Supervised Machine Learning Algorithms for the Prediction of Students Long Jump distances

1st Karim Midoul
New Technology Trends for Innovation
Faculty of Sciences
Tétouan, Morocco

2nd Badr Eddine EL Mohajir
New Technology Trends for Innovation
Faculty of Sciences
Tétouan, Morocco

3rd Outman El Hichami
New Technology Trends for Innovation
Higher Normal School, Abdelmalek University
Tétouan, Morocco

4nd Adnan Souri
New Technology Trends for Innovation
Faculty of Sciences
Tétouan, Morocco

***Abstract*—This study delves into the predictive analysis of factors affecting jump distance in students, employing various machine learning models. The models used include Linear Regression, Ridge Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, and Extra Trees Regression. The performance of these models was evaluated based on metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Explained Variance Score (EVS), Mean Squared Log Error (MSLE), R^2, and computation time. The analysis aims to identify the most effective model for predicting jump distances, thereby providing insights for athletic training and performance enhancement and to develop a reliable and accurate model that can assist physical education teachers in evaluating the jumping abilities of their students and designing appropriate training programs.**

***Index Terms*—long jump, predecting distance, performance, Sports analytics, Machine learning models, Comparative analysis.**

I. Introduction

The most athletics exercises are simple (fundamental) movements. However, the technique of competitive types of track and field isn't always easy to study and demonstrate qualitatively [1], [2], [3], [4].

In the context of the long jump, the primary objective is to generate speed on the running track and execute a jump that maximizes the distance from the takeoff board.

Amidst the numerous factors influencing performance, the horizontal velocity emerges as a pivotal biomechanical determinant of flight distance [5], [6]. Notably, some elite long jumpers, including Carl Lewis and Marion Jones, are recognized for their proficiency as high-level sprinters [7].

Contrary to the assumption that the fastest sprinters necessarily excel in long jump, it is posited that the best long jumpers are indeed the fastest ones. This correlation is substantiated by the biomechanical analysis report of the 2009 IAAF World Athletics Championships, where top-ranking athletes exhibited higher run-up velocities than their counterparts [8]. Observations of the world's top three athletes revealed a consistent horizontal velocity of 11 m/s [9]. Evidently, run-up velocity stands out as the foremost determinant of long jump performance [10],[11].

Research affirms a robust relationship, with a correlation coefficient of 0.96, between horizontal velocity and jump distance [12].Consistent findings across various studies further establish the association between velocity and jump [13]. Artificially manipulating run-up velocity has been shown to result in a notable increase in jumping distance. Calculations indicate that a mere 0.1 m/s increase in velocity corresponds to a rise in jump distance ranging from 6 to 12.8 cm [14].

This study investigates the factors influencing jump distance in athletes, utilizing a dataset encompassing various physiological and performance metrics. Key variables analyzed include age, sex, weight, height, body mass index (BMI), and speed. Through exploratory data analysis (EDA) and correlation analysis, the study identifies significant determinants of jump performance, offering insights for coaches and athletes aiming to enhance jumping capabilities.

The remainder of this paper is organized as follows: Section 2, we start by reviews related works, providing context for factors influencing jump distance. Next, Section 3, we will present details the materials and methods employed in the study. In section 4, presents the experiment results, while the Discussion section interprets findings. The paper concludes with the Perspectives section, offering insights into potential avenues for future research and applications.

II. Related Works

The investigation into the most efficient machine learning algorithm for predicting the long jump distance of male athletes, incorporating age and velocity variables, is a significant contribution to the field. In this study by Uçar as mentioned earlier [2]. Data from 328 valid jumps by 73 Turkish male athletes were analyzed, employing various performance metrics, including MAE, RMSE, MSE, R2 score, EVS, and MSLE, to assess algorithmic performance. The findings highlighted the precision in determining long jump performance through carefully chosen independent variables. The utilization of the 5-fold cross-validation technique further enhanced the evaluation of model performance, identifying the Gradient Boosting Regression Trees (GBRT) algorithm as the most successful

with an achieved MSE value of 0.0865. The study suggests that the proposed machine learning approach has valuable potential for trainers in evaluating the long jump performance of male athletes, providing a foundation for further research in this domain.

III. MATERIALS AND METHODS

A dataset of 2,123 athletes was utilized, encompassing variables like age, sex, weight, height, BMI, speed, and jump distance. The data was split into training and testing sets. Six machine learning models were trained and evaluated. The models' performances were compared using metrics like MSE, RMSE, MAE, EVS, MSLE, R^2. Additionally, actual vs. predicted jump distances were visually compared using scatter plots .

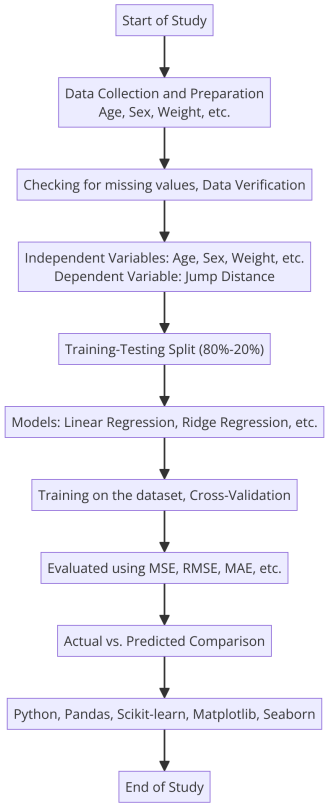

Fig. 1. Methodologie graph of this study

A. Data Collection and Preparation

- **Dataset Description**: The study utilized a dataset comprising 2,123 students, encompassing various attributes such as age, sex, weight, height, BMI, speed, and jump distance. Age: Ranges from 11 to 19 years with an average of approximately 14.5 years. Sexe: Appears to be a binary variable (0 for female and 1 for male), indicating two genders. Weight_kg: Varies from 28 to 94 kg, with an average weight around 48.4 kg. Height_m: Heights range from 1.34 to 1.91 meters, with a mean height of 1.61 meters. BMI: The Body Mass Index ranges from 12 to 33, with an average of 18.6. Speed_m/s: Speed ranges from 3.08 to 7.59 meters per second, with an average speed of 5.37 m/s. Distance_m: Jump distances vary from 1.52 to 5.66 meters, with an average distance of 3.02 meters.

Variables	n	Mean	Min	Max
Age (year)		14.5	11	19
Weight$_{kg}$		48,4	28	94
Height$_{m}$	2123	1,61	1,34	1,91
BMI		18,6	12	33
Speed$_{m/s}$		3,08	5,37	7,59
Distance$_{m}$		3,02	1,52	5,66

TABLE I
MEAN AND STANDAR DEVIATION VALUES OF THE VARIABLES FOR THE SAMPLES

Fig . 2 shows histograms that provide a visual representation of the distribution of each variable in the dataset: **Age Distribution**: Shows a relatively uniform distribution across the age range, with a slight increase in frequency for middle ages. **Sex Distribution**: Indicates a nearly equal distribution between the two categories (0 and 1), suggesting balanced representation of sexes. **Weight Distribution**: The weight appears to be normally distributed, centered around the mean weight. **Height Distribution**: Like weight, height also shows a roughly normal distribution. **BMI Distribution**: BMI values are somewhat normally distributed, with a concentration around the mean value. **Speed Distribution**: The speed at which individuals jump shows a normal distribution, centered around the mean speed. Distance Distribution: Jump distances are somewhat normally distributed, with a slight skew towards shorter distances.

- **Preprocessing**: The dataset was checked for missing values and inconsistencies. No missing values were found, and data was verified for correctness. For machine learning purposes, the data was left in its original form, with 'Sexe' treated as a categorical variable.

B. Feature Selection

- Independent Variables: Age, sex, weight, height, BMI, and speed .
- Dependent Variable (Target): Jump distance.

Fig. 3 shows the relationships between these variables, especially focusing on how factors such as age, sex, weight, height, BMI, and speed relate to the jump distance ('Distance_m'). Speed vs. Distance: There is a strong positive correlation (0.74) between speed and jump distance, indicating that higher speeds are associated with longer jump distances. Age vs. Distance: No clear trend is visible, aligning with the weak correlation observed. Sex vs. Distance: There seems to be a slight difference in jump distances between the two sexes, but it's not very pronounced(0.64). Weight vs. Distance: The relationship is not very strong, but there's a slight tendency for lighter

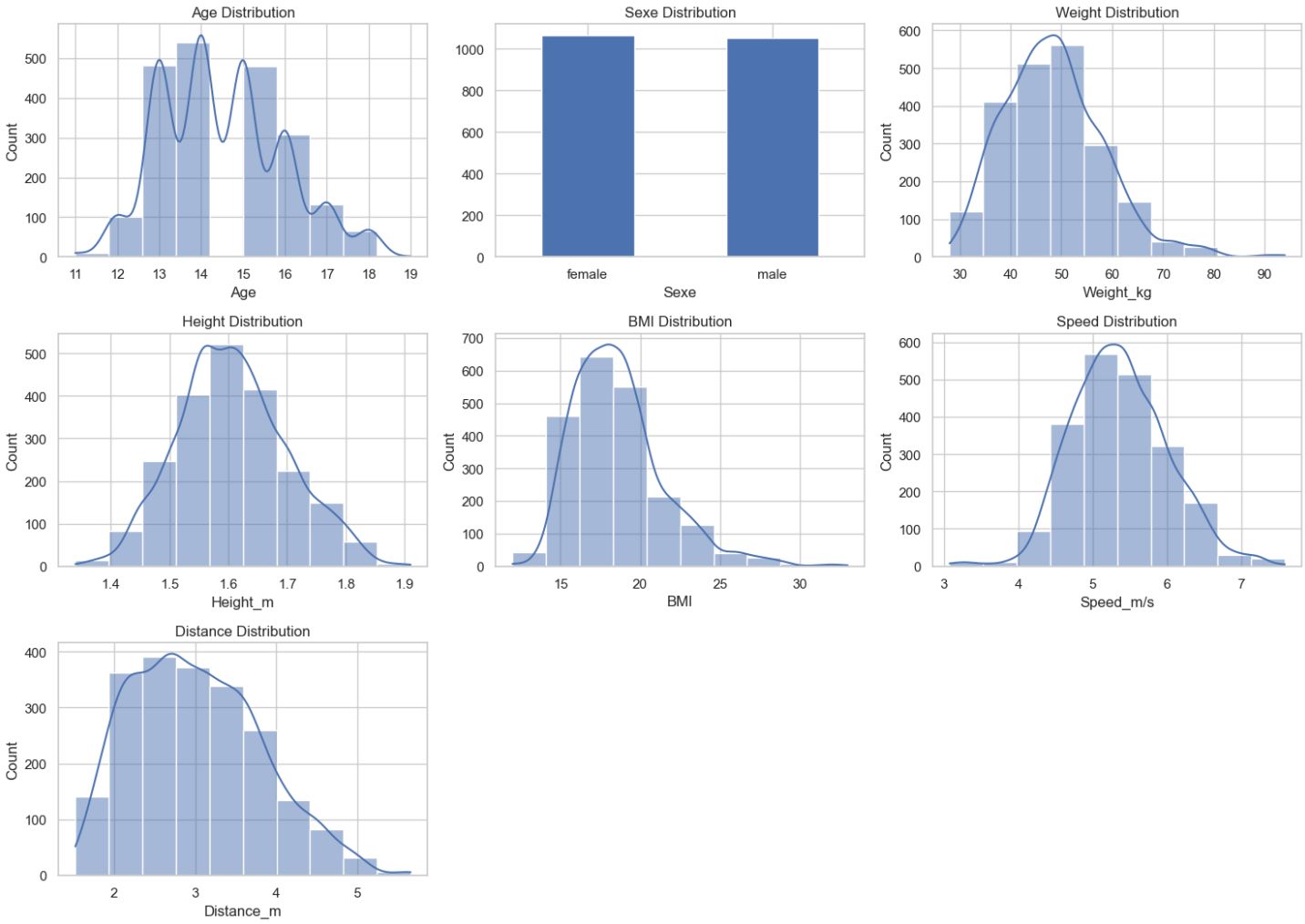

Fig. 2. Histograms Distibutions of Various Variables

individuals to jump farther. Height vs. Distance: Taller individuals show a tendency to jump farther, as reflected in the correlation. BMI vs. Distance: The trend is not very strong, but it slightly leans towards lower BMI being associated with longer jumps.

C. Splitting the Data:

Training and Testing Sets: The data was split into training (80%) and testing (20%) sets using stratified sampling to maintain the distribution of key variables.

D. Model Selection:

In this paper six popular machine learning techniques were used: Linear Regression, Ridge Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, and Extra Trees Regression .

- **Linear Regression** is a supervised learning algorithm that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation. It aims to find the best-fit line that minimizes the sum of squared differences between observed and predicted values [15].
- **Ridge Regression** is a linear regression technique that introduces regularization by adding a penalty term to the ordinary least squares objective function. It helps prevent overfitting by penalizing large coefficients [16]. **Decision tree regression** is a non-linear regression algorithm that models data using a tree-like structure. It recursively splits the dataset based on feature values to predict the target variable [17]. **Random forest regression** is an ensemble learning method that constructs a multitude of decision trees during training and outputs the average prediction of the individual trees for regression tasks [18]. **Gradient Boosting Regression** is an ensemble learning technique that builds a predictive model by combining the predictions of multiple weak learners, often decision trees. It sequentially corrects errors made by previous models [19]. **Extra Trees**

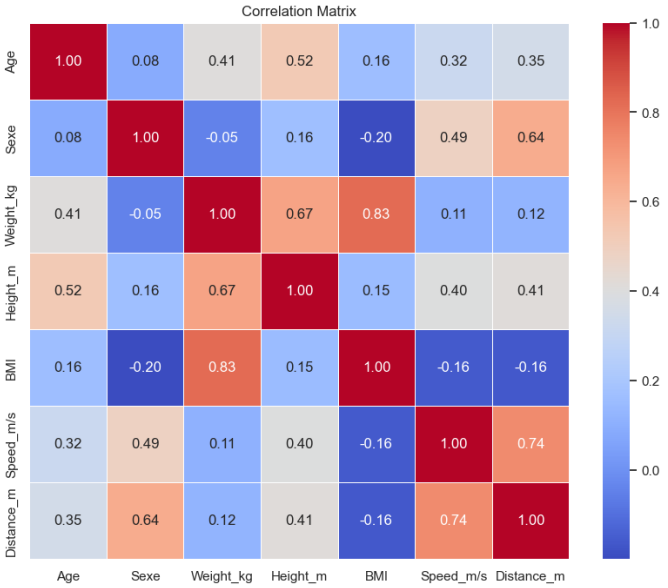

Fig. 3. Correlation Matrix between variables

(Extremely Randomized Trees) regression is an ensemble learning method similar to random forests. It introduces additional randomness during the tree-building process, leading to a more diverse set of trees [20].

E. Model Training :

- Procedure: Each model was trained on the training dataset. Default hyperparameters were used initially, with the scope for tuning in future iterations of the study.

F. Model Evaluation :

Metrics: Models were evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Explained Variance Score (EVS), Mean Squared Log Error (MSLE), R^2, and computation time.

- **Mean Squared Error (MSE):** measures the average squared difference between actual and predicted values, providing a measure of the model's accuracy [21].
- **Root Mean Squared Error (RMSE)**: is the square root of MSE, offering a measure of the average magnitude of errors in the model's predictions [22].
- **Mean Absolute Error (MAE):** calculates the average absolute difference between actual and predicted values, providing a robust measure of model accuracy [23].
- **Explained Variance Score (EVS):** quantifies the proportion of variance in the dependent variable explained by the model. It ranges from 0 to 1, with 1 indicating a perfect prediction. [24].
- **Mean Squared Log Error (MSLE)**: measures the average squared logarithmic difference between actual and predicted values, often used when the target variable spans several orders of magnitude [25].
- **R^2 (Coefficient of Determination):** measures the proportion of variance in the dependent variable explained by the independent variables. It ranges from 0 to 1, with 1 indicating a perfect prediction [26].

IV. Experiment Results

In the conducted experiments, the data were partitioned, with 80% allocated for training purposes and the remaining 20% for evaluation. To gauge the predictive capabilities of the algorithms, six error measurement techniques were employed, allowing for a comprehensive assessment of their performance.

A. Validation

Table II shows the comparison of the machine learning models using the specified metrics:
Extra Trees Regression has the lowest MSE, RMSE, MAE, and MSLE values, indicating the highest precision and lowest error among all models. It also has the highest R^2 and EVS, showing its strong predictive accuracy and ability to explain the variance in the data. However, it takes relatively more time to train.
Random Forest Regression also performs well with low error metrics and high R^2 and EVS, but it has the longest training time.
Decision Tree Regression shows good performance with relatively low error metrics and high R^2, but it may be prone to overfitting.
Linear Regression and **Ridge Regression** have moderate performance with similar metrics, suitable for simpler linear relationships.
Gradient Boosting Regression balances performance and training time, offering a good compromise between model complexity and predictive accuracy.
The choice of model would depend on the specific requirements of accuracy, interpretability, and computational efficiency. For high accuracy and complex pattern recognition, Extra Trees and Random Forest are preferred, while for quicker, simpler models, Linear and Ridge Regressions are suitable. **Actual vs. Predicted Comparison**: Scatter plots were created to visually compare the actual and predicted jump distances for each model, providing a qualitative assessment of model performance. Fig. 4 shows scatter plots displaying the actual versus predicted jump distances for each machine learning model:
Linear Regression, Ridge Regression, Gradient Boosting Regression: These models show a fairly good alignment of predictions with actual values, but

Model	MSE	RMSE	MAE	EVS	MSLE	R^2
Linear Regression	0.198503	0.445536	0.351819	0.693015	0.013006	0.692725
Ridge Regression	0.198571	0.445614	0.351870	0.692908	0.013007	0.692619
Decision Tree Regression	0.137174	0.370370	0.180094	0.787988	0.009134	0.787659
Random Forest Regression	0.063057	0.251111	0.181813	0.902410	0.004119	0.902391
Gradient Boosting Regression	0.143268	0.378508	0.301771	0.778290	0.009209	0.778226
Extra Trees Regression	0.025478	0.159617	0.094374	0.960562	0.001611	0.960562

TABLE II
PERFORMANCE COMPARISONS OF MACHINE LEARNING ALGORITHMS AND OTHER MODELS FOR DISTANCE PREDICTION

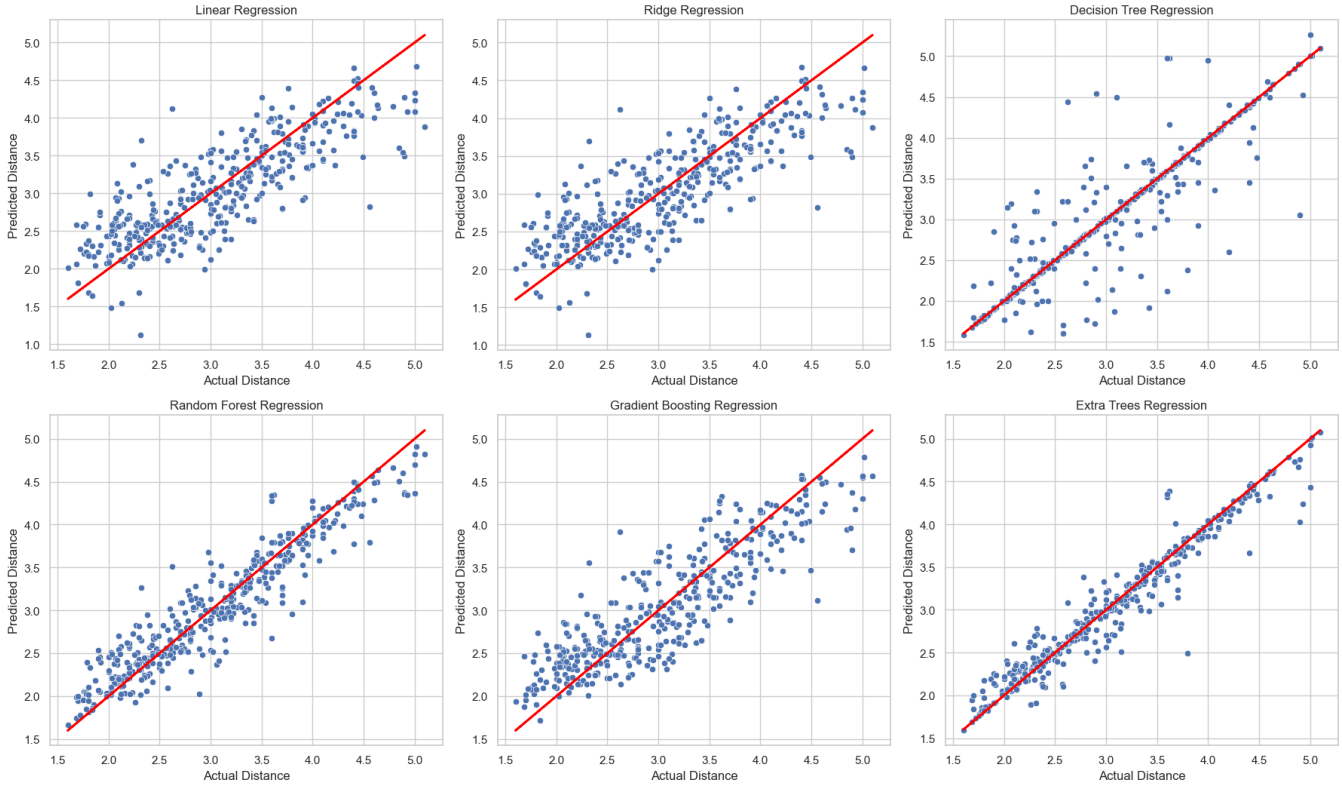

Fig. 4. Scatter Plots Comparaison of Machine Learning Model Performance

there's some deviation, especially for higher values of jump distances.

Decision Tree Regression: The plot indicates a relatively good fit, but with some overfitting, as seen by the points clustering on specific lines. This is typical for decision trees, which can create overly complex models capturing noise in the training data.

Random Forest Regression: This model shows a strong alignment between the predicted and actual values, with the points closely following the diagonal line, indicating high accuracy.

Extra Trees Regression: The plot for this model demonstrates an excellent fit, with predictions closely matching the actual values. The points are very near to the diagonal line, reflecting the high accuracy and generalization capability of this model. In all the plots, the red line represents the line of perfect prediction. The closer the scatter points are to this line, the better the model's predictive performance. The Extra Trees and Random Forest models stand out for their close alignment to the line, indicating superior predictive capability in comparison to the other models.

B. Tools and Technologies Used:

Software and Libraries: The analysis was conducted using Python, with libraries such as Pandas for data manipulation, Scikit-learn for machine learning models and evaluation, and Matplotlib and Seaborn for data visualization.

C. Results:

The Extra Trees Regression model exhibited superior performance across most metrics, followed closely

by the Random Forest Regression model. Linear and Ridge Regressions showed moderate performance, suitable for simpler linear relationships. The Decision Tree and Gradient Boosting models offered a balance between complexity and accuracy. The Extra Trees Regression model's predictions closely matched the actual jump distances, indicating its high predictive accuracy and generalization capability.

D. Discussion:

The results demonstrate the efficacy of ensemble methods, particularly Extra Trees Regression, in modeling complex relationships in athletic performance data. These models capture intricate patterns and interactions among variables, which linear models may not effectively address. The findings suggest that while simpler models offer quicker training times, they may not achieve the high accuracy and nuanced understanding provided by more complex models.

CONCLUSION

The study underscores the necessity of focusing on speed and height in training regimens aimed at improving jumping ability. While other factors like weight and BMI have some influence, they are less pivotal. These insights can guide athletes and coaches in optimizing training strategies to enhance jump performance.

In Addition, this study underscores the potential of machine learning in sports analytics, particularly in predicting and improving athletic performance. The Extra Trees Regression model, with its high accuracy and reasonable computation time, emerges as a valuable tool for athletes and coaches. Future research could explore the integration of these models into real-time training environments and investigate their applicability in other sports performance metrics.

In future work, we plan to develop an application based on this research to predict the performance of the long jump (the distance) in students, aiming to identify young talents and also to design tailored training programs.

Conflict of Interest: The authors declare that they have no conflict of interest.

Data Availability Statement : Data available on reasonable request.

REFERENCES

[1] Andrii Yefremenko, Svitlana iatysotska, and Viktor Pavlenko. The comparison of students' long jump study programs. *Slobozhanskyi Herald of Science and Sport*, 27(3):110–117, 2023.

[2] UÇAR Murat, Mürsel Ozan İNCETAŞ, Işık Bayraktar, and Murat Çilli. Using machine learning algorithms for jumping distance prediction of male long jumpers. *Journal of Intelligent Systems: Theory and Applications*, 5(2):145–152, 2022.

[3] Lee-Kuen Chua, Judith Jimenez-Diaz, Rebecca Lewthwaite, Taewon Kim, and Gabriele Wulf. Superiority of external attentional focus for motor performance and learning: Systematic reviews and meta-analyses. *Psychological Bulletin*, 147(6):618, 2021.

[4] Yuta Takanashi. The relationship between jump ability and athletic performance in athletic throwers. *Sport Mont*, 19(1):71–76, 2021.

[5] Jaimes G Hay. The biomechanics of the long jump. *Exercise and sport sciences reviews*, 14:401–446, 1986.

[6] Nicholas P Linthorne. Analysis of standing vertical jumps using a force platform. *American Journal of Physics*, 69(11):1198–1204, 2001.

[7] Kyle Davis, Stephen Rossi, Jody Langdon, and Jim McMillan. The relationship between jumping and sprinting performance in collegiate ultimate athletes. *Journal of Coaching Education*, 5(2):24–37, 2012.

[8] Rolf Graubner and Eberhard Nixdorf. Biomechanical analysis of the sprint and hurdles events at the 2009 iaaf world championships in athletics. *Positions*, 1:10, 2009.

[9] Chris McCosker, Ian Renshaw, Daniel Greenwood, Keith Davids, and Edward Gosden. How performance analysis of elite long jumping can inform representative training design through identification of key constraints on competitive behaviours. *European journal of sport science*, 19(7):913–921, 2019.

[10] Zijian Shang. Research on the factors influencing long jump. In *Proceedings of the 2022 5th International Conference on Mathematics and Statistics*, pages 36–40, 2022.

[11] Teerawat Kamnardsiri, Worawit Janchai, Pattaraporn Khuwuthyakorn, and Wacharee Rittiwat. Implementation and validity of the long jump knowledge-based system: Case of the approach run phase. *PeerJ Preprints*, 7:e27524v1, 2019.

[12] Lisa A Bridgett and Nicholas P Linthorne. Changes in long jump take-off technique with increasingrun-up speed. *Journal of sports sciences*, 24(8):889–897, 2006.

[13] Adrian Lees, Philip Graham-Smith, and Neil Fowler. A biomechanical analysis of the last stride, touchdown, and takeoff characteristics of the men's long jump. *Journal of applied Biomechanics*, 10(1):61–78, 1994.

[14] Agoston Schulek. Long jump with supramaximal and normal speed. *NEW STUDIES IN ATHLETICS*, 17(2):37–46, 2002.

[15] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

[16] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[17] Trevor Hastie, Robert Tibshirani, Jerome Friedman, Trevor Hastie, Robert Tibshirani, and Jerome Friedman. High-dimensional problems: p n. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, pages 649–698, 2009.

[18] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[19] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[20] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63:3–42, 2006.

[21] Fabian Pedregosa. Scikit-learn: Machine learning in python fabian. *Journal of machine learning research*, 12:2825, 2011.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.