Predicting Football Match Result in France Ligue 1

Rahul Saw Anish Dhanuka Shashank Shekhar Das

17BCE1179 17BCE1256 17BCE1325

***Abstract--*** Nowadays, the craze for sports among the people have increased exponentially according to recent survey FIFA world cup 2018 saw a total viewership of around 3.572 billion[1] which is half the world population. Football results forecasts are of extraordinary enthusiasm to fans and sports press. Over the most recent couple of years, it has been the focal point of a few investigations. In this paper, the analyst proposes Multi-layer perceptron, Support vector machine and Adaboost algorithm to anticipate the conclusive outcome of football matches. The specialists used the past results, Home and Away team, Average Position, Odds of team winning, average quality of each side for predicting the result. The procedure was applied to the 2019-2020 Ligue 1 matches and it was validated for 2018-2019 Ligue 1 matches. The outcomes acquired utilizing this model had a genuinely decent precision.

1. **INTRODUCTION**

Football is one of the most popular games in the world which is widely played all over the world. Prediction system typically works by learning from the past results the information is gathered. The information out there follows a specific pattern. the task of a prediction system is to look at the info and confirm the pattern therefore on predicting the results in a long run. In this paper, we have tried to predict the result of the matches of Ligue1 using Multi-Layer Perceptron, Support Vector Machine and AdaBoost algorithm and we have found out the accuracy of different algorithms.

Ligue 1 is a French skilled league for comfort station soccer clubs.[2] At the highest of the French league system, it's the country's primary soccer competition. Administrated by the Ligue American state soccer Professional, Ligue 1 is contested by twenty clubs and operates on a system of promotion and relegation with Ligue2. Seasons run from August to May. Clubs play 2 matches against each of the opposite groups within the league – once home and once away – totaling to thirty-eight matches over the course of the season. Most games happen on Saturdays and Sundays, with many games compete throughout weekday evenings. Play is often suspended the last weekend before Christmas for 2 weeks before returning within the second week of January. Ligue one is one in all the highest national leagues, presently hierarchical fifth in Europe behind Spain's La Liga, England's Premier League, Germany's Bundesliga and Italy's Serie A.

It has become a major area of study nowadays to increase the accuracy of predicting the result of a sport game. One common talk of the town is the result of the football match happening in different leagues spread all over the world. Everyone likes guessing the winner of the game. Many are in this profession of betting their money in different matches either online or offline using their knowledge or luck. They use different methodologies like winning streak of the team or current form .But actually that are not enough as football is much more complex game than it is thought.

Prediction of the result of a football match predicts whether a team will win or lose a particular match. There have been many different methodologies that have been worked until now. Most research papers use the current form of the team, its total winnings at home and away. Also, the accuracy of the system varies from the classifier to classifier and features to features. Features are primarily the parameters that have an effect on the end result of a soccer match. The style of a team within the previous few matches is one in all the necessary options. If a team has been losing all the five matches, there's a high chance that it'll lose the consequent match and the same is that the case of winning. However, this isn't the sole feature to be thought-about. Others are shots on target, goals conceded, red cards, yellow cards, injury of main players, home and away matches, goal distinction, etc. it's additionally discovered that the information out there in games like FIFA are terribly near to the particular ones. These attributes embody lightness, free-kick, dribble and lots of such options of individual players. Every player is allotted points for each of the attributes, which might be used as a parameter to check the potency of the players at completely different purpose within the game. This comparison will be used mutually with the options for prediction. A combination of varied options ends up in variable accuracies. the best combination will be selected on the basis of those accuracies.

The accuracy in most of the research papers in this field of study has been around 60 percent. While we are using one more unique feature over the rest like we are taking the FIFA ranking of the players in each team which will denote the current form of each player and thus each team. And, we have been successful in getting the accuracy around 64 percent using Multi-Layer Perceptron.

1. **BACKGROUND**

[3]This paper predicts the result of soccer match in English Premier League using AI and ML algorithms. A feature set is created which includes the game day performance of the player from historical data. After gathering the data set, an average of the performance of all the players was calculated which is used for categorization of the data set. Then a random graph of a player was selected from the given category. A cumulative team performance was calculated by the performance of all the players of the team. Then the performances of the two teams were compared to predict the final result. The prediction was in one of three classes for each game: win, draw, or loss**.** The dataset incorporated the execution of the players during games at each 15 minutes interval of the match. The diagrams were contrasted to decide how a player would play during a given game. After the exhibitions of the considerable number of players were resolved, they were utilized to ascertain the performance of the team during that match. The performance of the two teams was compared to figure out who might win. The paper used K-fold cross validation method to get a maximum accuracy of 41.2% for K value equal to 10.The accuracy could be improved by using a larger dataset

.

[4]Model training was done on a self-made dataset from public sources, consisting of thirteen seasons of Dutch Eredivisie match data. Several combinations of classification algorithms were tested on the public data training set in a structured way.

Models have been created for a betting odds feature set and a hybrid feature set (the union of public data and betting odds features).

The highest accuracy for the public data model was seen when the Naive Bayes or Multilayer Perceptron classifier was used in combination with a Principle Components Analysis (with 3 or 7 Principle Components), which achieved an accuracy of 54.702%. Spann et al showed it is possible to design a profitable betting strategy for a prediction model which was less accurate than our public data model (53.98%).

Although, it is necessary to note that taking into account all the features that have great influence on the results of football league matches is a difficult fit but this study has been able to find most features that are significant. By making use of machine learning and database that is more precise, this research[5] aims to proffer a solution that will help to make predictions more accurate and precise. Being aware of what the future holds is something that will be a lot easier by making use of this concept. Regardless, it is significant for wagering houses and bookies to utilize AI as well as ensure that it is joined by human control and direction in order to guarantee the outcomes acquired are exact and clear. At last, it very well may be reasoned that the examination needs increasingly sensible information and satisfactory capability in order to make the model created, progressively advance and steady.

There are different ways to predict the outcome of the football match such as Naïve Bayesian system ,k-nearest neighbor and others.

The choice of any technique depends on the application domain as well as the feature sets. The priority of a system developer or designer in most cases is to obtain a high prediction accuracy. A sum of 38 features were included and it yielded 53.3% accuracy using SVM[6]

.

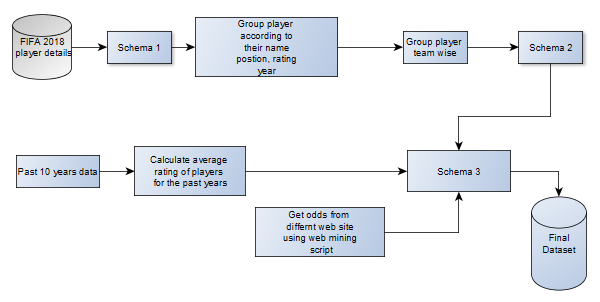
In this paper[7], the researcher propose Poisson relapse model to anticipate the last consequence of football matches. The analysts anticipated the normal objectives scored by each group by accepting that the quantity of objectives scored by a group in a match pursued a univariate Poisson conveyance. Poisson relapse model was defined from four covariates: the objective normal in a match, the home-group advantage, the group's hostile power, and the rival group's protective power. The system was connected to the 2017-2018 English Head League. The outcomes got utilizing this model had a genuinely decent precision. Accuracy for our modeling was 64%.

In this paper[8] they implemented 8 different models namely Baseline, Gaussian Naïve Bayes, Hidden Markov Model,Multinational RBF SVM, Random Forest, Linear SVM, One Vs All SGD.They got their best performing models as SVM with an RBF kernel, an SVM with a linear kernel, a Random Forest model, and a one-vs-all SGD model. These models achieved error rates between .48 and .52

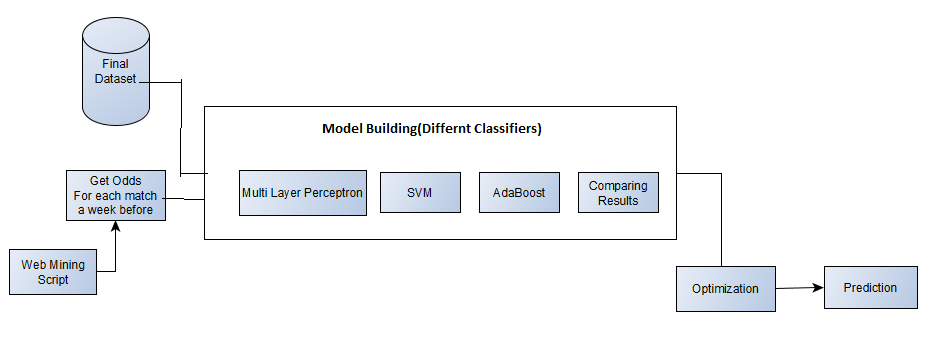
In the paper [9], the researcher present a methodology to assess football match results with Neural systems. At first, they have ordered a match into three classes utilizing a Learning Vector Quantization neural system to decide the quality differentiate between the two rivals. At that point they utilize explicit Back Propagation organizes on the information they have structured by the arranging result. They have prepared and tried their model on real football coordinate results from the celebrated Italian class, Serie A and have accomplished an incredible performing model with great precision

1. **PROPOSED ARCHITECTURE**

PHASE 1:- LEARNING



PHASE 2:- PREDICTING



1. The datasets used for predictions are past 10 years Ligue 1 results, FIFA datasets from 2014 to 2020 and odds which will are present in past results and will be dynamically fetched for future matches. We create 3 schemas namely Schema 1, Schema 2 and Schema 3. The final schema, namely schema 3 is fed as input to the model whereas the other two schemas, the web mining script to dynamically fetch odds and a input of the average finishing position of all teams create schema 3.
2. The first step is to clean our different datasets and ensure all the different datasets are in “sync". Quite often the names of teams or players are shortened or given a “pet name", for example Paris Saint Germain are quite often called PSG and players like Sergio Aguero are called Kun Aguero, these needs to be brought to a standardized format. In our problem, the names of players are not useful rather the name of team is considered, we are standardizing each team name by replacing shortened names by the full name of the team concerned.
3. For creating the first schema, from the final standings of each team for the past 10 years, we deduce the set of all teams which have been a part of Ligue 1 for each year since 2014, using this names of teams we find players who have played for a team in ligue 1 in that year from the past years FIFA datasets, we are extracting names, position, team, rating and year from which the data was generated of all players who played for a team in Ligue 1 in that particular year. Quite often in football, it is seen that players can play in a different position than what’s noted in our FIFA Dataset, for example Cristiano Ronaldo may play as a ST, LW or even a CF, so we need to standardize this too, to do this we are creating 4 groups namely Goal Keeper (GK), Defence (DF), Midfield (MF) and Attack (ATT) and are assigning players to these positions. This creates our Schema 1.
4. For creating schema 2, we take schema 1 and for each particular year we find all players that play for a certain team, then we find the sum of ratings for all four positions(namely GK, DEF, MID and ATT) and average them out individually to create a single entry for our Schema 2. This is repeated for all teams that have played in Ligue 1 for any year between 2014 to 2019. The attributes of Schema 2 are the team name, the average rating of attack, midfield, defence and goalkeeper.
5. Now we take the past 10 years data of Ligue 1 results and calculate average position of all teams participating in Ligue 1 between 2014-2019. Since Ligue 1, like many big leagues uses Promotion and demotion, we can’t guarantee a single team would play all the years, hence if a team is demoted and then again promoted, we give it a position of 21 for all the years it was demoted, we have assigned 21 as there are 20 teams and since it played in a lower level, its ranking should be lower than all the 20 teams so a ranking of 21. Example on how we are calculating average ranking, if in the year 2015, we want to calculate average position of a certain team, we check for the results from 2009 to 2014 and add the position finished and then average it, this gives us the average position of a certain team in a specific year, similarly if we want average position for a team in 2017, we take years from 2009 to 2016.
6. For past matches, we already have odds for each match, for future matches we mine the web to get odds of the home team winning, drawing or losing. This web mining script would automatically collect odds before the match and feed it to Schema 3. We are taking odds from 6 different companies to better solidify our model and its prediction.
7. To create our final schema, we take all matches from 2014 to 2019 and fill 3 columns , the standardized name of the home and away team and the final result, in the format H, D or A (Home team win, Draw or Away team win). Then for past matches we already have odds, so we include it directly from past results, for future matches the web mining script feeds data to the schema. Then, we derive data from schema 2, namely the average rating of each team in a definite year for each of the four above mentioned positions. This creates our schema 3 and this will be used as our training data.
8. We decided to use Multi-Layer Perceptron, Support Vector Machines and AdaBoost based on the performance of these models on prediction in similar genre. We are feeding our data to the model to predict the final class, I.e. H, D or A. To test these models, we divided our dataset into 2 parts, matches before 2018-19 and after 2018-19, the matches before 2018-19 served as our training set and the later the testing set. In our predictions on different models, we found that MLP gave us the best results followed by SVM and Adaboost. For predicting, present day matches, the entire set of past matches serves as our training data and the present match inputs serve as testing data. For matches in the present seasons, we have achieved reasonably good predictions, comparatively better than when we were testing our model with 2 separations.
9. **IMPLEMENTATION**
10. **Dataset Used:-**The following dataset were used for the prediction FIFA 2018 player rating, Ligue 1 last 10 year data, odds extracted from different website
11. **Software Used:**- Google Colaboratory for implementing all the machine learning models
12. **Models Used:**-
13. **Support Vector Machine:** A support vector machine (SVM) is machine learning algorithm that analyzes data for classification and regression analysis. SVM is a supervised learning method that looks at data and sorts it into one of two categories. An SVM outputs a map of the sorted data with the margins between the two as far apart as possible. SVMs are used in text categorization, image classification, handwriting recognition and in the sciences.
14. **Multi-Layer Perceptron:-** A multilayer perceptron (MLP) could be
15. a category of feed forward artificial neural network (ANN). The term MLP is employed equivocally, generally loosely to talk to any feed forward ANN, generally strictly to talk to networks composed of multiple layers of perceptrons (with threshold activation); see § language. Multilayer perceptrons area unit generally informally named as "vanilla" neural networks, particularly once they have one hidden layer
16. **AdaBoost Algorithm:-** Adaptive Boosting is a boosting algorithm that is used to boost the accuracy of the prediction.  AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. AdaBoost is sensitive to noisy data and outliers. AdaBoost refers to a particular method of training a boosted classifier. It can be used in conjunction with many other types of learning algorithms to improve performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier
17. **RESULTS AND DISCUSSION**

We have implement 3 models namely SVM, AdaBoost, MLP. Those models and their accuracy rates are illustrated in the given figure.

From the above illustration it is clear that the model having higher accuracy is Multi-layer perceptron the accuracy is around 64% which comparable with the accuracy rate of 62% for Ben Ulmer’s Predicting Soccer Match Results in the English Premier League[8].

Out of the models tested we found that SVM and MLP are showing a pretty good accuracy. The confusion matrixes of the following models are listed below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted win | Predicted draw | Predicted loss |
| Actual win | 117 | 17 | 43 |
| Actual draw | 13 | 33 | 29 |
| Actual loss | 12 | 23 | 93 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted win | Predicted draw | Predicted loss |
| Actual win | 103 | 13 | 43 |
| Actual draw | 37 | 47 | 23 |
| Actual loss | 20 | 23 | 71 |

1. **CONCLUSION AND FUTURE WORK**

The model which we have devised on the analysis of football matches to predict whether team will win, lose or draw was able to make fair accurate predictions. Unlike the other works, we have included the individual player performance using FIFA player rankings which gives better and conjugate results as team performance as a whole depends on all of its player and winning shouldn’t be decided just by the previous records or the current trends.

Although the accuracy of our model is quite good, there always remain a scope of improving it further and therefore, there is a lot of future scope in this regard. We could bring in features like trending hash-tags on twitter on match day, the posts from fans on social media or other trending football ideas in the social media which will require a more detailed research which could further help us in increasing the accuracy

.

1. **REFRENCES**
2. Forecasting English Premier League Match Results(February 2016)

By Kusum Lata,Prabha Gupta

1. Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach

By Niek Tax and Yme Joustra

3. Using Machine Learning to Predict Winners of Football League for Bookies

4. Support Vector Machine–Based Prediction System for a Football Match result(June 2015) By Chinwe Peace Igiri

5. Predicting Final Result of Football Match Using Poisson Regression Model

By H R Azhari

6. Predicting Soccer Match Results in the English Premier League

By Ben Ulmer

7. Predicting outcome of soccer matches using machine learning

By Albina Yezus

8. A Review Paper on Cricket Predictions Using Various Machine Learning Algorithms and Comparisons Among Them by GJ Harshit and Rajkumar S.