Football Prediction System using Gaussian Naïve Bayes Algorithm

Conference Paper · March 2023

DOI: 10.1109/ICEARS56392.2023.10085510

3 authors, including:

Athish Venkatachalam
Clemson University
5 PUBLICATIONS **1** CITATION
SEE PROFILE

Rajeswari Devarajan
SRM Institute of Science and Technology
42 PUBLICATIONS **164** CITATIONS
SEE PROFILE

Proceedings of the Second International Conference on Electronics and Renewable Systems (ICEARS-2023).
IEEE Xplore Part Number: CFP23AV8-ART; ISBN: 979-8-3503-4664-0

2023 Second International Conference on Electronics and Renewable Systems (ICEARS) | 979-8-3503-4664-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICEARS56392.2023.10085510

Football Prediction System using Gaussian Naïve Bayes Algorithm

Athish V P[1]
Dept. of Data Science and Business Systems
SRM Institute of Science and Technology
Kattankulathur, Tamilnadu, India
vpathish@ymail.com

Rajeswari D[1*]
Dept. of Data Science and Business Systems
SRM Institute of Science and Technology
Kattankulathur, Tamilnadu, India
drajiit@gmail.com

Sree Nandha S S[2]
Dept. of Data Science and Business Systems
SRM Institute of Science and Technology
Kattankulathur, Tamilnadu, India
sreenandha24601@ymail.com

*Corresponding author

***Abstract*— Soccer (or, more colloquially, football) is among the most popular sports around the globe, with a thriving economy valued at more than $400 billion and billions of supporters (estimated) worldwide. Predicting match results has always piqued people's interest, and studying game results has shown to be extremely beneficial for corporate success and player development. There are many machine learning approaches for game prediction, however, it is believed the Bayesian approach could be very helpful in this scenario (given reliable historical data). The proposed model has been enforced on authentic squad information induding match results collected from kaggle.com and other websites like Sofifa.com. Observations indicate that the Gaussian Naive Bayes Approach is capable of predicting the results of a football match with an accuracy of 85.43%, which is a bit higher than the 79.81% accuracy that is achieved using the Decision Tree Classifier**

Keywords*— football, Gaussian naïve bayes, machine learning, Big data, goals, outcomes

I. INTRODUCTION

In order to forecast the outcome of a football game, this work would like to create a prediction system employing machine learning with big data. The Gaussian Naive Bayes model was used to fit the dataset and determine the chances of each team winning a game and ultimately the competition. The prediction methods have gained a lot of traction and molded careers in a variety of industries, including stock markets, weather forecasts, and many more. In order to create a system, this research work employs categorization tools. It is referred to as supervised learning since the class variables are taken into account along with the data processing component when building the model. Betting can be made more profitable by employing these methods, and team performances can be assessed using them, which helps coaches improve their game plans and purchase new players who would be the ideal replacements for the squad's outgoing members. In order to assess the odds of a team winning and place better bets, users may use a large number of elements in football prediction, such as past outcomes of games, player information such as speed, shooting power, passing accuracy, and so on. Predicting which club will win a tournament is the project's end goal since aware of the enormous sum of money involved in both online and offline sports betting. Once a prediction is made, the next issue is deciding how much money to gamble. The authors personally disclaim all liability for anyone who loses money using the proposed technique in the hopes that they would gamble sensibly. Additionally, sport managers seek models that could really forecast the playing patterns and styles of the opposition's teams and develop a plan of attack. From a data scientist's perspective, the fact that social media and other sources are being increasingly flooded with data on clubs and players, both personal and professional, simply serves as further proof that demand for such a prediction systemis rapidly expanding.

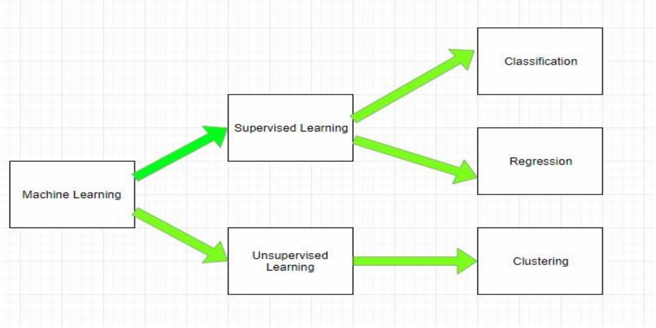


Fig. 1. Supervised Learning Vs Unsupervised Learning

II. LITERATURE SURVEY

A. Abel Hijmans [1] here saw that the authors study more than one technique particularly in data mining and their results were directed to invent predicting model to predict the fixtures of the Netherlands's football team. They have used some primary models like the K-nearest neighbor, Generalized Boosted modes and Naive Bayes classification. While the other models were not as accurate, Using GBM, they got the accuracy of 60.22%. The data-sets that only included data about the Netherlands's national team were the bases of this paper to make the prediction but more importantly no information was taken about the opponent other than their ranking in FIFA. If wish to improve it need to include more data sets and more information about the opponent teams and also better use of variables.

Hucaljuk, J & Rakipović [2], were trying to make a prediction system to predict the Union of European Football Associations Champions League (UEFA) with an accuracy of approximately 60%. The algorithms used were as follows - Bayesian networks, Naive Bayes, Random forest and k-nearest neighbors they used all these Major models so as to get the best mixture of classifiers and features needed to make correct predictions of the matches. Java and Weka API were used to build the software system. The authors have also stated that, the process of selection of features can be done a lot better to produce a better predicting system. Larger data-sets for learning for the machine would enhance the accuracy of the systemin a huge way.

Farzin, Parinaz, & Faezeh [3], have used the BNM model to make prediction just for one team and that is FC Barcelona which plays in the Spanish league of La Liga and that too just for the season of the 2008-2009. For that season They got an accuracy of 92%, the fact that they in season fixtures which is a limitation in this paper, they achieved this by dividing the data-set in various non-psychological aspects like result of last ten fixtures played, average goals scored, weather conditions, and injury prone players and so on. As only a single team was taken into consideration it poses as a weakness of the paper.

Darwin, P & Dra, H [4], have used logistic LR model to predict the English Premier League (EPL) for the 2015-16 season fixtures achieving a correctness of about 70%.The model was built by taking data-sets form sofifa.com and BPL website by using some major variables: Home Attack, Away Defense, Away Offense and Home Defense. Football predictor was the software used for implementing this. They worked towards predicting who is going to win a fixture, though it only uses four variables still the model gives a very good prediction , also it gives the odds for betting on a team.

E. Davoodi, A. Khanteymoori [5], Here saw that authors have are trying to predict Horse races using Artificial Neural Networks (ANN). For real horse racing data they have used many techniques like Quasi-Newton, Conjugate Gradient Descent, Levenberg-Marquardt, and Back-Propagation learning algorithms. Races included 100m, 500m and many more. Only authentic data was used.

T. Cheng, et.al [6], provides a way to predict football match results using NN (Neural Network). First, they divided the fixture in some classes with the help of machine learning techniques to get the difference in strengths of the opposing teams. Then the data that they have already designed was undergone back propagation to classify result. Actual football outcomes from the famous Italian league, Seria A were used to train and test and they got very good and accurate model.

Delen, D. Cogdell, and N. Kasap [7], wanted to determine by utilizing the score data to make a prediction of American football played at university level. Still, know that the scores provide a lot of detail and hence the importance of each static should be determined pre calculation of the results. Finally, also know that the winner occurs in a variety of ways. So they also took that into consideration also.

III. Proposed System

In the proposed system, trying to predict the outcome of a tournament like the world cup by feeding in the data into our system and apply machine learning algorithms like Naïve Bayes and others and find the best among them to get the best results.

In Fig.2 the architecture diagram, shows the life cycle of our project. Firstly the input, which is the raw data collect form the various sources like kaggle, wiki and many more and give it to our model as the input. After getting the input have to clean the data, and make it into information, this includes making labels and features for our machine learning model to perform its calculations at. After this have to split the data and place it in the exact location where it is needed in other words have to fit the data in our model. Then this data is fed into our machine learning classifier (model) and calculations are performed by the model on the data provided. Optimization is the next step where do some changes if needed due to some unexpected circumstances. And the finally have to predict the outcome and present it in front of the user in an understandable format.

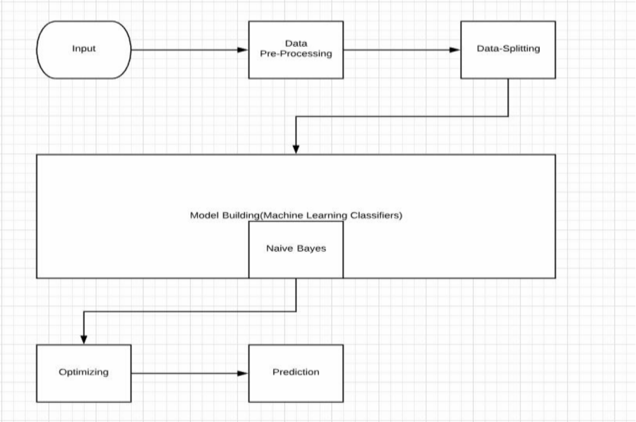


Fig. 2. Architecture Diagram

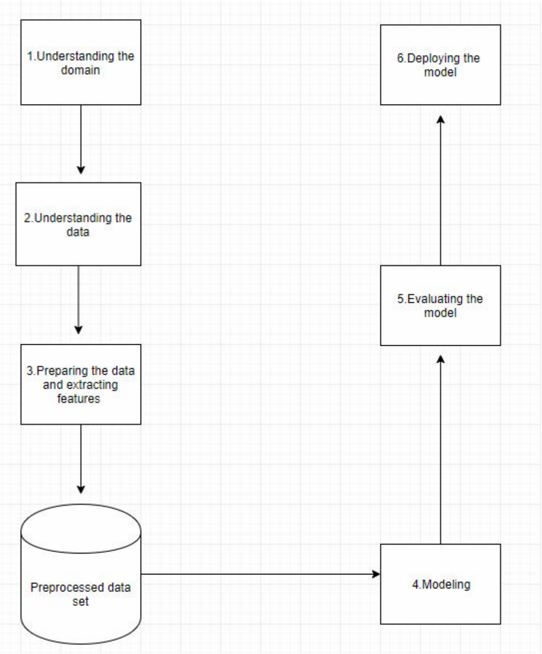


Fig. 3. Entity Relationship Diagram

In the above ER-diagram tried to explain the following -:

A. *Understanding the domain*

It includes understanding the problem, the goals of the model and the details of the game itself. The understanding of the game, how it is played and the points that are important to determine the result of the fixtures.

B. *Understanding the data*

- Often the data for sports prediction is available online and some sources have been automated, which means that the data is automatically taken up online and then feed into a record or a database.

- Then it is important to consider the granularity of the data. Team level data has been used previously as the training data, but now many models have been built on player level data.

C. *Preparing the data and extracting features*

- Many sub sets were created form the available features
- External and match related features were distinguished, which was required for preprocessing the data.

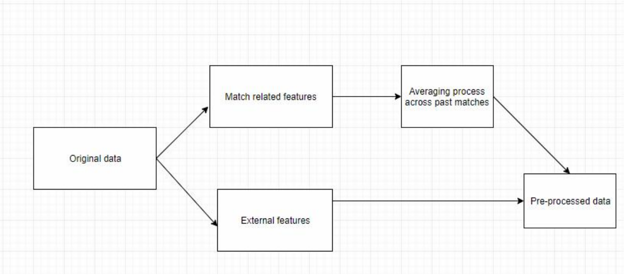


Fig. 4. Preparing Data and Extracting Features

D. *Modeling*

- Models based on the literature survey were selected.
- These models underwent different experiments based on machine-selected and human selected features sets

E. *Evaluating the model*

If data is not imbalanced select the measure of performance that is the accuracy of the model.

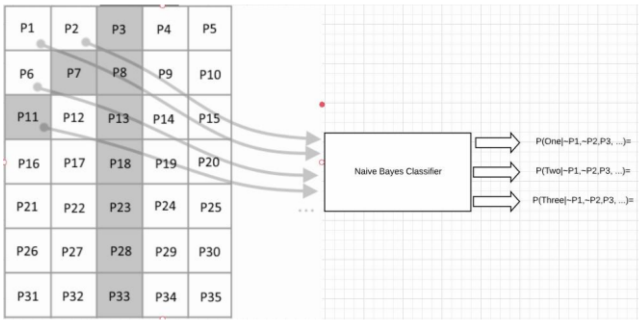
F. *Deploying the model*

- Automate the source data extract and data pre-process if possible.
- Re-train model based on fresh data.
- Generate predictions for upcoming fixtures.

IV. ALGORITHM USED

The algorithm used is called the Gaussian Naive Bayes algorithm. It is a machine learning algorithm coming under the supervised algorithm's category. It is mainly used for classification. Text classification is the domain in which it performs the best and used the most. Sentimental analysis and spam filtration are a few examples of its application. It is fast, easy to implement, requires less training data and is highly scalable.

Naïve Bayes is more effective than other classification techniques when there is conditional independence in data- Nearest neighbor is a supervised lazy classifier which cannot be used for real-time prediction. Random Forest deals with complex behavior of data and hence, if not carefully built, leads to over fitting.



P1	P2	P3	P4	P5
P6	P7	P8	P9	P10
P11	P12	P13	P14	P15
P16	P17	P18	P19	P20
P21	P22	P23	P24	P25
P26	P27	P28	P29	P30
P31	P32	P33	P34	P35

Fig. 5. Gaussian Naïve Bayes Implementation

V. RESULT AND IMPLEMENTATION

A. *Creating Groups*

Consists of Grouping for the tournament's group stages as taken out by the official FIFA draw and prints the respective groups.

B. *Preliminary Processing Of The Data*

Function to get the squad details for country from the Dataset:

```
function get_country_data(Country_name):
        return squad
```

Function to find the best 11 players for a given formation:

```
function_best_best_players(formation,nationality,df,measurement='Potential'):
    return squad_rating, squad_list, squad_stats.
```

C. *Computing The Winning Probability Matrix*

- Import GaussianNB from sklearn library.
- Read the features and Labels from WC_Competition_stats.csv
- Create a naïve bayes Classifier.
- Fit the features and labels to model the data using fit function of gaussianNB.
- Import team data from WC18_processed.csv
- Create a 32 x 32 matrix to match each team with every other team and determine the winner using predict function of gaussianNB.

D. *Printing the Final Winning Probability Of Each Team*

Given a matrix of pair wise beliefs about who will win/lose between teams generate a distribution over outcomes of the tournament:

```
function rr_outcomes(n):
    • Get all possible outcomes of a round robin tournament with n teams.
    • Result will be an nxn matrix where outcome(i,j) = i beats j.
    • Diagonals will be 0.
    return outcomes
```

```
function _mapper(ids):
        return head to head win-loss probs as a 32x32
        matrix where teams will be identified by rows
```

Generates a lookup table for any (winner, runner-up) combination. Generating beliefs:

- generate belief out initial bracket
- for inital bracket, col1 is winners and col2 is runners-up
- reorder so that the appropriate teams play each other
- 0 is group A, 1 is group B etc
- format: winner of groups 0-1 plays winner of 2-3, winner of 3 -4 plays 4 -5, etc

```
function next_level(bracket):
        return new_bracket
        //compute beliefs about winner
        print_mapper(l)
```

TABLE I. COMPARISION OF ALGORTIHMS

2018 World Cup Season	*#Wins*	*Accuracy*
Whatifsports.com	36	56.25%
Christodoulou's first approach	21	65.62%
Christodoulou's Second approach	44	68.75%
Neural Network	45	70.31%
Support Vector Machine	48	75.00%
Decision Tree	51	79.81%
Gaussian Naïve Bayes	**54**	**85.43%**

VI. CONCLUSION

Users can use this project to know which team is most likely to win a particular tournament the result win not only show the winning team but the percentage or rather the chances of all the teams with respect to winning the tournament. The user can if they want to bet on the teams using our project, but they should do that on their own risk as me as well as my partner do not guarantee that the team will surely win the tournament or not. This model gives an accuracy of 85.43% which is slightly higher than that of the 79.81% accuracy given by the Decision Tree Classifier. To conclude hope that anyone doing research in this domain will find our work helpful and hope that they can make a better one.

REFERENCES

[1] Hucaljuk, J & Rakipović, A. (2011). Predicting football scores using machine learning techniques. 2011 Proceedings of the 34th International Convention MIPRO, 1623-1627

[2] Abel Hijmans. Dutch football prediction using machine learning classifiers (unpublished).

[3] Darwin, P & Dra, H (2016). Predicting Football Match Results with Logistic Regression. 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA).

[4] Farzin, O., Parinaz, E., &Faezeh, S. M. Football result prediction with Bayesian network in Spanish league-Barcelona team. International Journal of Computer Theory and Engineering, 5(5), 2013, 812-815.

[5] T.Cheng, D. Cui, Z. Fan, J. Zhou, and S. Lu, "A new model to forecast the results of matches based on hybrid neural networks in the soccer rating system," Proceedings Fifth International Conference on Computational Intelligence and Multimedia Applications. ICCIMA 2003.

[6] E. Davoodi, A. Khanteymoori Horse racing prediction using artificial neural networks Recent Adv. Neural Networks, Fuzzy Syst. Evol. Comput., 2010 (2010), pp. 155-160

[7] D. Delen, D. Cogdell, N. Kasap A comparative analysis of data mining methods in predicting NCAA bowl outcomes Int. J. Forecast., 28 (2) (2012), pp. 543-552

[8] J.Edelmann-Nusser, A. Hohmann, B. Henneberg Modeling and prediction of competitive performance in swimming upon neural networks Eur. J. Sport Sci., 2 (2) (2002), pp. 1-10

[9] M. Fernandez, B. Ulmer, Predicting Soccer Match Results in the English Premier League, 2014

[10] D. Sikka and R. D, "Basketball Win Percentage Prediction using Ensemble-based Machine Learning," 2022 6th International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India, 2022, pp. 885-890, doi: 10.1109/ICECA55336.2022.10009313.

[11] D. Rajeswari, S. R, R. S and P. M, "Intelligent Refrigerator using Machine Learning and IoT," 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2022, pp. 1-9, doi: 10.1109/ACCAI53970.2022.9752587.

[12] N. Rajadhyaksha, "Modelling Veracity of Football Player Trade Rumours on Twitter Using Naive Bayes Algorithm," 2021 International Conference on Artificial Intelligence and Machine Vision (AIMV), Gandhinagar, India, 2021, pp. 1-5, doi: 10.1109/AIMV53313.2021.9670932.