MACHINE LEARNING TECHNIQUES FOR PREDICTING RESULTS OF BRAZILIAN VOLLEYBALL LEAGUE

Abhinav Lalwani, Aman Smitesh Saraiya, Apoorv Singh, Aditya Jain
Major Project – BITS F464

Abstract

In this study, we develop machine learning models, using methods such as Artificial Neural Networks and Support Vector Machines, to predict results of the Brazilian Men's Volleyball League (SuperLiga). Data used to train and test the models were obtained from the league results of the past ten years.

We used features such as average points scored and conceded, form in the last 5 games, current league standings as input parameters, and match result as output parameter. As per the author's knowledge, there are no prior studies on volleyball match prediction. Related studies mostly focus on sports such as football, soccer, cricket and basketball.

Best results were achieved by using MinMaxScaler preprocessing and a SVM model with Gaussian RBF kernel and a regularization parameter of 1.0. The optimal accuracy was 77.9%, which is better than the median accuracy achieved by previous machine learning papers for each sport. Hence, we can conclude that volleyball results can be accurately predicted using our model.

Table of Contents

Introduction

With the increase in the amount and the type of data collected for various sports and increased computational power, it is now possible to gain deeper insights into complex information about the matches. Machine Learning algorithms aim to assist coaches and sport managers, to track a player's or team's performance, possible player's injury, scouting, and making sport-betting decisions. The use of machine learning algorithms to predict the result of a future match can be of significant use to make the right betting decisions. All these factors drive the research on machine learning in sports.

We attempt to develop Machine Learning models for predicting the winner of the Volleyball matches based on the standings and the performances of the teams in the previous matches. Raw data of the Brazil SuperLiga volleyball league is scraped from flashscore for all matches over the past 8 years. Appropriate processing and averaging methods are used to store the relevant features for every match.

This report explores some of the developed machine learning models, preprocessing techniques, appropriate model metrics, and the result obtained on the collected test and training dataset.

Related Works

One of the earliest studies to consider AI in the analysis of sports performance was done by Lapham and Bartlett in 1995 [9]. They showed that methods from artificial intelligence could be a rewarding future direction for the discipline.
Since then, many different AI techniques have been applied for sports prediction.

Purucker[14] attempted to predict results in the National Football League (NFL) using a Neural Network Model. Kahn [15] built on the work of Purucker, achieving 75% accuracy across the matches of week 14 and 15 of the NFL. In 2015, Tax et al. [7] combined dimensionality reduction techniques with Machine Learning algorithms to predict a Dutch football competition. Herbinet et al. [10] used a variety of machine learning techniques to predict the score and outcome of football matches, using in-game match events. Vaswani et al. [11] used an autoencoder based approach to simulate the results of the UEFA Champions League knockout rounds.

Although not as popular as football or soccer, volleyball is a team sport that has a national as well as international level competitions in almost every country. Wenninger et al.[12] predicted results of beach volleyball using machine learning methods. Tümer and Koçer [13] used artificial neural networks to predict league rankings in volleyball. However, there has been little to no prior research on predicting volleyball match results.

Dataset

Data Collection

Features in sport result data can be divided into several different subsets. Miljkovic et al. [1], for example, split the features into match-related(for basketball these include field goals made and attempted, 3 pointers, free throws, rebounds, blocked shots, fouls, etc.) and standings(number of wins and loses, home and away wins, current streak) features. We decide to use only standings related data for our task. We scrape data from the website FlashScore.in to build our dataset. We collected data from the seasons 2010-11 to 2018-19 of the Brazilian Volleyball League (SuperLiga), totaling 1289 matches. While scraping the data, we create embeddings for each team, so that for each match, we can use the embeddings to import contextual data from the previous matches for each of the two teams playing the match.

Feature Selection

For each match, we collect the following features, inspired by a study conducted by [7] and the features used in [8] :

1. **Matches won**: Miljkovic et al [1] showed the win ratios of basketball teams to be relevant features in predicting National Basketball League (NBA) basketball matches.

2. **Average points scored and conceded** : Baio et al [2] proposed Gamma distribution mixture model for Italian Serie A match prediction, relying on the amount of goals scored and conceded by both teams. We use exponential averaging to estimate the average points scored and conceded, so as to give more importance to recent matches.

3. **Performance in earlier matches**: Aranda-Corral et al [3] found the previous matches between the same teams to have medium to high correlation with the match result in their study focusing on the Spanish national soccer league. We encode this by calculating the exponential average of the set difference of the previous matches played between the two sides.

4. **Form**: Goddard[4] concluded that losing streaks result in an increased winning probability and winning streaks result in a decreased winning probability. We attempt to take into account whether the team is on a winning or losing streak by calculating the form in previous 5 matches(this is estimated using exponential averaging of the set difference of previous matches).

5. **Home/away form**: Palomino et al [5] showed that home advantage plays an important role in predicting match outcomes. Hence, we add separate columns for home and away form to our data, as this will estimate how good a team is playing at home/away from home.

6. **Importance of Match**: Goddard [4], showed specific end-of-season matches to be significant for match outcome. We assign an importance of 0 to league matches, 1 to quarter-final games, 2 to semi-final games and 3 to final games.

7. **Rest**: Constantinou et al [6] uses fatigue as one of the features in his prediction model. We attempt to gauge fatigue as the number of days played since the previous match. A team having 7 days of rest is considered to be well-rested and thus if the number of days

since the previous game is more than 7, we reduce it to 7.

8. **Team Ranking**: The team's position on the league ladder based on a list of the teams. This feature's use is obvious as a team with a high ranking is expected to defeat a team with a low ranking.

9. **Previous Year's Ranking**: We use this feature as a team who performed well in the previous year is expected to perform well in the next year as well.

10. **Performance in Previous Game**: the team's performance in their most recent game. In the first round this value is typically taken from the previous season's final game.

11. **Result**(Output Parameter): This is set as 1 if home team won the match and it set as 0 if the away team won the match.

Methods Applied

Preprocessing

The data collected is in raw form and contains numbers in a wide range, some feature's value ranges between 0 and 1, while some have a large range. Thus preprocessing is required, otherwise proposed models may not capture the essence of this wide range of data and may give less accurate results.
The preprocessing used in the models are MinMaxScaler and StandardScaler. The MinMaxScaler scales the data in the range [0,1] or in [-1,1] in case of negative values. It compresses all the inliers in a narrow range. The StandardScaler scales the data such that the distribution has mean value 0 and standard deviation of 1. It also ensures that the model does not give excessive attention to one feature and leave out other features which have small values. Preprocessing distributes the data in a specific range which helps the model in better prediction.

Model Evaluation Metrics

A comprehensive set of model evaluation metrics to assess the performance of our model, are described here.We first divide the dataset into two parts -- training and test through an 80-20 split. We train our model on the training set, tune the hyper-parameters using cross validation on the training set, and then we evaluate our model on the test set. All metrics provided here are as evaluated on the test set.

Accuracy

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

Accuracy is a simple metric that allows us to understand the performance of our classification models by seeing what proportion of samples it has correctly predicted. The accuracy is always between 0 and 1, and better performance is achieved for higher accuracy.

F1-Score

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

Where,

- $precision = \frac{TP}{TP + FP}$,
  TP is the number of true positive predictions (number of matches correctly predicted as win)
  FP is the number of false positive predictions (number of matches incorrectly predicted as win)
- $recall = \frac{TP}{TP+FN}$
  TP is the number of true positive predictions.
  FN is the number of false negative predictions. (number of matches incorrectly predicted as lose)

The F1-Score is in general recognized to be more useful than the accuracy as it takes false positives and false negatives into account. It is especially useful for us as we have slightly unequal class distributions. (Home wins are more common than away wins)

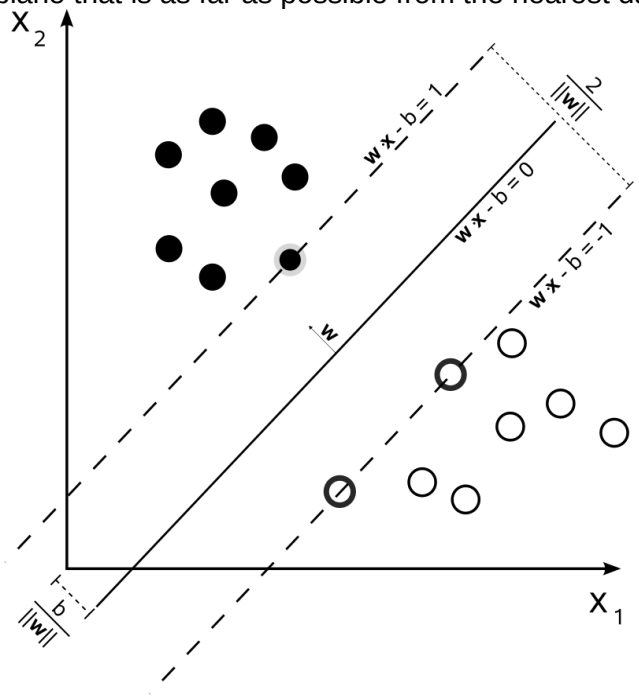
AUROC

AUROC stands for Area Under the Receiver Operating Characteristic (ROC) curve. The ROC curve is a plot of True Positive Rate vs False Positive Rate, as calculated at different classification thresholds. The area under this curve is the AUROC score, or simply, the AUC score. The AUC score provides an aggregate measure of performance across all possible classification thresholds. Essentially, the AUC score is the measure of the ability of a classifier to distinguish between classes

Model Architectures

Support Vector Machine

Support Vector Machines (SVMs) are Supervised Machine Learning models for both classification and regression. It is a non-probabilistic binary linear classifier. An SVM model represents the training data as points in space so that examples falling in different categories are divided by a hyper-plane that is as far as possible from the nearest data point.



As our data is not linearly separable, the kernel trick can be used, by using different possible kernel functions such as Radial Basis Functions (RBF), polynomial functions etc., in order to map the data into high-dimensional feature spaces and find a suitable high-dimensional hyper-plane.
We used a Gaussian Radial-basis function kernel, and a regularization parameter of 1.0

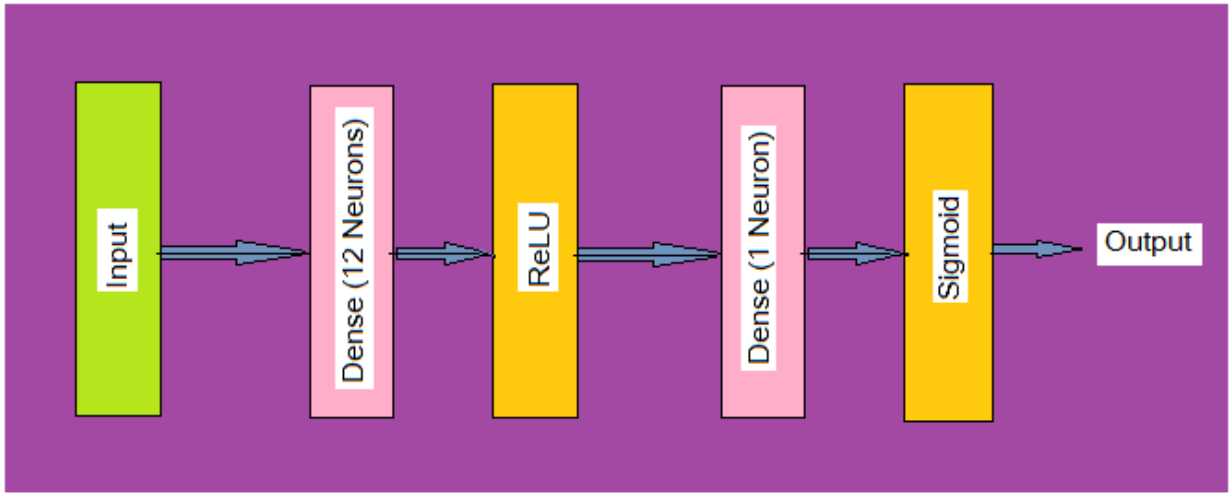
Gaussian RBF: It is a stationary kernel. It is parameterized by a length scale parameter $l > 0$, The kernel is given by:

$$k(x, y) = exp(-\frac{d(x,y)^2}{2l^2})$$

Where, l is the length scale of the kernel and $d(x, y)$is the Euclidean distance. This kernel is infinitely differentiable, which implies that GPs with this kernel as covariance function have mean square derivatives of all orders, and are thus very smooth.

Artificial Neural Network

The second model that we used for prediction is the Artificial Neural Network (ANN). Artificial Neural Networks are collections of neurons (nodes) at an algorithmic level. Multiple layers of neurons are connected together to form a feed forward network. Every neuron in one layer is connected to all the neurons in the next layers. ANN's are very good at capturing complex patterns in the input features.



For our prediction model we used an ANN with 2 layers. Due to less data a smaller network is used to prevent over-fitting. First layer has 12 neurons, it receives the preprocessed input, each neuron performs computation and the output is passed through ReLU (Rectified Linear Unit) activation function. The output of the first layer is passed to the output layer with 1 neuron followed by sigmoid activation function. Sigmoid function returns a value from (0,1). It is interpreted as the probability of the home-team winning the match. A threshold value of 0.5 is used to decide the result of the match. All the hyper-parameters for the model were tuned manually to improve the performance of the model.

Optimizer

The Adam optimizer with the default parameters β_1=0.9, β_2=0.999 and ϵ =10^{-7} was used. Adaptive learning rate is used.

Choice of Loss Function

We use the binary cross-entropy loss function for training.

$$Loss = Y(-log(\hat{y})) + (1 - y)(-log(1 - \hat{y}))$$

Other Models

We also tried using the following models:

- Linear-Discriminant-Analysis
- Logistic regression
- Random Forest Classifier
- Decision Tree
- K-Nearest Models
- Naive Bayes

However, the results obtained on these models were not as good as those achieved by SVM and Neural Network models.

Results

The project results are based on the two models Artificial Neural Networks (ANN) and Support Vector Machine (SVM). Two different sets of preprocessing, MinMaxScaler and StandardScaler were used in obtaining the following results.

Results with MinMaxScaler pre-processing applied.

Models/Evaluation Metrics	Accuracy	F1-Score	AUC-ROC
SVM	0.779070	0.797153	0.777119
Linear Discriminant Analysis	0.751938	0.768116	0.751453
Logistic Regression	0.740310	0.758123	0.739407
Artificial Neural Network	0.736434	0.757143	0.734504
Naive Bayes	0.713178	0.691667	0.724395
Random Forest	0.701550	0.725979	0.699031
Quadratic Discriminant Analysis	0.689922	0.639640	0.706961
K-Nearest Neighbors	0.666667	0.703448	0.660896
Decision Tree	0.612403	0.645390	0.608898

Results with StandardScaler pre-processing applied.

Models/Evaluation Metrics	Accuracy	F1-Score	AUC-ROC
Artificial Neural Network	0.771318	0.787004	0.770642
SVM	0.763566	0.784452	0.760835
Linear Discriminant Analysis	0.751938	0.768116	0.751453
Logistic Regression	0.748062	0.765343	0.747215
Random Forest	0.720930	0.737226	0.720884
Naive Bayes	0.713178	0.691667	0.724395
K-Nearest Neighbors	0.705426	0.734266	0.701271
Quadratic Discriminant Analysis	0.689922	0.639640	0.706961
Decision Tree	0.589147	0.613139	0.588136

The best accuracy (77.9%) was achieved by the SVM model with MinMaxScalar applied.

The results obtained from the models were also compared with a study conducted by Horvat et al. [16], which calculated the median accuracy achieved by various machine learning papers for sports prediction. We found that the accuracy of our model is better than the median accuracy for all sports analyzed in the paper, i.e., football, cricket, basketball, baseball and soccer.

Conclusion

Our main objective of building an expected winner model by exploring different Machine Learning techniques has been accomplished.
Predicting the match results before the match on the basis of past results is a difficult problem because of the number of factors in the volleyball league. In this study, many Machine Learning algorithms such as ANN, SVM, and Random Forest etc. were used in order to try and solve this problem. The statistics from 1290 matches between 2010 and 2019 were used to test and train the models, where the test and train set were split with a ratio of 1:4. The highest accuracy measured was 77.9% using SVM with MinMaxScaler preprocessing. This is slightly higher than the second-best accuracy which was 77.1% using ANN with StandardScaler preprocessing. Overall, we can see the results of this study as a positive sign that volleyball match results can be predicted accurately using Machine Learning techniques.

References

1. D. Miljković, L. Gajić, A. Kovačević, Z. Konjović, The use of data mining for basketball matches outcomes prediction, in: Intelligent Systems and Informatics (SISY), 2010 8th International Symposium on, IEEE, 2010, pp. 309–312.
2. G. Baio and M. Blangiardo, “Bayesian hierarchical model for the prediction of football results,” Journal of Applied Statistics, vol. 37, no. 2, pp. 253–264, 2010
3. G. A. Aranda-Corral, J. Borrego-D´ıaz, and J. Galan-P ´ aez, ´ “Complex concept lattices for simulating human prediction in sport,” Journal of Systems Science and Complexity, vol. 26, no. 1, pp. 117–136, 2013
4. J. Goddard, “Who wins the football?” Significance, vol. 3, no. 1, pp. 16–19, 2006.
5. F. Palomino, L. Rigotti, and A. Rustichini, “Skill, strategy and passion: An empirical analysis of soccer,” in Proceedings of 8th World Congress of the Econometric Society, 2000, pp. 11–16.
6. A. C. Constantinou, N. E. Fenton, and M. Neil, “pi-football: A bayesian network model for forecasting association football match outcomes,” Knowledge-Based Systems, 2012.
7. Tax, Niek & Joustra, Yme. (2015). Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach. 10.13140/RG.2.1.1383.4729.
8. McCabe, A., & Trevathan, J. (2008). Artificial Intelligence in Sports Prediction. Fifth International Conference on Information Technology: New Generations (itng 2008). doi:10.1109/itng.2008.203
9. A.C. , & Bartlett, R.M. (1995) The use of artificial intelligence in the analysis of sports performance: A review of applications in human gait analysis and future directions for sports biomechanics, Journal of Sports Sciences, 13(3), 229-237.
10. C. Herbinet, "Predicting Football Results Using Machine Learning Techniques," Department of Computing, Imperial College of Science, Technology and Medicine, London, 2018.
11. Vaswani, Ashwin & Ganguly, Rijul & Shah, Het & Ranjit, Sharan & Pandit, Shrey & Bothara, Samruddhi. (2020). An Autoencoder Based Approach to Simulate Sports Games.
12. Wenninger, Sebastian & Link, Daniel & Lames, Martin. (2020). Performance of machine learning models in application to beach volleyball data.. International Journal of Computer Science in Sport. 19. 24-36. 10.2478/ijcss-2020-0002.
13. Abdullah Erdal Tümer & Sabri Koçer (2017) Prediction of team league’s rankings in volleyball by artificial neural network method, International Journal of Performance Analysis in Sport, 17:3, 202-211, DOI: 10.1080/24748668.2017.1331570
14. M.C. Purucker. Neural network quarterbacking. IEEE Potentials, 1996. pages 19
15. J. Kahn. Neural network prediction of NFL football games. World Wide Web Electronic Publication, 2003. pages 19
16. Horvat, T, Job, J. The use of machine learning in sport outcome prediction: A review. WIREs Data Mining Knowl Discov. 2020; 10:e1380. https://doi.org/10.1002/widm.1380