ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 8, August 2024

A Comparative Analysis of Machine Learning Algorithms for Basketball Winning Team Prediction

[1] Babitha Ganesh, [2] Janardhana Bhat K, [3] Sanketh K H

[1] [2] [3] Department of Computer Science and Engineering, Canara Engineering College, Mangalore, Karnataka, INDIA
Corresponding Author Email: [1] babitha.ganesh@canaraengineering.in, [2] janardhanabhatk@gmail.com, [3] sankethkalige369@gmail.com

Abstract*— For several reasons, including team and player growth as well as coach and sports expert decision-making, sports prediction is vital. The main objective of this article is to apply machine learning (ML) techniques to build a data-driven model that can forecast the results of NBA league games. The work done in this article starts with data processing and continues with model construction using five different ML models including Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF) and Artificial Neural Networks(ANN). These five models were then evaluated using different metrices such as accuracy, precision, recall and F1-score. This study's systematic approach offers a flexible framework that may be used in a range of sports analytics contexts. Based on past extended index averages, the model projects symmetric extended indices for both teams playing in upcoming games. Testing and training sets from multiple seasons are used to evaluate the suggested model. The variables pertaining to teams, players, and opponents are the main emphasis of this research. These variables include field goal attempts, three-point attempts, made free throws, attempted free throws, and offensive rebounds, among others. The algorithms performed differently, as demonstrated by the results, with the RF Algorithm coming out on top with 84% accuracy. This shows how well RF predicts the likelihood of a winning team, outperforming SVM (74%), LR (82%), ANN (82%), KNN (79%). The proposed is intended to help the coaches to select the right set of team members to improve the chance of winning in the tournament. This work is still limited in terms of data quality and model restrictions. Nonetheless, sports professionals looking for practical insights should immediately consider the ramifications of the findings. This work points to future directions for research, encouraging the creation of more intricate algorithms, detailed feature analysis, and the integration of temporal patterns for comprehensive predictive accuracy.

Index Terms*— Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), National Basketball Association (NBA), Predictive Analysis, Machine Learning, Sports Analytics.

I. INTRODUCTION

The most appealing aspect of competitive sports, which draw a lot of interest and popularity, is their unpredictable nature. The outcomes of the high-intensity, quick-paced games might alter, frequently in a matter of minutes or even seconds, and a variety of factors influence the victors and losers. The study of sports competition analysis has grown in importance in recent years. Owners of individual teams invest thousands of dollars in their teams every year in the hopes of avoiding spending money on players who don't improve the squad. Fans want to win game bets all season long, and analysts spend a lot of effort attempting to forecast which team will win the championship each year, therefore making bigger financial advantages commercially[1].

A growing number of studies have concentrated on predicting basketball game results using information gathered from leagues all over the world. Depending on what the method focuses on and how the predictions are provided, these forecasting techniques can be separated into two categories: generative methods and discriminant methods. Generative models are useful for a variety of tasks, such as generation and classification, and they focus on comprehending the underlying data distribution. Discriminative models, on the other hand, are mostly applied to classification tasks and learn the decision border between classes directly. The particular task at hand, the data that are accessible, and the available computing power will determine which of these two methods is best. [2]. Use of AI algorithms have gained more popularity these days due to the fact that it produces better results compared to other existing techniques.

In order to identify the significant feature set that influences NBA game outcomes, the study carried out in [3] suggests a novel intelligent machine learning framework for game outcome prediction. Through this, the scientists were able to determine what important elements influence game outcomes and how historical data from prior games played can be used to estimate the outcome of an NBA game using machine learning techniques. Various machine learning algorithms were utilized, such as Naïve Bayes, artificial neural networks, and decision trees. Through a comparison of the models' performance against various sets of basketball-related variables, it was shown that the prediction model's efficiency and accuracy determine which approach is most appropriate.

Different player stereotypes have been established by the authors in [4], who are entirely unaffected by how players are categorized in relation to the five traditional basketball positions. Through the identification of various playing styles in the league, they were able to ascertain the effectiveness of specific team configurations in creating favorable synergies

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 8, August 2024

that result in increased winnings. The optimal input space for predictors has been created by the application of clustering algorithms.

The multi-layer perception model used in [5] was able to predict the results of NBA game. The purpose of [6] is to develop a solid yet straightforward model that uses past game outcomes to estimate the point differential for each game. It also uses data from five seasons to calculate team ability and individual home advantages. When compared to the bookmaker's point spread, the model's analysis of games played between 2012 and 2016 yields accurate forecasts based on the evaluation of several important variables, including home advantage and team ability.

In [7], deep learning is also used for prediction in addition to regular machine learning. However, for their dataset, Deep Learning performed less well than standard Machine Learning. It makes sense that Deep Learning's capacity to handle a sizable volume of Big Data was constrained when their data were relatively modest and organized with a small number of predictor factors. The final findings from the Regression and Classification Analysis showed that, for any team, scoring is the most crucial component from the starting players as well as the preferred style of play for basketball..

This article demonstrates the results of applying different ML models such as LR, SVM, KNN, RF and ANN on the NBA dataset with the statistics of the year 2010 to 2023 available in Kaggle.

This article is divided into the following sections. Section II presents the methodology followed in the work. Section III presents the results and discussion. Section V concludes the article and gives a brief on the future scope of work.

II. METHODOLOGY

This section of the article explains the data sources and methodology used to predict NBA results with various machine learning models. Figure 1 depicts the general process for predicting results. It consists of data collection and interpretation, pre-processing, feature selection, the application of ML/DL models, and model evaluation. This section provides a detailed description of these steps.

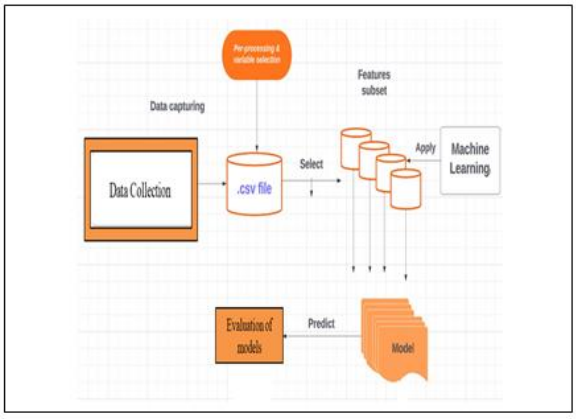

Fig. 1. Methodology followed in prediction of NBA results

A. Data Collection and Interpretation

The first step in creating a strong NBA statistics-driven prediction model is data collecting. To collect the data, the kaggle portal (https://www.kaggle.com/datasets/nathan lauga/nba-games),. The main dataset includes basic information for every team throughout regular seasons and covers seasons from 2010 to 2023. Apart from Kaggle, information is obtained from a number of reliable sources, including sports analytics platforms, official league websites, and the NBA API. To guarantee the accuracy and applicability of the prediction model, it is necessary that the data acquired be precise, current, and span a sizable period of time. The different datasets used in the proposed work are as follows.

Games database: Data about every NBA game from the 2004 season till the most recent update is included in this. It contains details about the season, participating teams, date of the game, and some game-specific statistics like results. The data includes:

- Game date
- Game ID
- Game status (Final)
- Home team ID
- Visitor team ID
- Season
- Additional details like number of points

Player's database: This dataset includes the following details of the individual player.

- Player Name
- Team ID
- Racking
- Points scored

There are 1749 unique values in the "Player Name" column, indicating there is information for that many players

Ranking database: This database likely contains information about NBA team rankings on a specific date. The columns of this CSV file are:

- Date: This column would specify the date for which the rankings are provided.
- Team ID: This column would likely link to a corresponding team in the teams.csv file using a unique identifier.
- Conference: This column would indicate the conference (East or West) that the team belongs to.
- Rank: This column would represent the team's ranking within its conference on the given date (e.g., 1st, 2nd, 3rd etc.).
- Additional Fields (Optional): The file might include additional columns with metrics used to determine the ranking, such as win-loss record, total points scored, or point differential.

Teams database: This contains the information of different teams that have participated in the tournament that was held before.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 8, August 2024

Here's a possible breakdown of the data it might contain:

- Team ID: This column would act as a unique identifier for each team and might correspond to the Team ID in the ranking.csv file.
- Team Name: This column would specify the full name of the NBA team (e.g., Golden State Warriors, Los Angeles Lakers).
- Conference: This column would indicate the conference (East or West) that the team belongs to.
- City: This column would specify the city where the NBA team is located.

By combining the data from ranking and teams database, one can get a comprehensive picture of how NBA teams stack up against each other. It clearly shows which teams are ranked highly within their conferences, how the rankings change over time, and potentially explore the reasons behind the rankings using additional data sources.

B. Preprocessing

The different data pre-processing methods applied on the NBA databases include, handling the missing values, encoding categorical variables, standardization and scaling. Missing values have been handled by methodical cleaning procedures that make use of strategies like imputation based on mean, median, or machine learning algorithms in order to preserve data integrity and completeness. The categorical variables have been encoded by applying one-hot encoding, frequency and label encoding depending on the type of features. It is also required to bring all the values to the same range to avoid large deviation in the range of values of different features.

C. Building Machine Learning Models:

Building the machine learning model involves designing, selecting, and compiling models suited for predictive analysis. Basketball game outcomes are predicted using a variety of ML techniques, including LR, SVM, KNN and RF. The deep learning algorithm ANN also has been applied on the dataset and the performance of all of these have been measured. This section describes the different techniques used to predict the outcome of the Basket-ball tournament.

a. LR:

To investigate the relationship between several independent factors and a dependent variable—which is typically binary—a statistical method known as linear regression is utilized. Using the logistic function, which always makes use of the sigmoid function, one can utilize logistic regression to find the association between variables. Since the logistic function's output is a number between 0 and 1, it is widely utilized and very appropriate for predicting the link between winning and losing. The regularization parameter, regularization strength, solver, maximum number of iterations, tolerance, class weight, fit intercept, and intercept scaling are just a few of the hyper-parameters for the LR implemented in Sci-kit-learn[8].

b. SVM

Support Vector Machines (SVM) work in basketball result prediction by creating a model that distinguishes between different outcomes (e.g., win or loss) based on historical data. The fundamental principle of SVM is to identify a hyperplane that clearly classifies the data points in an N-dimensional space, where *N* is the number of features. A decision boundary that divides the data points into various classifications is called a hyperplane. The number of features determines the hyperplane's dimension[9].

c. Random forest (RF):

Powerful ML Classification and regression issues are handled by RF. Because it can handle complex datasets and produce feature importance rankings, machine learning practitioners enjoy it.

At the end, the chosen model is put together, specifying components such as the optimizer, loss function, and evaluation metrics that will be used throughout the training phase. Using past performance as a guide, the system is trained on historical data to predict basketball game outcomes. When the model is trained, it may generate accurate predictions, which aids stakeholders and NBA teams in making strategic decisions and analyzing player performance.

The key parameters that can be tuned to optimize the performance include no. of estimators, maximum depth of trees, samples to split node, samples per leaf etc[10].

d. H. KNN:

KNN is a kind of instance-based learning, often known as lazy learning, in which the model memorizes the training dataset instead of learning a discriminative function from the training data. KNN computes the distance between each new data point and each training data point in order to categorize or predict the value of a new data point. The algorithm predicts basketball game outcomes by analyzing the similarities between past games and the current game in question. Initially, a historical game dataset is assembled with attributes such as team statistics, individual performance measures, and game circumstances. Based on these attributes, the KNN algorithm determines the 'k' games in the dataset that are most similar to the present game when predicting the result of a new game. The forecast, which shows the percentage of the closest neighbors that led to a win, is sometimes stated as a percentage. This percentage gives a probabilistic evaluation of the game's likely outcome and represents the algorithm's confidence in the forecast.

e. Artificial neural network (ANN):

Artificial Neural Networks (ANNs) provide many benefits in machine learning, such as robustness against noise, flexibility in solving different problems, and the capacity to learn complex and non-linear correlations. Their proficiency

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 8, August 2024

in managing unstructured and high-dimensional data renders them perfect for use in picture and speech recognition, natural language processing, among other fields. ANNs can leverage parallel processing technology to expedite training and inference, and they can automatically extract pertinent features from raw data, eliminating the need for human feature engineering. Moreover, ANNs can learn continuously, scale well, and function well with partial or missing data. They are essential in domains like computer vision, healthcare, and autonomous vehicles due to their adaptability in a variety of tasks like classification, regression, clustering, and anomaly detection. ANNs are being utilized to estimate victory probabilities and forecast basketball game outcomes [12][13].

III. RESULTS AND DISCUSSION

An experiment carried out as part of the project "Basketball Result Prediction Using Machine Learning Techniques" yielded convincing insights into the predictive capabilities of various ML and DL regarding winning probability of basketball tournament. The ML and DL classifiers were carefully compared using a common dataset which has the information of players, teams and the results of historical matches.. The initial partitioning of the data set into training and test sets enabled a comprehensive evaluation of the generalization capabilities of the algorithms. Through the use of confusion matrices, performance metrics were carefully analyzed, including accuracy, true positive, true negatives, false positives, and false negatives.

Table 1 show the performance values of different classifiers. Figure 2 shows the comparison of these values. The results revealed different performances between the algorithms, with the RF emerging as the best with an impressive 97% accuracy. This superiority over Logistic Regression with 82%, SVM with 74%%, ANN with 82% and KNN with 79% underlines the effectiveness of Random Forest in predicting the probability of Team Winning. Not only in terms of accuracy, while considering other performance indicators such as Precision, Recall and F-1 score RF has come out as the best classifier for the dataset considered in the study.

For a number of important reasons, the RF approach frequently outperforms other single classifiers. A significant benefit is the decrease in overfitting. Overfitting to the training set is less likely with Random Forest because it aggregates the output of multiple decision trees. While individual decision trees are prone to overfitting, the model becomes more robust and broadly applicable when the predictions are smoothed out by the ensemble approach. Variance minimization is another important component. By sampling with replacement, RF uses bootstrap aggregating, also known as bagging, to generate different subsets of the training data. Because each decision tree is trained on a distinct subset, the model's stability and variance-reduction are improved. By ensuring that the model does not rely excessively on any one tree, this technique produces findings that are more consistent.

RF's aggregation technique helps them be resilient to noise. Random Forests have greater resistance to noise in the training data due to its majority voting process in classification and their averaging across numerous trees in regression. As a result, performance is more consistent and dependable because outliers and noisy data points have less of an impact on the final forecast. Furthermore, Random Forests are adaptable and scalable, working well with a variety of datasets and scaling well as data size and feature dimensions increase. By revealing information about the significance of a feature, Random Forests also offer perceptive judgments. This skill aids in feature selection for more straightforward model construction as well as in comprehending the underlying data patterns. Furthermore, by using the proximity of data points and averaging predictions from several trees—each of which may have a distinct missing value handling strategy—Random Forests manage missing values successfully.

In general, less parameter adjustment is needed for Random Forests to operate well. They are user-friendly and accessible for practitioners with different levels of competence because of their simplicity of usage. Because of the combination of these elements, Random Forests are a strong and dependable ensemble technique that frequently outperforms single classifiers, offering robust performance and excellent accuracy in a wide range of machine learning applications.

In general, shallow machine learning models perform worse than deep learning frameworks. Nevertheless, the outcome of this study demonstrates how poorly the ANN model performs. ANNs are prone to overfitting, especially when the training dataset is short or the noise level is high. This is especially true for deep neural networks with numerous layers and parameters. When a model learns too many details and noises from training data, it might become overfit and lose its capacity to generalize to new data. Random Forests, on the other hand, are more robust since they somewhat reduce overfitting through ensemble averaging and feature unpredictability.

Results suggests that ML models especially the ensemble models that combine the results of multiple classifiers effectively captures basketball dynamics, offering valuable insights for decision-making. Feature engineering was crucial for prediction accuracy, and model selection emphasized the importance of algorithm choice. Despite challenges, our study underscores machine learning's potential in sports analytics. The goal of this effort is to provide the coaches of basketball games with the information they need to make informed player selections. Basketball fans get even more engaged and excited when they can predict the results of games. As supporters cheer for their preferred outcomes and follow their predictions, it makes

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 8, August 2024

viewing sports more involved. This increased involvement can add to the excitement and fun of watching games.

Table 1: Values of performance metrics

	Accuracy	Precision	F1 Score	Recall
SVM	74	77	78	82
LR	82	83	84	87
RF	84	84	86	88
ANN	82	82	86	85
KNN	79	78	84	80

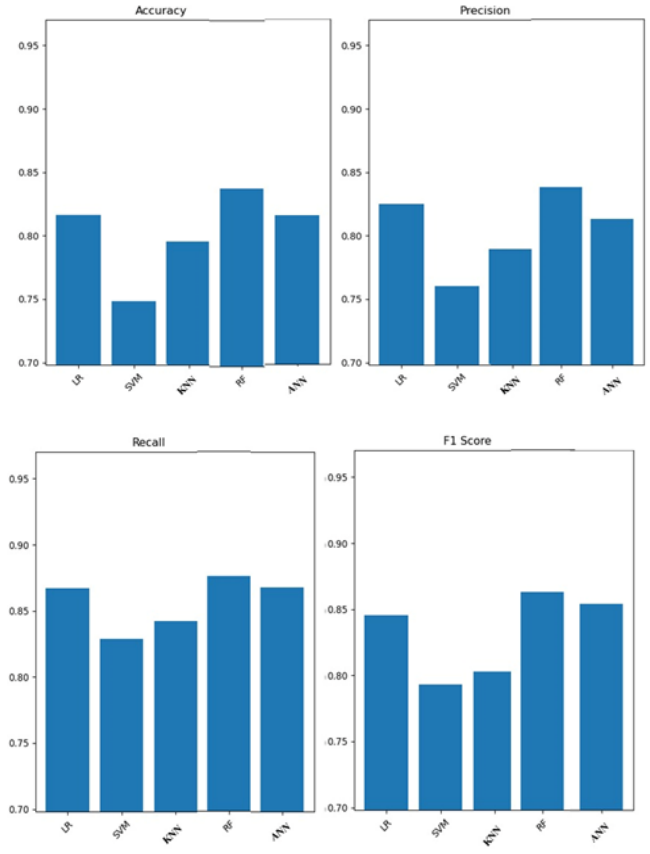

Fig. 2. The histogram comparison of the four models under Accuracy, Precision, Recall, F1 Score, and AUC value.

Table of classification models

IV. CONCLUSION AND FUTURE WORK

In summary, by using ML approaches to forecast NBA game outcomes based on player performance and team statistics, this study advances the area of sports analytics. Through the study, we were able to examine the effectiveness of different ML classifiers. By comparing the performance of these classifiers with the help of metrics such as accuracy, precision, recall and F-1 score, it has been observed that ensemble RF model performs well compared to the single classifiers.

This study makes a significant contribution to sports analytic by using machine learning to predict NBA game outcomes based on player performance and team statistics. The study identifies key factors that influence game outcomes, providing valuable insights for teams and analysts. Future enhancements in basketball result prediction include integrating real-time data sources like player tracking data for enhanced accuracy, exploring ensemble learning methods to improve prediction robustness, continuously refining feature engineering for nuanced predictions, developing interpretable models to enhance user understanding, incorporating domain-specific knowledge from basketball experts, and designing user-centric applications for enhanced engagement. These advancements aim to further leverage machine learning's potential in sports analytics, offering valuable insights for coaches, analysts, and sports enthusiasts.

V. ACKNOWLEDGMENT

We would like to express our profound gratitude to ***Canara Engineering College*** for their providing the eco system for conducting this study.

REFERENCES

[1] F. Thabtah, L. Zhang, N. Abdelhamid. "NBA Game Result Prediction Using Feature Analysis and Machine Learning," in Annals of Data Science, vol. 6, no. 1, pp. 103–116, 2019.

[2] J. Libed. "Basketball Game Analysis using Naive Bayesian Classification Algorithm Basketball Game Analysis using Naive Bayesian Classification Algorithm,", no. February 2017, 2020.

[3] C. Osken, C. Onay. "Predicting the winning team in basketball: A novel approach," in Heliyon, vol. 8, no. 12, pp. e12189, 2022.

[4] J. Perricone, I. Shaw, W. Swie¸ chowiczswie¸ chowicz. "Predicting Results for Professional Basketball Using NBA API Data,", 2017.

[5] B. Georgievski, S. Vrtagic. "Machine learning and the NBA Game," in Journal of Physical Education and Sport, vol. 21, no. 6, pp. 3339–3343, 2021.

[6] J. Lu, Y. Chen, Y. Zhu. "Prediction of future NBA Games' point difference: A statistical modelling approach," in Proceedings - 2019 International Conference on Machine Learning, Big Data and Business Intelligence, MLBDBI 2019, pp. 252–256, 2019.

[7] Horvat, T., et al. "A Data-Driven Machine Learning Algorithm for Predicting the Outcomes of NBA Games," in Symmetry, vol. 15, no. 4, pp. 1–18, 2023.

[8] Thorn, J. (2022b,August30). Logistic regression explained - towards data sci- ence. Medium.

[9] Jadhav, A., et al. "Predicting the NBA Playoff Using SVM," in of Web Development and vol. 1, no. 1, pp. 1–6, 2016.

[10] Chaya,"Random Forest Regression Level Up Coding". (2022, April 14).

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 8, August 2024

[11] Vamsi Saladi "Deep Shot : A Deep Learning Approach To Predicting Basketball Success," in no. 33, 2019.
[12] Loeffelholz, Bernard, Earl Bednar, and Kenneth W. Bauer. "Predicting NBA games using neural networks." Journal of Quantitative Analysis in Sports 5.1 (2009).
[13] Dertat,"Applied Deep Learning - Part 1: Artificial Neural Networks". Medium, (2022, November 21).