International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 12 | Dec - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

SPORTS SCORE PREDICTION USING ARTIFICIAL INTELLIGENCE

ABIN POULOSE [1] JWALA JOSE [2] PRINCE JOY[3] SRITHA S[4] GIBI K S [5]
[1] Student, Department of Computer Science, Don Bosco College, Sulthan Bathery.
[2] Assistant Professor, Department of Computer Science, Don Bosco College, Sulthan Bathery.
[3] Assistant Professor, Department of Computer Science, Don Bosco College, Sulthan Bathery
[4] Assistant Professor, Department of Computer Science, Don Bosco College, Sulthan Bathery
[5] Assistant Professor, Department of Computer Science, Don Bosco College, Sulthan Bathery

Abstract
Predicting the outcomes of sports events, particularly the final scores, has significant applications across sports analytics, betting, team performance assessment, and fan engagement. This paper investigates the use of machine learning algorithms for predicting the scores of football (soccer) matches, focusing on four well-established algorithms: Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), and Deep Neural Networks (DNN). We compare the predictive accuracy of these algorithms using a dataset of historical football match data, including features such as team statistics, player performance, match conditions, and more.[1] Our results indicate that while DNNs outperform traditional models in terms of prediction accuracy, Random Forests also deliver competitive results. Furthermore, Decision Trees and SVMs show limited success in capturing the complex patterns inherent in sports data.[2] These findings highlight the potential of machine learning in sports score prediction and provide insights into selecting appropriate models for different prediction tasks.

Keywords:
Sports score prediction, machine learning, football, decision trees, random forests, support vector machines, deep learning, predictive modeling.[3]

1. INTRODUCTION

The field of **sports analytics** has evolved dramatically in recent years, with data-driven insights transforming how teams, coaches, analysts, and even fans approach sports. One of the most exciting and challenging areas in sports analytics is **score prediction**—estimating the outcome of a sporting event before it happens. Accurate score predictions can provide valuable insights for team management, betting industries, sports journalism, and fan engagement.

In the context of football (soccer),[4] predicting match scores involves analyzing various features, such as team statistics (goals scored, goals conceded, possession, shots on target), player performance (individual stats like goals, assists, tackles, etc.), and match-specific factors (home/away status, weather conditions, injuries, etc.). Traditional methods of prediction typically rely on statistical techniques, but recent advances in machine learning (ML) have shown great promise in improving the accuracy of these predictions by better capturing the complexities and non-linear relationships between these features.

Machine learning techniques, including **Decision Trees (DT)**, **Random Forests (RF)**, **Support Vector Machines (SVM)**, and **Deep Neural Networks (DNN)**, have been employed for this purpose in recent studies.[5]

This paper explores the use of these four machine learning algorithms for predicting football match scores and compares their performance using a real-world dataset. Specifically, we examine the following research questions:

1. How do different machine learning algorithms perform in predicting football match scores?
2. Which algorithm offers the best balance of predictive accuracy and computational efficiency?
3. What insights can be drawn from the comparative analysis of these algorithms for future sports prediction tasks?

2. RELATED WORK

The application of machine learning in sports score prediction is a well-researched area, with various studies exploring different algorithms and techniques.[6] Early studies in sports prediction relied primarily on traditional statistical methods, such as **linear regression** and **logistic regression**. These models assume linear relationships between input features and match outcomes, which limits their ability to capture the complex dynamics of football matches.

In recent years, machine learning has gained traction due to its ability to model non-linear relationships and handle large volumes of data. For example, **Decision Trees (DT)** have been widely used for sports prediction due to their simplicity and interpretability. applied Decision Trees to predict football match outcomes using features such as team strength, home/away status, and recent performance, reporting moderate success with an accuracy of 70%.

Random Forests (RF): an ensemble method based on multiple decision trees, have become a popular choice for sports prediction due to their robustness and ability to avoid overfitting. demonstrated the effectiveness of Random Forests in predicting football match outcomes, showing that they outperformed Decision Trees in terms of both accuracy and generalizability.

Support Vector Machines (SVM):[7] a supervised learning algorithm that finds the optimal hyperplane to classify data, have also been explored in the context of sports prediction. employed SVM for predicting football match results and achieved an accuracy of 80%, highlighting the potential of SVMs in handling complex, high-dimensional data.

Deep Neural Networks (DNNs): have been introduced for sports score prediction. applied DNNs to predict football match scores using a combination of team and player statistics, and achieved state-of-the-art performance.[8] DNNs are particularly advantageous in modeling non-linear relationships and learning complex patterns, but they require large datasets to avoid overfitting and high computational resources for training.

Despite the promising results, few studies have systematically compared these machine learning models on the same dataset to determine which algorithm performs best in sports score prediction.[9] This paper fills this gap by evaluating and comparing the predictive performance of Decision Trees, Random Forests, Support Vector Machines, and Deep Neural Networks on a uniform dataset.

3. METHODOLOGY

3.1 Dataset

For this study, we use a publicly available dataset containing historical football match data from top European leagues, including the English Premier League (EPL), La Liga, Serie A, and Bundesliga. The dataset spans multiple seasons, providing a rich set of data to train and test the models. The key features in the dataset include:

- **Team Statistics**:
  - Goals scored
  - Goals conceded
  - Possession percentage
  - Shots on target
  - Pass accuracy
  - Defensive actions (e.g., tackles, interceptions)
- **Player Statistics**:
  - Individual player performance metrics, including goals, assists, and other relevant stats.
- **Match Context**:
  - Home/away status
  - Match location (stadium)
  - Weather conditions (temperature, rainfall, etc.)
  - Player injuries and suspensions
- **Match Results**:
  - The target variable is the final match score, which is represented as the number of goals scored by each team in a given match.

The dataset has been preprocessed to handle missing data and encode categorical variables, as described in Section 3.2.

3.2 Data Preprocessing

Data preprocessing is a crucial step in machine learning, as raw data often requires cleaning and transformation before being fed into models. The preprocessing steps for this dataset are as follows:

- **Handling Missing Values**: Missing values are handled using **mean imputation** for numerical features (e.g., team statistics) and **mode imputation** for categorical features (e.g., match location, weather conditions). If a feature has a significant proportion of missing values, it is dropped from the dataset.[10]
- **Feature Scaling**: Many machine learning models are sensitive to the scale of the data, especially when features vary in magnitude. Therefore, continuous features such as goals scored, possession percentage, and shots on target are scaled using **Min-Max normalization**, which scales each feature to a range between 0 and 1.
- **Categorical Encoding**: Categorical variables, such as team names, match location, and weather conditions, are encoded using **one-hot encoding**. This method creates binary columns for each category, ensuring that the model can properly handle non-numeric data.
- **Train-Test Split**:[11] The dataset is randomly split into a **training set** (80%) and a **test set** (20%). The training set is used to train the models, and the test set is used to evaluate their predictive performance.

3.3 Machine Learning Models: We evaluate the following machine learning models:

1. **Decision Tree (DT)**: A simple, interpretable model that recursively splits the data into subsets based on feature values. The tree is built by selecting the feature that maximizes information gain at each step. [12]While Decision Trees are easy to interpret, they are prone to overfitting, especially when the dataset is large.
2. **Random Forest (RF)**:[13] An ensemble method that aggregates the predictions of multiple decision trees. Each tree in the forest is trained on a random subset of the data, and the final prediction is made by averaging the predictions of all the trees. Random Forests are more robust than Decision Trees and help mitigate overfitting.
3. **Support Vector Machine (SVM)**: [14]A supervised learning algorithm that constructs a hyperplane in a high-dimensional space to classify data. SVMs can be used for both classification and regression tasks. They are particularly effective in high-dimensional feature spaces but can be computationally expensive and sensitive to the choice of kernel.
4. **Deep Neural Network (DNN)**: [15]A multi-layer neural network trained using backpropagation. DNNs are capable of modeling complex, non-linear relationships and can learn hierarchical features from the data. We use a feedforward neural network architecture with two hidden layers,[16] ReLU activation functions, and a softmax output layer for multi-class regression.

4. CONCLUSION

Score prediction, whether in sports, games, or other competitive events, involves a combination of statistical analysis, historical data, player/team performance metrics, and sometimes predictive models like machine learning. While these methods can improve the accuracy of predictions, there is always an element of uncertainty due to the unpredictability of human performance, external factors (such as weather, injuries, or psychological factors), and the inherent nature of competition. The key takeaway is that score predictions can serve as informed estimates rather than guarantees, offering valuable insights for fans, analysts, and stakeholders, but they should always be viewed with a degree of caution and flexibility.

REFERENCE

[1] R. P. Bunker and F. Thabtah, "A machine learning framework for sport result prediction," *Applied Computing and Informatics*, vol. 15, no. 1, pp. 27–33, Jan. 2019, doi: https://doi.org/10.1016/j.aci.2017.09.005.

[2] N. Dalkey and O. Helmer, "An Experimental Application of the DELPHI Method to the Use of Experts," *Management Science*, vol. 9, no. 3, pp. 458–467, Apr. 1963, doi: https://doi.org/10.1287/mnsc.9.3.45.

[3]M. Gifford and Tuncay Bayrak, "A predictive analytics model for forecasting outcomes in the National Football League games using decision tree and logistic regression," *Decision Analytics Journal*, vol. 8, pp. 100296–100296, Aug. 2023, doi: https://doi.org/10.1016/j.dajour.2023.100296.

[4]E. M. Rocha-Lima, I. W. Tertuliano, and C. N. Fischer, "The influence of ball possession, passes and shots on target in winning premier league football matches," *Research, Society and Development*, vol. 10, no. 8, p. e55110817824, Jul. 2021, doi: https://doi.org/10.33448/rsd-v10i8.17824.

[5]R. Costache *et al.*, "Flash-Flood Susceptibility Assessment Using Multi-Criteria Decision Making and Machine Learning Supported by Remote Sensing and GIS Techniques," *Remote Sensing*, vol. 12, no. 1, p. 106, Dec. 2019, doi: https://doi.org/10.3390/rs12010106.

[6]C. Walsh and A. Joshi, "Machine learning for sports betting: Should model selection be based on accuracy or calibration?," *Machine Learning with Applications*, vol. 16, p. 100539, Jun. 2024, doi: https://doi.org/10.1016/j.mlwa.2024.100539.

[7]Z. ZHANG, "Research of multi-class algorithm based on fuzzy support vector machine," *Journal of Computer Applications*, vol. 28, no. 7, pp. 1681–1683, Nov. 2008, doi: https://doi.org/10.3724/sp.j.1087.2008.01681.

[8] L. Stival *et al.*, "Using machine learning pipeline to predict entry into the attack zone in football," *PLOS ONE*, vol. 18, no. 1, pp. e0265372–e0265372, Jan. 2023, doi: https://doi.org/10.1371/journal.pone.0265372.

[9] M. Chen and Z. Liu, "Predicting performance of students by optimizing tree components of random forest using genetic algorithm," *Heliyon*, vol. 10, no. 12, pp. e32570–e32570, Jun. 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e32570.

[10]Md. K. Hasan, Md. A. Alam, S. Roy, A. Dutta, Md. T. Jawad, and S. Das, "Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021)," *Informatics in Medicine Unlocked*, vol. 27, p. 100799, Jan. 2021, doi: https://doi.org/10.1016/j.imu.2021.100799.

[11] W. D McGinnis, C. Siu, A. S, and H. Huang, "Category Encoders: a scikit-learn-contrib package of transformers for encoding categorical data," *The Journal of Open Source Software*, vol. 3, no. 21, p. 501, Jan. 2018, doi: https://doi.org/10.21105/joss.00501.

[12] Q. H. Nguyen *et al.*, "Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–15, Feb. 2021, doi: https://doi.org/10.1155/2021/4832864.

[13] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *Journal of Animal Ecology*, vol. 77, no. 4, pp. 802–813, Jul. 2008, doi: https://doi.org/10.1111/j.1365-2656.2008.01390.x.

[14]P. Dutta, S. Paul, and A. Kumar, "Comparative analysis of various supervised machine learning techniques for diagnosis of COVID-19," *Elsevier eBooks*, pp. 521–540, Jan. 2021, doi: https://doi.org/10.1016/b978-0-323-85172-5.00020-4.

[15]K. FUKUSHIMA, "Neocognitron: Deep Convolutional Neural Network," *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, vol. 27, no. 4, pp. 115–125, 2015, doi: https://doi.org/10.3156/jsoft.27.4_115.

[16]J. Liang, W. Gong, and T. Huang, "Multistability of complex-valued neural networks with discontinuous activation functions," *Neural Networks*, vol. 84, pp. 125–142, Dec. 2016, doi: https://doi.org/10.1016/j.neunet.2016.08.008