Rahul-211627054

Anish Saha-211627071

Atharva Singh- 211627057

ABSTRACT

**F**ootball is loved and played by millions of people. The official governing body of football, FIFA (Federation Internationale de Football Association) estimated that there were about 260 million football players and over 1.3 billion are interested. Hence, it is only natural that people would love to speculate the outcome of upcoming matches. Many researchers have utilized various algorithms to represent individual player skills and team performance, which are then used to create predictive models using multiple machine learning algorithms. This paper presents an approach that uses machine learning algorithms combining various datasets to predict future match outcomes. The approach analyses patterns within multiple datasets ranging from all time club and country rankings to goals per season, top performers, individual match outcomes and many more. The paper explores individual player and team statistics, the impact of home or away conditions on the winning, using only recent match outcomes versus all matches in the training set, and the prediction accuracy when creating separate models for each league versus a single model for all leagues. This paper applies machine learning to predict football match outcomes using data from the UEFA Champions league which covers the best clubs from top leagues in the world as experimental data. This data is pre-processed, divided into training and testing sets and fed into supervised learning algorithms.

INTRODUCTION

Football Leagues are widely celebrated all around the world. Multi-level competition fixtures are made over different national and international leagues segregated into divisions. Multiple leagues and tournaments take place simultaneously with viewers and enthusiasts from all over the world keen to know how these games uncover. FIFA and multiple other federations in football have taken advantage of this enthusiasm by bringing out the concept of fantasy football where fans from all over the world can win all sorts of prizes by merely predicting the score line of games and picking the best players expected to perform. The population involved in this is huge and there are multiple organizations holding databases of game-to-game records to individual player stats and team history from decades before. Modern technology has only ensured more precision brought to this data analysis as new aspects like VAR, Goal line technology and motion sensors in the football provide massive amounts of data every second. It is also the one of the most competitive sports in the world. Managers and coaches use every mean possible to increase their chance of winning. Every single piece of information is extremely valuable when millions of dollars are at stake. Using machine learning to predict the outcomes of matches has been an interesting task for data scientists for years.

Researchers have made multiple attempts to identify significant correlations between key variables in football match data in order to create a precise predictive model. To address this challenge, both classification and regression techniques were utilized. In order to predict the outcome of a football game and the winner of the league or tournament based on their past and current performance, the comprehension of data is combined with the understanding of the game. The UEFA Champions League match data from 1992 to 2022 was used as the sample data for this model.

The UEFA Champions League (UCL) is the biggest club football tournament in the world. 32 teams from every league in Europe participate in this year long league and knockout tournament.

Based on the how much a factor or attribute contributes to the outcome of the match, match date, goals, home and away grounds, fouls, shots, shots on target, corners, yellow and red cards, match outcome, team history, individual player stats, form and referees the finalized list of attributes and features is segregated. The algorithm aims to consider these aspects obtained from the datasets in order to predict the outcome of a given game eventually leading to the prediction of the tournament outcome and winners.

In the first part of the paper we have discussed the different features and attributes that could lead to predicting outcome of a match. Determining the level of impact each attribute holds was also derived here using previous data found on UEFA and FIFA databases. For example, Form of a current team might be a greater match deciding factor than Home or Away Matches. Using a set of operations numerical data is derived for each attribute. After the thorough data is obtained and processed, it is divided in the ratio of 80:20 for training and testing respectively. This is done keeping in mind the various uncertain values and discrepant data that may fall through for some matches.

In the second part, we look at the fixtures and group divisions of the league. Because top two teams from each group proceed to the next stage, a head to head comparison is done between the teams of each group with every possible match fixture possible. For Head-to-Head comparisons we look into the team’s past performances against each other. Using the features listed above and comparing team performances using different regression operations we find out the possible winner of each pairing. This deduces the top teams of each group which are then placed in knockout pairings. Same process is repeated for every match in the knockout stages. In case of a majority draw result, a predication model is made for extra time and penalties performance for every team. This is then reduced to the finals where same model is used to predict the winner hence delivering the final outcome.

In the third part, we find the exact possibility of each team winning the league using the same operations and classifications. This is required to fetch better and efficient method in the future. Different methods of Machine Learning will provide different levels of accuracy. In this case we used 3 machine learning techniques namely: 1) Random Forest 2) Logistic Regression and 3) Decision Tree. A detailed comparison is done between the different machine learning techniques using graphs and tables. A suitable method is then chosen from the list.

Finally, for the conclusion we discuss the further uses and scope of the model in professional prediction scenarios and see how the prediction model holds up against the various variables and discrepancies of the real world.