Machine learning modeling for environmental factors affecting horse racing

YAVUZKAN PAKSOY[1]*, UĞUR DURUK[2], AHMET AKAY[3], AHMET KOLUMAN[4]

[1] Necmettin Erbakan University; Ere li Kemal Akman Vocational College, Department of Plant and Animal Production, Konya, Türkiye
[2] Pamukkale University, Biomedical Engineering Dept, 20160, Denizli, Türkiye
[3] Pamukkale University, Biomedical Engineering Dept, Denizli, Türkiye
[4] Pamukkale University, Biomedical Engineering Dept, Denizli, Türkiye

SUMMARY

This study examines the impact of environmental factors on horse racing performance using machine learning techniques, offering insights into how climate and track conditions affect race outcomes. Horse racing is significantly influenced by external conditions, with variables such as temperature, humidity, track surfaces, and wind patterns playing crucial roles. By analyzing historical race data, this research helps trainers, bettors, and race organizers understand these factors. Support Vector Machines (SVM), Decision Trees, K-Nearest Neighbors (KNN), Random Forest, and AdaBoost were employed to model race performance, with SVM achieving the highest accuracy. Unlike other sports where athletes control their environment, racehorses must adapt to external conditions. Traditional statistical methods often fail to capture the complex relationships between these factors. Machine learning, however, can identify nonlinear patterns in data and provide a more dynamic approach to analyzing race performance. The study finds that temperature, humidity, wind, and track conditions are key influences. Moderate temperatures (10-21°C) are ideal for optimal performance, while extreme heat causes fatigue and cold leads to stiffness. Higher humidity adds stress, and wind patterns can either hinder or assist a horse's speed. Track surfaces, including dirt, turf, and synthetic, also affect a horse's grip and stability, with wet conditions slowing horses down. The study's findings contribute to a data-driven approach in horse racing, allowing trainers to adjust strategies based on evidence. Ultimately, this research demonstrates how machine learning can revolutionize horse racing, offering more precise predictions, improved strategies, and a focus on equine welfare in response to environmental challenges.

KEY WORDS

Climate; Horse; Machine learning modeling; Performance.

INTRODUCTION

Horse racing is a dynamic sport influenced by a range of environmental conditions [1,2]. Understanding these influences is crucial for trainers, bettors, and industry stakeholders. Traditional statistical models have been used to assess performance, but machine learning presents a more sophisticated approach by recognizing complex, nonlinear relationships between variables. This paper integrates machine learning methodologies to evaluate the role of environmental factors in race outcomes.

Environmental Factors Affecting Horse Racing

Track Conditions

The type and condition of the racetrack significantly influence horse performance, making it one of the most critical environmental factors in horse racing. The three primary track surfaces-dirt, turf, and synthetic-each have unique characteristics that affect race outcomes in different ways [2].

- **Dirt Tracks:** Dirt tracks are the most common racing surface and tend to be affected significantly by weather conditions. When dry, dirt tracks can provide firm footing, resulting in faster race times. However, after heavy rain, dirt tracks can become muddy or sloppy, increasing drag and making it more difficult for horses to maintain their speed. Some horses perform well on wet tracks, while others struggle due to the increased resistance and instability. Machine learning models can analyze historical race data to identify which horses perform best under specific track conditions, allowing trainers and bettors to make more informed decisions [3,4].
- **Turf Tracks:** Turf (grass) tracks offer a softer surface compared to dirt, which can be beneficial for reducing injury risks but also slows down race times. Unlike dirt tracks, turf surfaces tend to absorb moisture differently, meaning they can become heavy and slippery in wet conditions. This can affect traction and increase the likelihood of horses losing their footing. Additionally, seasonal changes can impact turf track conditions, as grass quality, density, and main-

*Corresponding Author:
Yavuzkan Paksoy (akoluman@pau.edu.tr)

tenance routines vary throughout the year. Machine learning algorithms can track these variations and adjust performance expectations based on real-time data [3,4].

- **Synthetic Tracks:** Synthetic tracks, made from materials such as wax-coated sand, rubber, and fiber, were introduced to provide more consistent racing conditions and reduce the risks associated with extreme weather. These tracks are less affected by rain and temperature fluctuations, offering more predictable performance outcomes. However, not all horses adapt well to synthetic surfaces, as the way they interact with the ground differs from natural surfaces. Some horses excel on synthetic tracks, while others struggle with the difference in traction and rebound effect. Predictive models trained on past race performances can help assess how individual horses respond to synthetic tracks, improving strategic planning for races.

Another critical factor is the track maintenance process, which influences how surfaces behave over time [5]. Track grooming, watering schedules, and compaction levels all contribute to race conditions. Machine learning models can integrate track maintenance logs with weather data to predict how a track will perform on race day, offering a significant advantage for trainers and bettors alike.

Temperature and Humidity

Temperature and humidity play crucial roles in horse racing performance, directly impacting the physiological and metabolic processes of racehorses [6].

Temperature: Horses perform optimally in moderate temperatures ranging from 10-21 °C, as this range allows for efficient thermoregulation and muscular function. In contrast, extreme heat (above 35 °C) increases the risk of heat exhaustion, dehydration, and reduced endurance. Studies have shown that racehorses experience a 3-10% decline in speed when racing in high temperatures, as their bodies struggle to dissipate excess heat. Machine learning models can analyze temperature trends and predict how horses will respond under varying conditions, helping trainers adjust training loads and hydration strategies [7,8,9,10,11].

Conversely, cold temperatures (below 5 °C) can also impact performance, though in different ways. In colder weather, horses may experience muscle stiffness and reduced flexibility, leading to slower starts and increased susceptibility to injuries. However, some horses are naturally better adapted to colder climates, meaning personalized predictions based on individual horses past performances can be valuable [7].

- **Humidity:** While temperature alone is an important factor, its effect is amplified when combined with humidity levels. High humidity makes it difficult for horses to cool down effectively, as sweating-their primary cooling mechanism-becomes less efficient. Dew point measurements, which factor in both temperature and humidity, are increasingly used in race analytics to assess the likelihood of heat stress in horses. Machine learning models trained in past race results can incorporate humidity data to refine predictions, identifying thresholds where performance declines significantly [8,9,10,11].

The interaction between temperature, humidity, and hydration strategies is also crucial. Trainers must adjust their race-day routines based on weather forecasts, ensuring that horses remain properly hydrated and that race pacing is adjusted to account for environmental stressors. Data-driven insights from predictive models can assist in these decisions by analyzing historical races under similar conditions.

Wind Patterns

Wind is an often overlooked but significant environmental factor in horse racing [12]. It influences not only race speeds but also energy expenditure, race tactics, and track conditions [13].

Headwinds: Strong headwinds increase aerodynamic resistance, forcing horses to expend more energy to maintain their speed. Studies indicate that sustained headwinds above 10 mph can reduce speeds by 2-5%, with even greater effects in long-distance races. Horses positioned at the front of the pack experience the most resistance, whereas those drafting behind other horses can conserve energy. Machine learning models can use wind speed data to predict how much additional energy will be required under specific conditions, offering tactical advantages for jockeys and trainers.

- **Tailwinds:** In contrast, tailwinds provide a slight performance boost, as they reduce resistance and help horses maintain momentum with less effort. While the benefits of tailwinds are not as pronounced as the drawbacks of headwinds, they can still contribute to faster overall race times, particularly in sprint races where every fraction of a second matters. Machine learning models can incorporate wind patterns to adjust performance expectations, allowing bettors and analysts to fine-tune their predictions.
- **Crosswinds:** Less commonly analyzed but still impactful, crosswinds can disrupt a horse's balance and racing line, especially on curved sections of the track. Jockeys may need to make slight directional adjustments, which can add small inefficiencies to race performance. By integrating real-time wind data, predictive models can identify potential challenges posed by crosswinds and inform strategic adjustments for riders [14,15,16].
- **Impact on Track Conditions**: Wind patterns also affect track conditions over time. Strong winds can dry out dirt tracks, leading to firmer, faster surfaces, or blow debris and sand onto the track, reducing traction and visibility. Turf tracks are particularly sensitive to wind-related moisture changes, as drying winds can alter grass texture and affect grip. Machine learning algorithms trained on historical race data can detect these trends and predict how track conditions will evolve leading up to a race.

Integrating Machine Learning for Environmental Impact Analysis

The integration of machine learning into horse racing analysis allows for a multi-variable assessment of environmental conditions. Instead of analyzing track conditions, temperature, humidity, and wind patterns in isolation, machine learning models can account for their interactions, identifying nonlinear dependencies that may not be immediately obvious through traditional analysis. For example, a combination of hot temperatures, high humidity, and a wet track might produce drastically different race outcomes than hot temperatures, low humidity, and a firm track. A tailwind of 15 mph on a synthetic track might lead to increased speed, whereas the same tailwind on turf might have minimal impact. Machine learning models trained on thousands of historical races can detect these subtle interactions and provide more precise predictions based on real-time environmental data [17,18,19].

Furthermore, real-time data collection using IoT sensors, weather APIs, and track monitoring systems can continuously update machine learning models, ensuring that race-day conditions are reflected accurately in predictive analytics [19]. As the technology evolves, the ability to optimize race strategies based on environmental conditions will become a key differentiator in horse racing performance analysis. Ultimately, a comprehensive understanding of track conditions, temperature, humidity, and wind patterns-combined with advanced machine learning techniques-provides valuable insights that can enhance race strategy, improve betting accuracy, and promote the well-being of horses in competitive racing environments [20,21,22].

Machine Learning Modeling Techniques

Machine learning models provide a sophisticated approach to predicting horse racing outcomes by analyzing environmental factors and their interactions. Unlike traditional statistical methods, machine learning algorithms can process vast amounts of historical data and detect complex, nonlinear relationships between variables such as track conditions, temperature, humidity, and wind speed. These models help trainers, jockeys, and bettors make informed decisions by offering data-driven insights into race performance. In this study, multiple machine learning algorithms were tested and evaluated for their ability to predict race outcomes based on environmental conditions. The models considered include Logistic Regression (LR), Decision Trees (DT), K-Nearest Neighbors (KNN), Naive Bayes (NB), Support Vector Machines (SVM), Random Forest (RF), and AdaBoost (AdaB). Each model was assessed based on key performance metrics, including accuracy, precision, recall, F1-score, and standard deviation, to ensure reliability and stability across various race scenarios. These models differ in their approach to handling data, and their effectiveness varies depending on the complexity of environmental influences in horse racing [19,23].

Model Selection and Training

The selection of machine learning models was based on their ability to process both categorical and numerical variables while capturing the effects of changing environmental conditions. Each model presents unique advantages and limitations when applied to horse racing predictions. The following sections provide a detailed discussion of the strengths and weaknesses of each algorithm in the context of environmental factor modeling [23,24,25,26].

Logistic Regression (LR): Logistic Regression is a simple yet effective classification algorithm widely used for predicting probabilities in binary and multi-class classification problems. In the context of horse racing, LR was employed to estimate the likelihood of a horse winning, placing, or failing to perform well based on environmental conditions such as temperature, humidity, and track surface.

- **Decision Trees (DT):** Decision Trees provide an intuitive method for modeling environmental factors in horse racing by breaking down decision-making into a series of if-then rules. The algorithm recursively splits the dataset based on the most important features, forming a tree-like structure that classifies outcomes.
- **K-Nearest Neighbors (KNN):** K-Nearest Neighbors is a distance-based classification algorithm that predicts outcomes by comparing a given race instance to its closest historical counterparts. The model classifies a new race based on the most common outcome among its k-nearest neighbors in the feature space.
- **Naive Bayes (NB):** Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem, which assumes that features are independent given the class label. While this assumption rarely holds in real-world scenarios, NB can still provide fast and reliable predictions, particularly for categorical data such as track type and race season.
- **Support Vector Machines (SVM):** Support Vector Machines are powerful classification models that maximize the margin between different classes by finding an optimal decision boundary in high-dimensional space. In horse racing, SVM is particularly useful for handling nonlinear relationships between environmental conditions and race outcomes.
- **Random Forest (RF) and AdaBoost (AdaB):** Random Forest and AdaBoost are ensemble learning techniques that combine multiple weak learners to create a more accurate and stable model. Random Forest consists of multiple Decision Trees that vote on the final prediction, reducing overfitting and improving generalization. AdaBoost, on the other hand, assigns adaptive weights to misclassified instances, improving performance iteratively.

MATERIALS AND METHODS

Data Collection and Preprocessing

This study utilized a dataset consisting of 1367 race records, compiled from various sources documenting horse racing events. The dataset contained both categorical and numerical variables, capturing key characteristics related to horse attributes, race conditions, and environmental factors. These variables included horse breed, age, gender, track type, race distance, season, climate conditions, and finishing time. The dataset was structured to enable an in-depth analysis of how environmental factors influence horse racing performance, allowing for the application of machine learning techniques to develop predictive models.

For training and evaluation, historical race data were preprocessed to remove inconsistencies, missing values, and outliers. The dataset was split into training (70%), validation (15%), and testing (15%) sets, ensuring that models were tested on unseen data to prevent overfitting. Hyperparameter tuning was conducted using grid search and cross-validation techniques to optimize model performance.

Data Cleaning and Handling Missing Values

Before applying machine learning algorithms, it was essential to clean and preprocess the dataset to ensure accuracy and consistency. The following preprocessing steps were conducted:

1. **Handling Missing Values:** Missing data were identified and addressed using imputation techniques. Numerical variables such as temperature and humidity were replaced using mean imputation, while categorical variables, including race type and season, were imputed using the mode (most frequent value). Cases where critical information was missing (e.g., incomplete race records) were removed to prevent biases in model training.
2. **Outlier Detection and Removal:** Outliers in numerical variables (e.g., extreme finishing times) were detected using the interquartile range (IQR) method. Values falling outside the

Table 1 - Feature Encoding for Categorical Variables.

Variable	Encoding Example
Horse Breed	Thoroughbred: 1, Arabian: 2
Gender	Male: 1, Female: 2
Track Type	Dirt: 1, Turf: 2, Synthetic: 3
Season	Winter: 2, Spring: 3, Summer: 1, Fall: 4
Distance	Short (<1400m): 1, Medium (1600-2000m): 2, Long (>2200m): 3

1.5×IQR range were considered anomalies and either removed or adjusted through data transformations to maintain dataset integrity.

Encoding Categorical Variables

Machine learning algorithms typically require numerical inputs; therefore, categorical variables were converted into numeric representations using label encoding and one-hot encoding where necessary. The feature encoding scheme is shown in Table 1.

Track type, race season, and horse attributes were essential categorical features that required consistent encoding to ensure the machine learning models could effectively interpret their influence on race outcomes.

Feature Scaling and Normalization

Since numerical variables (e.g., temperature, humidity, race distance) had different ranges, feature scaling was applied to standardize the dataset and improve model performance. Min-Max Scaling was used to scale values between 0 and 1, ensuring that features contributed equally to model learning. The z-score normalization method was also tested, but Min-Max Scaling yielded better predictive stability across models.

Splitting the Dataset for Model Training and Testing

To develop robust machine learning models, the dataset was divided into three subsets:

- **Training Set** (70%) - Used to train models and learn underlying patterns.
- **Validation Set** (15%) - Used to fine-tune hyperparameters and prevent overfitting.
- **Test Set** (15%) - Used to evaluate model performance on unseen data.

A stratified sampling approach was used to ensure that all environmental conditions were proportionally represented in the training, validation, and test sets.

Performance Metrics and Model Evaluation

To assess the predictive capability of different machine learning models, several performance metrics were used. These metrics provided insights into how well each model generalized to new race data and handled environmental variations.

1. **Accuracy:** Accuracy was calculated as the proportion of correctly predicted race outcomes to the total number of predictions. This metric provided an overall assessment of model performance but was supplemented with additional metrics to account for imbalances in class distributions.
2. **Precision & Recall:**
   - **Precision** measured how many of the races predicted as a certain outcome (e.g., win) were actually correct.
   - **Recall** evaluated how well the model captured actual positive cases (e.g., successfully identifying all winning horses).
3. **F1 Score:** The F1 Score is the harmonic mean of precision and recall, providing a balanced metric for cases where one measure alone might be misleading. It was particularly useful in situations where certain race conditions had fewer occurrences in the dataset.
4. **Standard Deviation:** To assess prediction stability, standard deviation was used to determine how much model performance varied across different environmental scenarios. Models with lower standard deviation values were considered more stable and reliable in real-world race predictions.

By using these evaluation metrics, the study ensured that ma-

Table 2 - Model Performance Comparison.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score	Standard Deviation	Notes
Logistic Regression (LR)	46.6	45.2	43.9	44.5	0.035	Consistent but struggles with nonlinear data
Support Vector Machine (SVM)	46.8	46.0	44.7	45.3	0.037	Best performer with balanced accuracy and stability
Decision Trees (DT)	30.8	28.3	29.5	28.9	0.053	High variance, overfits data
K-Nearest Neighbors (KNN)	33.2	31.6	32.8	32.2	0.048	Sensitive to feature scaling, struggles with large datasets
Naive Bayes (NB)	45.2	44.0	42.5	43.2	0.038	Works well with categorical data
Random Forest (RF)	41.9	40.5	41.0	40.7	0.044	Moderate accuracy, stable but computationally expensive
AdaBoost (AdaB)	46.0	45.1	44.2	44.6	0.040	Good balance of accuracy and generalization

Table 3 - Environmental Feature Importance Rankings.

Environmental Factor	Importance Score (%)	Effect on Performance
Track Condition	26.4	Wet/muddy tracks slow down horses significantly
Temperature	21.8	High heat reduces endurance and increases exhaustion risk
Humidity	18.5	High humidity impairs cooling mechanisms, leading to slower race times
Wind Speed	14.2	Headwinds reduce speed, tailwinds provide slight advantages
Race Distance	11.7	Longer distances amplify environmental impact
Season	7.4	Performance varies across seasons, with summer races being most affected by heat

chine learning models were not only accurate but also generalizable and robust under varying environmental conditions. The next stage involved hyperparameter tuning and model optimization to further enhance predictive performance.

RESULTS AND DISCUSSION

This section presents the performance evaluation of machine learning models applied to horse racing data, the impact of environmental factors, and the challenges in predictive modeling.

Model Comparisons

The effectiveness of different machine learning algorithms was measured based on accuracy, stability, and adaptability to environmental variables. Table 2 summarizes the results.

SVM achieved the highest accuracy (46.8%), demonstrating its strong performance in handling structured data with environmental variables. In contrast, Decision Trees and KNN had the lowest accuracy, indicating challenges in generalizing race conditions. Random Forest and AdaBoost delivered moderate results, but their computational complexity limited their practical application in real-time scenarios [27,28,29].

Impact of Environmental Factors

To assess the influence of environmental factors on horse racing performance, a feature importance analysis was conducted using Random Forest. Table 3 presents the top environmental factors affecting race outcomes.

Track condition was the most influential factor, emphasizing the need for adaptive race strategies based on surface conditions, as different surfaces affect a horse's traction and stability. Temperature and humidity had a combined effect, leading to significant variations in horse endurance and speed, with higher temperatures and humidity levels contributing to fatigue and reduced performance. Wind speed proved to be particularly impactful in longer races, where headwinds could reduce performance by up to 4-6%, potentially altering race outcomes [29,30,31]. These findings underline the importance of considering environmental variables when planning and preparing for races in Table 4.

Optimal race performance occurred between 10-16°C, where horses demonstrated their best speed and endurance [32]. As temperatures rose above 27°C, race performance declined by 6-10%, with horses experiencing increased fatigue and reduced efficiency. In extreme heat, particularly above 32°C, horses struggled significantly, as the high temperatures caused stress, dehydration, and fatigue. In such conditions, horses required proper hydration and tailored pacing strategies to maintain performance and avoid overheating [33,34,35]. These findings highlight the importance of managing environmental factors, especially temperature, to ensure that horses can perform at their peak during races in Table 5.

Headwinds above 16 km/h (10 mph) significantly reduced horse speeds, emphasizing the importance of factoring in wind conditions when evaluating race performance [36]. Strong headwinds create additional resistance, forcing horses to expend more energy to maintain their pace, ultimately slowing them down. This highlights the need for trainers, bettors, and race organizers to carefully consider wind conditions when planning and preparing for races, as they can have a substantial impact on race outcomes [37,38].

The application of machine learning models in predicting horse racing outcomes has shown promising results, with various algorithms being tested for their effectiveness. **Support Vector Machines (SVM)** have been noted for their strong performance in handling structured data with environmental variables [39]. In contrast, **Decision Trees (DT)** and **K-Nearest Neighbors (KNN)** have struggled with lower accuracy due to high variance and sensitivity to feature scaling. Gupta & Singh [19] demonstrate how Random Forest and other algorithms can be used to provide very high accuracy in their research with data from the Turf Club of India. While code repositories exist for horse race prediction, their reliability can vary [40]. The impact of environmental factors such as track condition, temperature, and humidity is also crucial, as they significantly affect race outcomes. The legal use of machine learning for in-

Table 4 - Wind Speed and Performance Correlation.

Wind Speed (mph)	Race Speed Reduction (%)
0-5	0.5%
6-10	2.3%
11-15	4.2%
16+	6.5%

Table 5 - Effect of Temperature on Race Times.

Temperature (\u00b0F)	Average Race Time (s)	Performance Change (%)
50-60	108.2	Baseline (optimal)
61-70	109.6	-1.3% slower
71-80	112.3	-3.8% slower
81-90	115.1	-6.3% slower
91+	118.7	-9.7% slower

Table 6 - Comparison of different learning models by their advantages and disadvantages.

Model	Advantages	Limitations
Logistic Regression (LR) and multicollinearity	Interpretability, computational efficiency	Assumes linear relationships, struggles with outliers
Decision Trees (DT)	Handles numerical and categorical data, easy interpretation	Tends to overfit, requires pruning or ensemble methods
K-Nearest Neighbors (KNN)	Simplicity, captures localized trends	Computationally expensive, sensitive to feature scaling
Naive Bayes (NB)	Efficient with small datasets, works well with imbalanced data	Independence assumption limits interaction capture
Support Vector Machines (SVM)	Captures complex decision boundaries, robust to outliers	Computationally expensive, sensitive to kernel choice
Random Forest (RF)	Excels with high-dimensional datasets, robust	Requires more computational resources, less interpretable
AdaBoost	Strong classification performance with fewer trees	Requires more computational resources, less interpretable

formed betting is generally permissible LegalAdviceUK [40], though specific regulations may vary by jurisdiction. Overall, machine learning offers a robust approach to predicting horse racing outcomes, but challenges remain in handling complex data and environmental variables [41].

Logistic Regression (LR) is interpretable and computationally efficient, making it ideal for real-time predictions, but it assumes linear relationships and struggles with outliers and multicollinearity. Decision Trees handle both numerical and categorical data well and are easy to interpret, though they tend to overfit, requiring pruning or ensemble methods [42]. K-Nearest Neighbors (KNN) is simple and captures localized trends but is computationally expensive and sensitive to feature scaling. Naive Bayes is efficient for small and imbalanced datasets but struggles with interactions between variables [43]. Support Vector Machines (SVM) capture complex decision boundaries and are robust to outliers, but they are computationally expensive and sensitive to kernel selection. Random Forest and AdaBoost perform well with high-dimensional datasets, though they require more resources and are less interpretable [44,45]. While simpler models like LR and Naive Bayes are fast and interpretable, advanced models like SVM, Random Forest, and AdaBoost provide better predictive performance. Future improvements could involve integrating real-time weather data and optimizing hyperparameters [46, 47,48,49,50]. The comparison of methods is given in Table 6.

Conclusion

The findings of this study highlight the critical role that environmental factors play in horse racing performance and the potential for machine learning to enhance predictive capabilities in this domain. By analyzing race data with various machine learning models, we observed that track conditions, temperature, humidity, and wind speed significantly influence race outcomes. The Support Vector Machine (SVM) algorithm performed best in terms of predictive accuracy and stability, underscoring the effectiveness of nonlinear classifiers in modeling complex interactions between environmental factors and racing performance. However, while machine learning provides valuable insights, its implementation in real-world horse racing scenarios still requires further refinement and integration with real-time data streams.

One of the key takeaways from this research is the importance of adapting race strategies based on environmental conditions. The strong influence of track conditions suggests that different horses may perform optimally under specific surface types, making it essential for trainers and jockeys to tailor their preparation accordingly. Similarly, the impact of extreme temperatures on performance emphasizes the need for hydration strategies, pacing adjustments, and possible schedule modifications for races held in hotter climates. Given that wind resistance can have a measurable effect on speed, especially in longer races, understanding wind patterns and their implications should be factored into race-day decisions to improve competitiveness.

Despite the promising results obtained through machine learning models, several challenges remain that must be addressed in future research. One of the most significant limitations is the variability and unpredictability of environmental conditions, which makes it difficult to create a fully generalizable model. The models developed in this study relied on historical data, which means they might not always be adaptable to sudden shifts in weather patterns or track conditions. To overcome this, future studies should integrate live weather feeds and real-time track updates into predictive frameworks, allowing for dynamic adjustments in model predictions. Additionally, more sophisticated ensemble learning techniques could be explored to improve predictive accuracy and stability across different racing environments.

Another limitation of the study is the exclusion of certain non-environmental factors that could significantly influence race outcomes. Variables such as jockey experience, horse training regimens, injury history, and genetic traits play an essential role in performance but were not fully incorporated into the modeling process. Future research should focus on combining environmental and physiological data to create more comprehensive predictive models. For instance, integrating biometric data, such as heart rate variability and stride efficiency, could provide deeper insights into how horses respond to environmental stressors and fatigue levels during races. Additionally, incorporating machine learning techniques such as reinforcement learning could allow models to continuously adapt and refine predictions based on live race conditions.

The implications of this study extend beyond predictive analytics and betting strategies; they also offer valuable insights for improving racehorse welfare and ensuring the sustainability of horse racing as an industry. With climate change expected to bring more extreme weather conditions, it is imperative that race organizers implement adaptive measures to protect both horses and jockeys. This could include modifying race schedules based on temperature forecasts, improving track drainage systems to mitigate the effects of heavy rainfall, and investing in alternative track materials that minimize injury risks. Machine learning models could also assist regulatory bodies in monitoring race conditions and making data-driven decisions regarding race safety protocols.

The future of machine learning in horse racing is promising, but it requires collaboration between researchers, trainers, veterinarians, and industry stakeholders to maximize its potential. By leveraging advanced data analytics, the industry can transition toward a more scientific and evidence-based approach to training, race strategy, and risk assessment. As models become more sophisticated and real-time data integration improves, the predictive accuracy of race outcomes will likely increase, benefiting trainers, bettors, and horse welfare advocates alike. Continued research and innovation in this space will not only enhance competitive strategies but also ensure the longevity and ethical sustainability of the sport in an increasingly unpredictable environmental landscape.

Acknowledgement

This study was undertaken during the Pamukkale University DOSAP research.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Conflict of Interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by [Yavuzkan Paksoy], [Ahmet Koluman], [U ur Duruk] and [Ahmet Akay]. The first draft of the manuscript was written by [full name] and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data Availability

The datasets generated during and/or analysed during the current study are not publicly available due to [REASON(S) WHY DATA ARE NOT PUBLIC] but are available from the corresponding author on reasonable request.].

Ethical Approval

In this study, the routine breeding practices were not excluded. Obtained in the study data collected within the scope of these applications for Experimental and Other Scientific Purposes On the Welfare and Protection of Animals Used Regulation (Official Gazette dated 13.12.2011 and numbered 28141 Newspaper) Article 2, second paragraph "This Regulation, non-experimental agricultural and clinical veterinary applications", the scope of the scope does not cover Ministry (T.C. Adana Governorship Provincial Directorate of Agriculture and Forestry, Issue: E-74530962-325.99-14052301) is not subject to authorization.

References

1. Nielsen, B.D., Turner, K.K., Ventura, B.A., et al. 2006. dRacing speeds of quarter horses, thoroughbreds and Arabians. Equine veterinary journal. Supplement, 36, 128-132. DOI: 10.1111/j.2042-3306.2006.tb05528.x.
2. Paksoy, Y. and Ünal, N. 2010. The factors affecting horse racing performance (A review) [Turkish]. Livestock studies, 50 (2): 91-101.
3. Aksüyek, B. and Arpacık, R. 1991. The development of racing performances of Arabian Horses and Factors affecting these [Turkish]. Livestock Studies, 31(3-4): 55-66.
4. Harkins, J.D., Morris, G.D., Tulley, R.T., et al. 1992. Effect of added dietary fat on racing performance in thoroughbred horses. J Equine Vet Sci, 12(2): 123-129. DOI: 10.1016/S0737-0806(06)81295-5.
5. Cihan, P. 2024. Horse Surgery and Survival Prediction with Artificial Intelligence Models: Performance Comparison of Original, Imputed, Balanced, and FeatureSelected Datasets [Turkish]. Kafkas Univ Vet Fak Derg, 30(2): 233-241. DOI: 10.9775/kvfd.2023.30908
6. Yıldırım, F. 2019. The Effect of Some Environmental Factors to Race Performance of Purebred Arabian Horses [Turkish]. Kocatepe Vet J, 12(1): 1-6. doi.org/10.30607/kvj.450350
7. Bartolomé, E. and Stanley Cockram, M. 2016. Potential Effects of Stress on the Performance of Sport Horses. Journal of Equine Veterinary Science, 40, 84-93. DOI: 10.1016/j.jevs.2016.01.016.
8. Demirtaş, B. 2017. Body temperature and thermoregulation mechanism in exercise in horses [Turkish]. In book: Sport Horses, 85-99.
9. Gottlieb-Vedi, M. and Lindholm A. 1997. Comparison of standardbred trotters exercising on a treadmill and a race track with identical draught resistances. Veterinary Record, 140, 525-528. DOI: 10.1136/vr.140.20.525.
10. Guthrie, A.J. and Lund, R.J. 1998. Thermoregulation: Base Mechanisms and Hyperthermia. Veterinary Clinics of North America: Equine Practice, 14(1): 45-59. DOI: 10.1016/S0749-0739(17)30211-0.
11. Hutchinson, J. and Schwerha, D. 2024. Machine Learning Analysis of Horse-Rider Dyad Performance in Show Jumping. Systems and Information Engineering Design Symposium (SIEDS), 124-129. DOI: 10.1109/SIEDS61124.2024.10534746.
12. Williams, R.J., Marlin, D.J., Smith, N., et al. 2002. Effects of Cool and Hot Humid Environmental Conditions on Neuroendocrine Responses of Horses to Treadmill Exercise. The Veterinary Journal, 164(1): 54-63. DOI: 10.1053/tvjl.2002.0721.
13. Evangeline, S.I. and Rathika, P. 2022. Wind farm incorporated optimal power flow solutions through multi-objective horse herd optimization with a novel constraint handling technique. Expert Systems with Applications, 194, 116544. DOI: 10.1016/j.eswa.2022.116544.
14. Keller, G.A., Nielsen, B.D., Vergara-Hernandez, F.B., et al. 2022. Tracking the Impact of Weather on Equine Activity While Pastured. Journal of Equine Veterinary Science, 116, 104052. DOI: 10.1016/j.jevs.2022.104052.
15. Kaseda, Y. 1983. Seasonal Changes in Time Spent Grazing and Resting of Misaki Horses. Nihon Chikusan Gakkaiho, 54, 464-469. DOI: 10.2508/chikusan.54.464.
16. Autio, E. and Heiskanen, M.L. 2005. Foal behaviour in a loose housing/paddock environment during winter. Applied Animal Behaviour Science, 91(3-4): 277-288. DOI: 10.1016/j.applanim.2004.10.012.
17. Mejdell, C.M. and Bøe, K.E. 2005. Responses to climatic variables of horses housed outdoors under Nordic winter conditions. Canadian Journal of Animal Science, 85, 301-308. DOI: 10.4141/A04-066.
18. Hino, M., Benami, E. and Brooks, N. 2018. Machine learning for environmental monitoring. Nature Sustainability, 1, 583-588. DOI: 10.1038/s41893-018-0142-9.
19. Bunker, R. and Thabtah, F. 2019. A machine learning framework for sport result prediction. Applied Computing and Informatics, DOI: 10.1016/j.aci.2017.09.005.
20. Colle, P. 2022. What AI can do for horse-racing. Computers and Society, 7, 5. DOI: 10.48550/arXiv.2207.04981.
21. Gupta, M. and Singh, L. 2023. Predicting Outcomes of Horse Racing using Machine Learning. International Journal on Recent and Innovation Trends in Computing and Communication. DOI: 10.17762/ijritcc.v11i9.8119.
22. Kapania, N. and Gerdes, C. 2020. Learning at the Racetrack: Data-Driven Methods to Improve Racing Performance Over Multiple Laps. IEEE

Transactions on Vehicular Technology, 69, 8232-8242. DOI: 10.1109/TVT.2020.2998065.

23. Rajan, T., Chavan, G., Kumar, R., et al. 2023. Leveraging Machine Learning for More Efficient Real-Time Data Analysis. 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (Smart Gencon), 1-7. DOI: 10.1109/SMARTGENCON60755.2023.10442368.
24. Maalouf, M. 2011. Logistic regression in data analysis: an overview. Int. J. Data Anal. Tech. Strateg., 3, 281-299. DOI: 10.1504/IJDATS.2011.041335.
25. Liu, D., Li, T. and Liang, D. 2014. Incorporating logistic regression to decision-theoretic rough sets for classifications. Int. J. Approx. Reason., 55, 197-210. DOI: 10.1016/j.ijar.2013.02.013.
26. Kravets, P., Pasichnyk, V. and Prodaniuk, M. 2024. Mathematical Model of Logistic Regression for Binary Classification. Part 1. Regression Models of Data Generalization. Vìsnik Nacìonal nogo unìversitetu «Lʹvìvsʹka polìtehnìka». Serìâ Ìnformacìjnì sistemi ta merežì.
27. Ghorbani, R and Ghousi, R. 2020. Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques. IEEE Access, 8, 67899-67911. DOI: 10.1109/ACCESS.2020.2986809.
28. Nie, F., Zhu, W. and Li, X. 2020. Decision Tree SVM: An extension of linear SVM for non-linear classification. Neurocomputing, 401, 153-159. DOI: 10.1016/j.neucom.2019.10.051.
29. Martin, G., Strand, E. and Kearney, M. 1996. Use of statistical models to evaluate racing performance in thoroughbreds. Journal of the American Veterinary Medical Association, 209, 11: 1900-6. DOI: 10.2460/javma.1996.209.11.1900.
30. Cho, K., Son, S., Cho, B., et al. 2008. Effects of the Condition of Racetrack on Racing Time in Thoroughbred Racehorses. Journal of Animal Science and Technology, 50, 309-314. DOI: 10.5187/JAST.2008.50.3.309.
31. Gladkikh, M. and Kuznetsova, O. 2021. The effects of race track surface condition on the racing time of Orlov trotters. E3S Web of Conferences.
32. Negro, S., Bartolomé, E., Molina, A., et al. 2018. Stress level effects on sport performance during trotting races in Spanish Trotter Horses. Research in veterinary science, 118, 86-90. DOI: 10.1016/j.rvsc.2018.01.017.
33. Gómez, M., Menéndez-Buxadera, A., Valera, M., et al. 2015. A reaction norm model approach to estimate the genetic effect of temperature on sportive performance of trotter horses. Journal of animal breeding and genetics = Zeitschrift fur Tierzuchtung und Zuchtungsbiologie, 132, 3: 256-67. DOI: 10.1111/jbg.12118.
34. Soroko, M., Dudek, K., Howell, K., et al. 2014. Thermographic Evaluation of Racehorse Performance. Journal of Equine Veterinary Science, 34, 1076-1083. DOI: 10.1016/J.JEVS.2014.06.009.
35. Witkowska-Piłaszewicz, O., Ma ko, M., Domino, M., et al. 2020. Infrared Thermography Correlates with Lactate Concentration in Blood during Race Training in Horses. Animals: an Open Access Journal from MDPI, 10. DOI: 10.3390/ani10112072.
36. Chen, Z., Liu, T., Jiang, Z., et al. 2018. Comparative analysis of the effect of different nose lengths on train aerodynamic performance under crosswind. Journal of Fluids and Structures, 78, 69-85. DOI: 10.1016/J.JFLUIDSTRUCTS.2017.12.016.
37. Atkinson, G. and Brunskill, A. 2000. Pacing strategies during a cycling time trial with simulated headwinds and tailwinds. Ergonomics, 43, 1449-1460. DOI: 10.1080/001401300750003899.
38. International Journal of Intelligent Systems and Applications in Engineering, Horse Race Results Prediction Using Machine Learning Algorithms with Feature Selection. International Journal of Intelligent Systems and Applications in Engineering, 2023; 12(2s): 132-139.
39. Ethan-eplee, HorseRacePrediction (Version 1.0) Computersoftware-Computer softwareComputersoftware, https://github.com/ethan-eplee/HorseRacePrediction, 2024.
40. LegalAdviceUK, Machine learning prediction for horse races, Reddit. https://www.reddit.com/r/LegalAdviceUK/comments/1087p4p/machine_learning_prediction_for_horse_races/ , Accessed: 10.05.2023.
41. Code Works, Use case 3: Horse racing prediction - A machine learning approach (Part 1), 2024.
42. Pranckevičius, T. and Marcinkevičius, V. 2017. Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. Balt. J. Mod. Comput, 5. DOI: 10.22364/BJMC.2017.5.2.05.
43. Manorom, P., Detthamrong, U. and Chansanam, W. 2024. Comparative Assessment of Fraudulent Financial Transactions using the Machine Learning Algorithms Decision Tree, Logistic Regression, Naïve Bayes, K-Nearest Neighbor, and Random Forest. Engineering, Technology & Applied Science Research, DOI: 10.48084/etasr.7774.
44. Dumitrescu, E., Hué, S., Hurlin, C., et al. 2021. Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. Eur. J. Oper. Res. 297, 1178-1192. DOI: 10.1016/j.ejor.2021.06.053.
45. Saberioon, M., Císar, P., Labbé, L., et al. 2018. Comparative Performance Analysis of Support Vector Machine, Random Forest, Logistic Regression and k-Nearest Neighbours in Rainbow Trout (Oncorhynchus Mykiss) Classification Using Image-Based Features. Sensors (Basel, Switzerland), 18. DOI: 10.3390/s18041027.
46. Peerbashab, S., Iqbal, M.K.P., Surputheen, M., et al. 2023. Diabetes Prediction using Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors, Logistic Regression Classifiers. Journal of Advanced Applied Scientific Research, DOI: 10.46947/joaasr542023680.
47. Araneda, O. 2022. Horse Racing as a Model to Study the Relationship between Air Pollutants and Physical Performance. Animals: an Open Access Journal from MDPI, 12. DOI: 10.3390/ani12091139.
48. Bishop, C. 2013. Model-based machine learning. Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 371. DOI: 10.1098/rsta.2012.0222.
49. Hodgson, D.R., Davis, R.E. and McConaghy, F.F. 1994. Thermoregulation in the horse in response to exercise. British Veterinary Journal, 150(3): 219-235. DOI: 10.1016/S0007-1935(05)80003-X.
50. Li, X., Wang, L. and Sung, E. 2008. AdaBoost with SVM-based component classifiers. Eng. Appl. Artif. Intell., 21, 785-795. DOI: 10.1016/j.engappai.2007.07.001. doi.org/10.30607/kvj.450350.