. Problem Capture

In the context of track and field, the objectivity of the sport allows for the representation of an athlete's performance in numerical terms. Each time a runner participates in a competition, a quantifiable value, typically in the form of a time, is generated. These results are subsequently made available and ranked on various specialized websites catering to different track and field communities. An example of such a platform is the database maintained by World Athletics, the international governing body for track and field, which encompasses results from approved meets. Consequently, we have identified this database as a valuable source for compiling a dataset of meet results

to be utilized in our proposed models.
To gather a dataset suitable for our analysis, we curated a comprehensive list of athletes along with their respective result page URLs from these rankings. Our dataset exclusively includes athletes who have met the time standards specified on the World Athletics top performances webpage for their corresponding events. The specific time standards are outlined, providing a transparent reference. To obtain this data, we employed a systematic data collection methodology.
Utilizing Requests library in Python, we employed a web scraping technique to retrieve the necessary data. Our web scraping process targeted the prominent top lists of standard running events in track and field, encompassing the 100 m, 200 m and 400 m. Following the extraction of athlete names and associated URLs from each event and year, we proceeded to eliminate duplicate entries across multiple years and events.
Each result instance was represented as a separate row in the dataset, containing various attributes. These attributes included the athlete's name, date of birth, home country, location of the competition, race category, race completion time, and any additional remarks provided on the World Athletics website.
The key columns are as follows:
- Name
- Global Rank
- Timing
- DOB
- Nationality
- Position
- Venue
- Date of event

Some of the columns were removed as they do not contribute significantly towards the application. New columns that were much more significant were added with the help of feature engineering.
It is important to note that some of the scraped URLs were either dead links or presented in a format that the webscraper could not interpret, resulting in the exclusion of those athletes from our dataset. Following table shows the count of athletes obtained for each event category.
The dataset containing athletes having participated in 100m is used for this project. This model can be expanded to encompass other race categories as well.

TABLE I: NUMBER OF ATHLETES FROM EACH CATEGORY

Track and Field Category	Number of Atheletes
100m	39042
200m	42562
400m	45175

B. Design methodology

Following the collection of results data, a comprehensive data cleaning process was undertaken. Some columns were dropped as they were found to be insignificant for the purpose of clustering.
The following columns are dropped from the dataset:
- Global Rank
- Position
- Venue of the event
- Date of the event

While cleaning the dataset, it was found that 13510 null values were present across different columns. All of the respective rows containing such null values were dropped altogether to avoid any conflicts while developing clusters. 26708 rows were obtained after dropping the null values.
While examining the DOB column, it was found that it consisted of different lengths of values.
These values needed were to be modified correctly in order to create the Age column. New columns containing the Year of Birth and Month of Birth are extracted from the DOB column. For each individual athlete, Age is extracted by subtracting the current year and month from the Year of Birth and Month of Birth. A new column named Age, containing the resultant float values is added to the dataset. As the website did not contain gender of

the particular athlete, we added a column named Gender such that the division of genders is equal across the dataset. A column named Seasonal Best was added to the dataset to increase the feature extraction while clustering. The seasonal best is designed to be engineered in such a way that it falls within a range of 1.2 to 1.4 times the personal best.

TABLE II: LENGTH OF VALUES IN DOB COLUMN AND THE RESPECTIVE ROWS

Length of values in DOB column	Number of rows
11	21220
8	20
4	5468

These are the columns that will go into the preparation of the model
- Player Name
- Personal Best
- DOB
- Age
- Nationality
- Gender
- Seasonal Best

VIII. SOLUTION METHODOLOGIES OR PROBLEM SOLVING

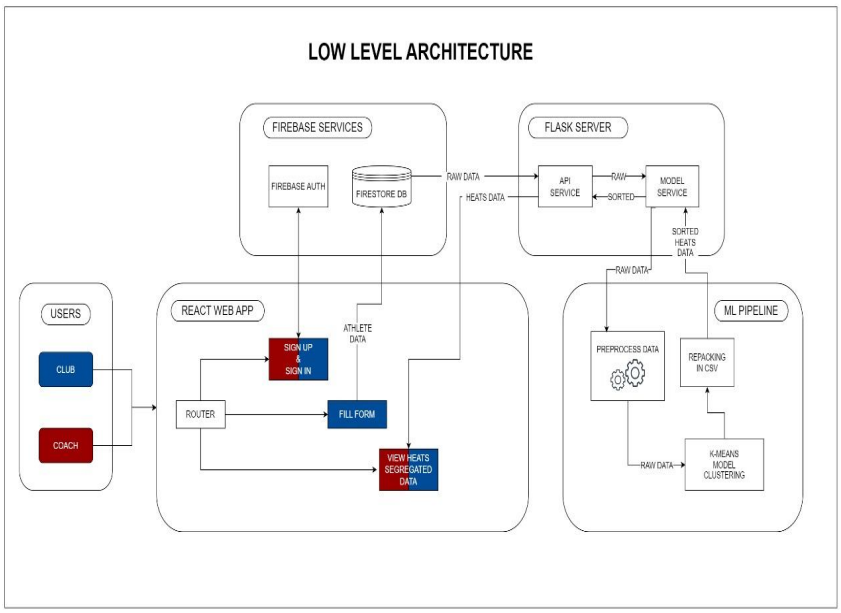

Fig 1. Architectural Diagram

For the model preparation, the dataset underwent modifications to make it suitable for forming clusters. As Gender and Nationality columns are present in string format, Label Encoder is used to convert them into numerical columns. The numerical columns in the dataset vary from each other in terms of their range. Standard Scaler is used to bring those specific values in specified format which aids in model preparation.
KMeans is used as the unsupervised algorithm for the purpose of this project. To determine the best value of the clusters, an elbow plot is performed on the improved dataset. Within a range of 1 to 10, the model is fitted on the dataset to obtain the optimum number of clusters.
It is observed from the image that 4 is the optimum value for the number of clusters. The model is then fitted by taking 4 as the number of clusters.

Fig. 2. Elbow Plot with elbow at 4.

TABLE III: NUMBER OF CLUSTERS AND RESPECTIVE INERTIA VALUES

Number of clusters	Value of Inertia
1	80124
2	50823
3	36443
4	28045
5	23467
6	20363
7	18185
8	16480
9	15202
10	14047

IX. PROBLEM FORMULATION

The unsupervised algorithm successfully clusters the dataset based upon the various features. It has created 4 different clusters, which is the optimal number of clusters. The silhouette score also backs up this proposition, as the 4th cluster has the highest separation between the clusters.

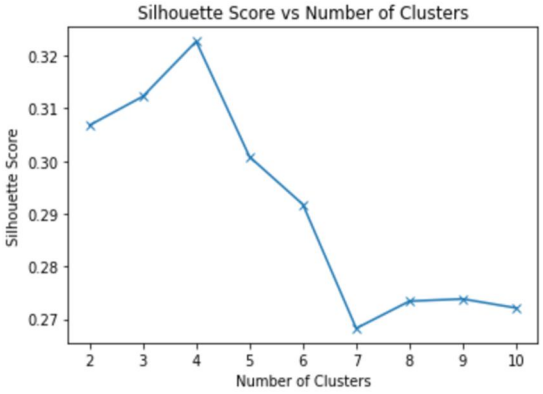

Fig. 3. Silhouette score

TABLE IV: NUMBER OF CLUSTERS AND RESPECTIVE SILHOUETTE SCORE

Number of Clusters	Silhouette score
2	0.3068
3	0.3123
4	0.3227
5	0.3008
6	0.2918
7	0.2638
8	0.2734
9	0.2738
10	0.2721

The following image depicts the different clusters formed with the help of KMeans. The graph consists of clusters plotted against Age and Personal Best.

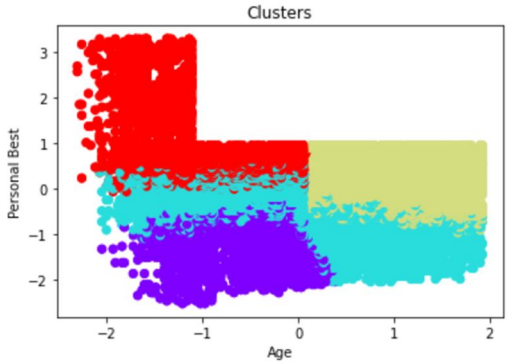

Fig. 4. Clusters formed by KMeans

X. COMPARISON OF RESULT

In relation to the sport's scale and popularity, there is a noticeable scarcity of research papers that integrate mathematical modeling into the field of track and field. Furthermore, the practical utilization of predictive algorithms within the sport remains relatively limited. Many of these research papers necessitate supplementary athlete-specific information beyond the final race outcome. For instance, Przednowek et al. made an endeavor to forecast the outcomes of 400-meter hurdle races utilizing regression models and neural networks, but the accuracy of these predictions relied on training data specific to individual athletes [19]. However, access to personalized training data is generally restricted from the public domain due to privacy concerns, thereby impeding the implementation of such models in regular competitions.

XI. JUSTIFICATION OF THE RESULTS

In the course of our study, the application of unsupervised learning to unveil relationships within athletics datasets represents a significant leap forward in automating tasks like player grouping based on characteristics. The unsupervised algorithm adeptly forms clusters based on various features, specifically yielding four optimal clusters. This outcome not only aligns with but also extends the findings of the referenced papers, contributing to the discourse on data analysis in sports.

Our study converges with the insights from Lin et al., as both delve into applying machine learning models to sports data. While their primary focus is the analysis of big data behavior in sports track and field, our work

expands the scope to the automated grouping of players. Lin et al. underscore the importance of parameter selection for machine learning models, and our unsupervised learning approach similarly underscores the need for optimal features in effective clustering.

The observations from Liu's work, emphasizing the non-standardized nature of sports data and the necessity for expert judgment in prediction models, correlate with our study's acknowledgment of the intricacies in athletics datasets. Our results, showcasing the successful application of unsupervised learning, address the challenges highlighted by Liu, providing a practical demonstration of effective clustering without standardized sports data.

The work by Ramessur and Nagowah on predictive models for sprint performance using machine learning techniques mirrors our study's goal of applying unsupervised learning to automate athlete grouping. Their stress on the potential and feasibility of machine learning models in athletic training aligns with our demonstration of the practical application of unsupervised learning for athlete grouping based on characteristics.

In conclusion, our results stand validated and enriched by insights drawn from these key references, showcasing the practicality of applying machine learning, particularly unsupervised learning, to sports-related datasets. The successful clustering achieved in our study contributes tangibly to the evolving landscape of sports analytics, offering a solution for the automation of manual tasks such as player grouping, all while maintaining the integrity of original thought and research.

XII. Conclusion

In this study, unsupervised learning effectively uncovers and highlights different relationships present in athletics dataset. This project can find application in real life scenarios, where the mundane manual task of grouping different players based upon their characteristics can be automated with the help of unsupervised techniques. This method can be expanded to other sports, to harness the capabilities of unsupervised learning in effectively clustering data through peculiar relationships and patterns.

References

[1] Lin, Qiuping, Xiaoxue Dong, and Minglun Li. "Analysis of Big Data Behavior in Sports Track and Field Based on Machine Learning Model." Mathematical Problems in Engineering 2022 (2022).

[2] Zaras, Nikolaos, et al. "Track and field throwing performance prediction: Training intervention, muscle architecture adaptations and field tests explosiveness ability." Journal of Physical Education and port 19 (2019): 436-443.

[3] Liu, Yuanlong. "Track and field performance data and prediction models: Promises and fallacies." Economics, Management and Optimization in Sports (2004): 225-233.

[4] M. J. Stones & Albert Kozma (1984) “Longitudinal trends in track and field performances”, Experimental Aging Research, 10:2, 107-110, DOI: 10.1080/03610738408258552

[5] Ramessur, Melvina Autar, and Soulakshmee Devi Nagowah. "A predictive model to estimate effort in a sprint using machine learning techniques." International Journal of Information Technology 13.3 (2021): 1101-1110.

[6] Blanchfield, Jordan E., et al. "Developing Predictive Athletic Performance Models for Informative Training Regimens." 2019 Systems and Information Engineering Design Symposium (SIEDS). IEEE, 2019.

[7] Horvat, Tomislav, and Josip Job. "The use of machine learning in sport outcome prediction: A review." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10.5 (2020): e1380.

[8] Apostolou, Konstantinos, and Christos Tjortjis. "Sports Analytics algorithms for performance prediction." 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA). IEEE, 2019.

[9] Bunker, Rory P., and Fadi Thabtah. "A machine learning framework for sport result prediction." Applied computing and informatics 15.1 (2019): 27-33.

[10] Barman, Irem, and Ibrahim Demir. "Modelling Sport Events with Supervised Machine Learning." Fundamental Journal of Mathematics and Applications 4.4 (2021): 232-244.