Design and Implementation of a Predictive Model for Nigeria Local Football League

ADEBISI John *adebisi_tunji@yahoo.com*
Engineering/Computer
University of Namibia
Ongwediva, Namibia

ALABI Damilola *oluwadamilolaalabi1@gmail.com*
Engineering/Computer
University of Lagos
Lagos, Nigeria

Abstract

Sports prediction has become more interesting especially in the era of statistical information about the sport, players, teams and seasons are readily available. Sport analysts have opted out in their traditional ways of analyzing sport events and tends to leverage on the advantages of sports data; this enables more realistic analysis beyond sentiments. However, football game was considered in this research. Data from Nigerian Professional Football League (NPLF) was used to predict result based on different conditions such as home win, draw and away win of teams in the league. Machine Learning, k-Nearest Neighbor and mathematical Poisson distribution algorithm was hybridized using data mining tools together with Anaconda packages. The model accuracy was compared with other online bookmarkers, and it yielded 93.33% accuracy which will be helpful in making substantial profits in within the economy through the betting industries. This model is practically based on the home and away matches coupled with historical trends of goals scored and winning of previous matches, by implication, Nigerian football league will be more enhanced to catch up with their international counterparts and the players tends to get more feasibility from match result predictions for international participation and employment opportunities.

Keywords: Football, Sport, Machine-Learning, Poisson Distribution, Data Mining.

1. INTRODUCTION

Sport is defined as an athletic recreation which include tennis, soccer, and golf to mention a few. Any form of professionalism in the act is often called athlete. In simple term, sport can be either played indoor or outdoor. Sport can be further grouped into individual or team sport. During the sport competition, those who watch either in the stadium or on the screen are called fans[25]. While other watch for fun, some were paid to watch and are called the spectators. In this research, football sport is considered as it is the most populous type of sport in Nigeria. Football is a sport that is played by two different teams and each team with eleven players. Football game involved varying degree of kicking and nodding of ball with the main aim of securing a goal. The game has some rules to abide to [30]. Every player on the field will all participate to the overall performance of the match. The game is played on the field in a stadium with fans watching and cheering their favorite team. Officials also perform their duties one of which is the referee who watches the game closely and ensure that no rule is violated. Others may include commentators. The winner of the game is determined majorly by the team that secure more goals amidst other conditions based on the stage of the game in a tournament/league. Fans often try to predict the winner of football match which has led to the rapid increase of sports prediction [5].

Sports prediction has been increasing in popularity for many years in Nigeria especially in the era where everyone has access to internet connection. The urge to predict correctly among football fans gave rise to organizations specialized in sport prediction for profit making, hence the development of a more efficient and effective prediction model is required especially for local leagues. In the last few years, many stakeholders have opted to sport prediction and get rewarded for I t[32]. This has increased the participant in the sport industry as they are more passionate about the rewards, they get from predicting the outcome accurately [6]. Due to this, the sport industries provide the best reward and maintain relationship with their participant to make the sport more relevant in the society. In most society, when the stakeholder tries to forecast the event of a sport played, they go back and draft the past histories of the sport teams, then judged based on the information provided to predict whose team is going to win or lose. In case of sports like racing, tennis where the factors affecting the outcome of the game is less, it would be easy to analyze without sentiments, but when it comes to sport like soccer, several factors need to be considered before prediction. The prediction was usually carried out manually thus waste a lot of time since every data in the game that would determine the winning or losing team will have to be put into consideration. This led to the development of sport result prediction monitoring system using Nigerian Professional Football League (NPFL) as a case study.

NPFL is the highest level of the Nigerian Football League System (NFLS), for the Nigerian Club football Championships. It was organized by the League Management Company (LMC). Currently, the NPFL has twenty and four (24) football clubs all over the region in Nigeria who are currently playing in the league season. Although it was formally known as Nigerian Premier League. The team will play against one another, and the ultimate champion will be rewarded with the Nigeria FA Cup. The best three (3) teams qualifies for championship for Confederation of African Football (CAF).

With the upsurge in modern technology, many stakeholders use various tools as means to extract useful past data to evaluate them effectively. To overcome the problems faced by majority stakeholders, sport result prediction monitoring system is implemented. The model is incorporated into a web format for public access through the internet. Some merits of the monitoring system are; continuous support of training with new data for better accuracy, it also allows stakeholders spend less time analysis sport event outcomes. The manual methods of analysis sport results will void since the system will generate the result within second and hence decision can quickly be made by stakeholders to for result determination even when the game is ongoing. In addition is, awareness creation on the team involved which is better than just guessing the outcome of the game even without prior knowledge of such team. Increases in public interest on football game especially with prediction feature goes a long way in the improvement and monitoring of the local football league, then with the implementation of this model, there will be more awareness of the NPFL matches. Result of prediction will be uploaded online to enhance prediction accuracy.

1.2 Significance

In addressing the identified problem, this research designs and implement a predictive model for monitoring system using NPFL data. The objectives are to retrieve NPFL historical data from public database, design an interactive web platform that allow local sport fans and managers to predict and monitor NPFL sport match outcome and implement the designed model. The implication of this research is to monitor sport event outcome using the NPFL. This will enhance sport managers victory strategies and further strengthen sport fans to watching more of this local league since this system predicts the results of each game for sport fans and in return get rewarded for. This brings a lot of advantages to the local league management with easy accessing of this predicting platform, more accuracy implies more awareness to the public and in turns give popularity to the leagues. This model will also maximize data retrieved during matches as it gets popular through regular update thus, boosting the online community of local sport fans and this can gain international recognitions.

1.3 NPFL Peculiarities

The draw of NPFL consist of two groups; Group A and Group B. each group consists of twelve (12) teams. The teams are paired up in their respective groups and will play against one another in their groups both in the home and away; by the end of the tournament, a team would have played against other twenty-three (23) teams in home and away each to complete a total match of (46). Points will be awarded according to the win draw or lose. For a win match the team will be awarded 3points, for a draw match team will be rewarded 1point, and lose with 0point. The best three (3) teams in each group qualifies for the next stage and the four teams with the lowest points in each group relegates (out of the tournament) and demoted to Division two (2) considered in determining the winner or loser. Rules in global football tournament also hold in NPFL.

2. THEORECTICAL FRAMEWORK

There are number of events that can be predicted in a football match such as the number of goal scored either for half time or full time, the player to score the next goal, the team to score first, or even to predict a fix match which is a determination of the exact goal that will occur in the match [9]. Nevertheless, football is full of unpredictable event and so this work will only be classifying our prediction into the win/drawn and lose. The research interest is in the supervised learning method of Machine Learning (ML) and in particular the classification methods such as Logistic Regression (LR) Model, Decision Tree (DT), Random Forest (RF), k-Nearest Neighbour (k-NN), and Support Vector Machine (SVM)[37]. Others include Artificial Neural Network (ANN), Lazy learning and Bayesian method. Machine Learning capacity is explored in this research alongside with others for effective prediction. ML allows computer decides by itself with little or no assistance. ML uses data and statistics for analyses just like statistics but uses different methodology. The major two categories of ML is the supervised and unsupervised learning. The Supervised learning is later grouped into regression and classification method. Some of most important ML methods are: “Supervised Learning (Classification and Regression) [22].

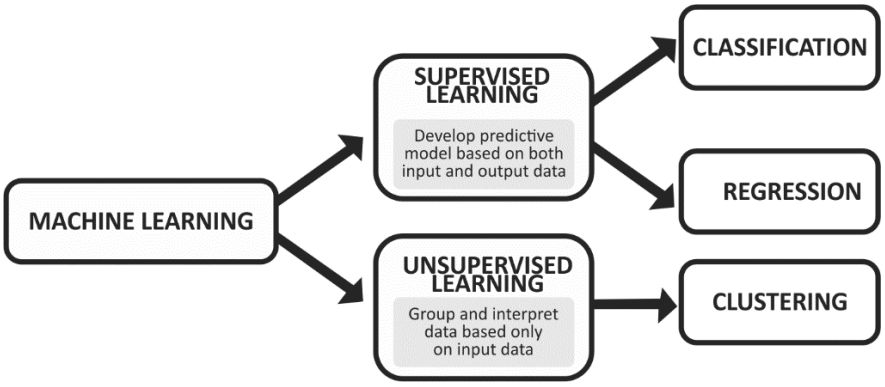


FIGURE 1: Supervised and Unsupervised L1earning (Source: https://de.mathworks.com/).

a) Artificial Neural Network

ANN mimic the human brain; the network has neurons. Each neurons have a weight value and are connected at the node. The network is consist of at least one input and output with some interconnected neurons in between the input and output nodes [4]. The core of neural networks, neurons, are just simple activation function that has multiple inputs and one output. The neuron can be seen as a composition of several other weighted neurons and the network can be described by the network function in equation 1.0 as described by [22]. The neuron output may be the input of another neuron and every neuron weight the input and calculate its activation value give;

$$f(x) = k((\sum_i M_i c_i(x)) \qquad (1.0)$$

Where *Mi* are weights, *Ci* are other functions and K is the activation function. Generally, ANN continuously changes the weight of each hidden node depending on the output weight.

b) Bayesian Method

The Bayesian model is one of the well-known supervised machine learning classification techniques. It is easy and effective in performing well with unrelated and distinct features. Bayesian classifier is a probabilistic forecast system that implies that all characteristics variable involves does not in any way depend on one another i.e. are independent from the class variable. There are some separate characteristics in each category. Based on prior information, it then predicts the outcome event information. In the existence of complexity and uncertainty, Bayesian networks are graphical models for estimation [6]. Bayesian network's primary concept is effectively derived from Thomas Bayes' called the rule of the Bayes. The Bayes' theorem is represented in Equation 2.0 or Equation 3.0 by[28].

$$P(\text{hypothesis} \mid \text{evidence}) = \frac{P(\text{evidence}|\text{hypothesis}).P(\text{hypothesis})}{P(\text{evidence})} \quad (2.0)$$

Where:
P(hypothesis | evidence) The likelihood of the hypothesis after proof is observed.
P(evidence |hypothesis) Describes the likelihood of evidence for a given hypothesis
P(evidence) as the evidence of the unknown cause in the event.
P(hypothesis) The probability of all event before observing their effects.

Generally, the formula can be interpreted in the equation below.

$$P(J|K) = \frac{P(K|J).P(J)}{P(K)} \quad (3.0)$$

Where:
P(J) is previous likelihood of J.
P(J|K) is the later likelihood of J given K
P(K|J) is Conditional likelihood of K given J
P(K) is the prior probability

c) Linear Models and Logistic Regression

Linear models are collection of regression techniques that assume that the output figure is a linear mixture of all input variables.

Consider the diagram above which is the graph of a straight line

$$y = \alpha + \beta x \quad (4.0)$$

In the straight-line Equation in 4.0, β is the slope and α is intercept on y. This relationship may not be true for large dependent and independent variable which lead to another equation when observing n sample of data as shown in Equation 5.0.

$$Yi = \alpha + \beta x_i + \varepsilon_i \quad (5.0)$$

The aim of this model is to determine the value for the α and β for which the output is formed on the best fit line on the other way, Logistic Regression (LR) is a distinguished classification method. Unlike linear regression, LR relies on linear feature mixture, which is then plotted by the logistic feature to a value between 1 and 0. Thus, dependent factors should have a constant significance that, in turn, is a function of event probability. There are two phases of logistical regression. First, estimate the probability of each group's characteristics and second, determine the cut-off points and categorize the characteristics appropriately by [23].

d) Decision Tree and Random Forest
DT is a common ML method for linking entry factors (input) depicted in the branches and nodes of the tree with an outcome value (output) displayed in the leaves of the tree. Trees can be used either in classification analysis, by producing a class tag or in regression analysis, by producing an actual number. Decision Trees can be installed using various methods, including the most common CART or ID3 DT systems. However, DT can often become incorrect, particularly when subjected to big amounts of information from practice as the tree becomes a victim of over fitting. This happens when the model fits the training information but cannot generalize to unforeseen data. Random Forest (RF) on the other hand is a combination of different DT in DT training output node is the input to another DT which form the RF classification. RF has better performance when dealing with over fitting.

e) Support Vector Machine (SVM)
SVM are both classification and regression ML algorithm. An SVM system reflects the training data as space points. New variables are plotted and categorized as the class they drop into (which becomes part of the hyper plane) in the same manner as the training data. The kernel trick can be used if the information is not linearly separable by using various feasible kernel features such as (RBF) or nonlinear features. The SVM algorithm looks for an ideal hyper plane that functions as a border of choice between the two categories. While training, SVM lasts longer than other techniques. The algorithm is highly accurate due to its elevated capacity to build non-linear, complicated choice boundaries.

2.1 Lazy Learning
It is a ML technique which has no actual model for the training. The overall model is trained based on the new input. It is best used for data set that has relatively small features but has a reasonable large amount of data. One of the algorithms that can be classified as Lazy learning is the k-nearest neighbor which can be used for regression and classification analysis. The k-value depends on the sets of data set. The larger the value of k the lesser the noise effect. In this model if the new data set (green object) is to be classified.

a) Poisson Distribution and Poisson Regression
This is used in the statistics field to determine the outcome of an event in form in probabilities. The Poisson model was first described by Simeon Denis Poisson who was a mathematician at that time in France named Haight in 1967. [11]. This model is used to determine the occurrence of an event that takes place in interval. Example where this model can be applied is in football game to check the goal probability that can occur in a match between two team using the past average goals scored by individual team. Using the mathematical model f (k; λ) [10].

$$\text{PMF} = \Pr(X = k) = e^{-\lambda}\frac{\lambda^k}{k!};\ \lambda > 0 \qquad (6.0)$$

Equation 6.0 is a Probability Mass Function (PMF). Where λ is the avg. number of event that can occur i.e. the value of X in the conditional variable event λ = E (k|X) > = 0 and the e is the Euler's number of 2.71828… and the k ϵ {0,1,2,…} is a positive integer in factorial. The Poisson Regression is used to generalize the Poisson distribution into a linear model.

2.2 Empirical Review
Various researchers have come with different methods and techniques to predict result. Meanwhile, the analysis serve as a guide line for the public interested in football both for the coaches viewers. In this section, this research consider various researchers who have done relevant works in this area. Graham and Stott[9] applied an ordered probit-model using one fixture in the team to determine the strength of the team individually, but the major drawback in this model is lack of dynamic update.

[2] developed a machine learning framework to expand areas of necessity for good predictive accuracy in sport prediction based on artificial neural network to formulate informative strategies

however not designed for football. [30, 31 and 32] presented a comprehensive overview of big data management which is not really in sport but some of its concepts are important to this work in social data area since sports are regarded as social activities. Impact of software architecture in product upgrade and maintenance become very important as system deployment spanning across decades risk increased complexity that could only be managed by proper maintenance, which is very useful to enhance the longevity of the work examined by this research. Exploring the power of real time result processing and application is the works of [1], an aspect of their methodology during data gathering was very useful for this research to ensure real time prediction result processing. [37] uses MATLAB for sport prediction unlike the web application developed for this work. The works of [35] and [36] applied machine learning to the prediction of the outcome of professional sports events and to exploit "inefficiencies" in the corresponding betting markets. Tenis was the major subject of discuss and not football as addressed in this research.

[27] predicted results in the National Football League (NFL) using an ANN model which was conducted during one of his initial studies. He selected five fixtures in the first eight rounds of the league, consisting of “yards gained, rushing yards gained, turnover margin, time of possession and betting line odds”. The research used unclassified part of ML to determine which team is best and which one is poor. Achieved 61% accuracy. However, limitation of this study was the limited number of fixtures, and the model cannot be used to classified match for a win/lose/draw.

[29] developed a model using Artificial intelligence hybridized with Multiple Linear Regression and expert human predictions to determine the outcome of soccer and rugby game. The English Premiership Football teams and 2 Premiership Rugby Union team were used as a case study however much details of the uniqueness and relationship between soccer and rugby was not presented unlike, McCabe and Trevathan[19] who presented an extension of earlier work after three years Reed and O’Donoghue[28] published theirs. Artificial intelligence was used for prediction of sport game event. A multi-layer perception was used to model the system. “The information used in this model has been drawn from various sources and includes four main sports in the league which were the Australian National Rugby League (NRL), the Australian Football League (AFL)” McCabe and Trevathan[19] explored Super Ruby and English Premier League Football (EPL) in 2002. The fixtures acquired were focused exclusively on information such as score line, latest results and "league ladder" location compared to other teams. An accuracy of 65.1%, 63.2%, 54.6% and 67.5% AFL, NRL, EPL, was obtained and Super Rugby League, respectively. The work tried to decrease the Bayesian hierarchical model's over-shrinking difficulty by implementing a blend model, making the system more complicated and time consuming. The system can forecast outcomes for only a team.

[33] implemented a system for predicting sport game. Data set of two successive seasons of the (NBA) League were used. Data were collected from NBA league then uses module in his system. Unfortunately, the system uses the referent classifier and thus, absence of comparison with similar research in which to compare the predictions on the same set. Also the system uses manual fixture selection, however, the prediction accuracy is quit reliable. Furthermore, a group of researchers also worked in the area of the National Basketball Association (NBA) sport in 2010, Miljkovic *et al* uses data mining to forecast the results of NBA league basketball matches. The model uses classification problem which include the native Bayes method. Also multivariate linear regression to determine NBA spread. The data set of 2009/2010 NBA season was used to evaluate their system and achieved an accuracy of 67%. The system only uses win/lose as required in NBA, so it cannot be used for sport competitions that give room for draws such as football.

[18] considered the Problem in the competitive horse racing framework and show how to adapt the RF Classifier to forecast the results. The assessment was focused on a dataset of 1000 games between 2005 and 2006 at Hong Kong racetracks. The major drawback in this method was the complexity of the model used for a simple sport as it win/lose is mostly determined by individual/ horse capability. [15] presented the application of the weighted probability strategy

using weighting systems that are simple to obtain. This method focuses on the amount of goal the two competitors secured. The data from the Champions League was used to show the capacities of the suggested approach. Although, this strategy makes it possible to predict the initial score adequately and accurately, it does not account for big or unexpected final score that may deteriorate parameter projections. It uses the goal scored by team; it does not give account for shocking goals that can weaken its approximations.

[4] suggested a predictive scheme for football games beating the likelihood of bookmakers (odd). The prediction for the matches' uses previous result of the team involved as a guideline. The match projections use the prior team outcome as a guideline. Data set of the last 15 years for the Dutch football competition were used. In their inquiry, some of the most significant ML algorithms were used such as the BN and a form of LR. Features such as number of home wins, number of goals etc. The accuracy was not reliable since it was never much higher than a mere 55%, also, there is absence of comparison with similar research. In the works of [24], the approach to forecast results of soccer matches using the NETICA software. The Spanish League-Barcelona team during the 2008-2009 season was used to test the performance of the technique. Factors which affect the outcome of football matches were evaluated which were divided into non-psychological factors such as average player's age, history of five previous matches and psychological factors like the weather. The number of goals conceded by each team was categorized by BN which determines the (win/lost/draw). Following the results; a comparison with 2008-2009 seasons and gives an accuracy of 92%. The limitation of this model was football data affected the outcome of match in terms weighting. Although, many factors were considered in the model but most of the factors has little effect on football outcome as used.

[11] focused on data mining techniques used for sport prediction. The research work reviewed various techniques such as ANN, Decision trees, SVM, Fuzzy Systems, Bayesian methods among others. They evaluated available literatures in this regard and detected two major challenges. In this respect, they assessed literatures and identified two significant difficulties. First, the small precision of projections demonstrated the need for further studies to achieve accurate result. Second, the absence of an extensive collection of statistics pushes the research to gather information data from sports pages". They propose a range of alternatives to address these problems one of which was to improve prediction accuracy through ML and data mining techniques that have not been used in the field of football prediction but have been used in other field and yielded good results. They concluded that the application of hybrid algorithms can increase prediction accuracy.

[12] uses data mining tools to implement the model and weigh and predict the outcome of a soccer game. Using data mining software such as Rapid Miner to mine football information. Fixtures like the number of goals scored, moving average of teams, performance within a season, players and managers' performance indices, history of team, and weather conditions were used in this research. The system uses nine fixtures with optimization of weight. The structure utilizes two distinct methods of information mining which were ANN and LR techniques. The data set comprises of 110 games in the 2014-2015 season of the English Premier League. Comparing the result outcome of their model, "a greater forecast precision was obvious when weighting optimized characteristics. The ANN technique yields 85% while the LR yield 93 %. Although, the LR model cannot predict if a match must be draw. In this case the ANN is more accurate when compared with LR technique when predicting if a match must end at draw. The major limitation with the ANN approach is that it requires major factors that affect the outcome of a match and need as many data as possible to be more accurate.

[2] did something quite similar to Dixon and Coles[5]. They proposed an advanced modeling strategy and predicted the result-based Poisson Auto-regression with exogenous covariates (PARX) of football games. Used the 2013/2014, 2014/2015 and 2015/2016 English Football Premier League information season. This research work too advantage of the goal intensity feature to determine the best team. Based on the performance of the model it yielded of 43.27%, 44.96% and 12.63% respectively for the seasons 2013/2014, 2014/2015 and 2015/2016. The

threshold was modified to 0.3 and they achieve a return greater than 87% for those three Premier League Season. The limitations behind this model is that sophisticated features and factors that determine outcome of football was not considered. Considering the fact that the accuracy was low and thus depends on threshold value which may not be determined in real life event. A return greater than 87% was achieved for those three Premier League Season with limited fixtures and factors that determine outcome of football.

[32] used hybridized ANN and Linear Regression to predict the score outcome for matches played in the Spanish La Liga over five seasons. Features were gotten from FIFA 18 game database. They were able to achieve 71.63% accuracy with LR. Their ANN model achieved an accuracy of 63.1% from the match history database and 69.2 percent from the combination of the match history in the Team database. The major limitation was that the ANN model takes time and also large and sophisticated data is needed for this approach. [31] presents the use of the Google Prediction API to analyze prior cricket game information and predict cricket match outcomes. System of predictions works on the principle of machine learning which uses Regression Algorithms and Classifiers. The India team and other teams were used as a case study. Supervised machine learning was used. The proposed model yielded outcomes depending on earlier data supplied. The more the system of data is trained, the more superior outcomes. The test information was used to verify the forecast precision between India and other teams with a total 9 out of 10 games properly predicted. However, the limitation to this model was that it requires a lot of data for better accuracy, if the data is less, the accuracy is also less.

[8] predicted soccer match outcome based on the chances of bookmakers by using k-nearest method using the super league of Turkey competition 2015/16 season. The neighboring k-nearest algorithm was chosen as the assessment method, used the estimated results were compared with the bookmaker as a reference. Their model depends on the bookmarkers' odd. It was observed that there may be inconsistency in result if the bookmarker prediction is inaccurate. [29] applied fuzzy predictive classifier and proposed a model to predict the Brazilian football match in local league and. It forecasted 71 of 97 (73.2 percent precision) victories and 21 of 44 (47.73 percent precision) losses. The Maximum Likelihood classifier estimated, however, only 9 comes out of 49, a poor precision of 18.37 %. It is limited to first order uncertainty.

Having reviewed quite a number of related works to this research, the uniqueness of this work lies in the geographical area of application with the peculiarities of the game administrative pattern obtainable in Nigeria. One of the major gaps bridged by this work is the gap between the international league and local football leagues. In practical this work offers a more reliable approach that is best suitable to local league football result predication. In addition is the combination of methods as discussed in the subsequent section of this work.

3. METHODOLOGY

The method used in this research are encompassing, most of which are deductive in nature. This method aims at leveraging existing approaches and apply some of its specifics towards achieving the set objectives of this work. This work used an hybrid of k-NN model for data related to goals and Poisson distribution/regression mathematical model for other information. It centers on how effective prediction of the Nigeria football league can be achieved using fixtures from past matches. The importance of match fixtures when it comes to prediction becomes inevitable, event outcome of the matches played, among others. Algorithms were formulated to calculate the probabilities of match outcome. The methodology comprises of four (4) modules.

a) Data Collection: This enhances the model for effective and accurate prediction, outcome of sport results. Data were collected manually and electronically from valid local sport website peculiar to the Nigeria League in line with the objectives. The data in *.csv/.xlsx* format for easy manipulation during implementation. NPFL sport data were collected from the official website and other secondary sources as shown in Table 1.0. A modified database was created from the public version for the purpose of this research and consolidates information from four distinct

sources that to suite its method. Below is sample raw data from NPFL website for 2017 league showing some highlight of all clubs in season 2017.

b) Fixture Selection: This module determines the most important fixtures. Data mining tools were used i.e. a combination of machine intelligence with human perception. The data available are majorly based on the goals scored between two teams in the league as obtained from www.rsssf.com/tablesn/nigchamp.html [last accessed 1/10/2019]. As a result of this, the goal scored were analyzed and used in this prediction model. Several indicators for the home team and for the away team were created. Output result FTR (Full Time Result) will be (H-Home win, D-Draw, A-Away win). Hence, data were separated into training and testing.

c) Classifier/Algorithm Selection: The algorithm used in this research solely depend on the data fixture which is majorly the goals scored in the match between two teams. Therefore, the best classifier for this dataset uses k-NN model. Although, the goal scored were analyzed with the Poisson distribution and regression mathematical model. Predictions were classified into (Home Win- "1", Draw-"X" and Away Win "2") which was the scope of this research. For effective result, these classifiers were hybridized since hybrid model has proven effective in past research.

d) K-NN Algorithm Analysis

The k-nearest neighbor was implemented due to the nature of data available. The k-nearest method uses Euclidean distance, (x_1, x_2) which find the closest distance between two data sets and classified it based on the value of k.X_1 is the target distance value while X_2 is the data to compare. Generally, the Euclidean distance is given by;

$$||X_1 - X_2|| \tag{7.1}$$

For probability of the Home *Win*, *Draw,* and Away *Win* is given as

$$\varphi = \varphi_{w,}\, \varphi_D, \varphi_L \tag{7.2}$$

Therefore, the probability of the class is between 0 to 1. The percentage of the probability can be

obtained as $P_{Per} = \frac{\varphi_i^{-1}}{\sum_{i=1}^{n} \varphi_i^{-1}}$ (7.3)

Generally, using the Euclidean distance formula, a standard deviation can be obtained for each classifier using 7.4 Murphy.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_1 - \overline{x_2})^2} \tag{7.4}$$

Where 'n' represents the total number of team features to be determined i.e. PTS, W, D, L, GF, GA. Detailed analysis shown in table 1.0. Using dataset from 2009 to 2019 of NPFL, the Points were calculated for Points (PTS), Won (W), Drawn (D), Loss (L), Goals for (GA), Goal against (GA) and Goal Difference (GD). For each variable, the average performance for the points (PTS) were calculated. The average point for the team from 2009 to 2019 (10 years) is given by $\frac{1}{n}\sum_{i=1}^{n} x$ where x is the variable to determine and calculated. Same for other parameters. After this has been determined for each team the Euclidean is calculated.

The formula is given as $\sqrt{\sum(x_1 - x_2)^2}$ (7.5)

Where x_1 and x_2 is the Euclidean distance where x1 is the parameter of the first team and x2 is the parameter for the second team to be compared.

e) Analysis of Goal Scored with Poisson Distribution

With Poisson distribution and regression formula, odd in percentage of home win, draw and away win for each team are determined respectively. Previously identified equations are applied to predict the outcome of football matches using the goals scored during the match. The variable in the equations are be determined from the scope of the data. Football team is measured on the basis of the "**Attack/Defence Strength**" which is in this instance determined using the goals scored in general. Where the attack strength is the ability of a team to score and the defence strength is the ability for a team concede goal.

$$\text{Attack Strength of NPFL} = \frac{\text{total goal scored in the NPFL}}{\textit{total number of match played}}. \quad (8.0)$$

$$\text{Defence Strength of NPFL} = \frac{\text{total goal concedes in the NPFL}}{\textit{total number of match played}} \quad (8.1)$$

In this case, it involves two teams which is either **Home** or **Away** donated as (H or A). Equations 8.0 and 8.1 are used to predict the probability of the Home Win, Draw and Away Win of a team basically using the illustration of two teams **Team "A"** and **Team "B". Team A** is **Home** while **Team B** is **Away**

Attack Strength of **Team "A" -Home =**

$$\frac{\text{total goal scored in the Home}}{\textit{total number of Home Match}} \div \text{Attack Strength of NPFL} \quad (8.2)$$

Defence Strength of **Team "B"-Away =**

$$\frac{\text{total goal away conceded from Home}}{\textit{total number of Away Match}} \div \text{Attack Strength of NPFL} \quad (8.3)$$

Equations 8.0 – 8.3 are used to determine the possible goals that **Team "A" - Home** is likely to score in NPFL which is given as;

Attack Strength of **Team "A** × Defence Strength of **Team "B" ×** Attack Strength of NPFL - (8.4)

The same steps were applied to determine the goal **Team "B" –Away** is likely to score.

Attack Strength of **Team "B" -Away =**

$$\frac{\text{total goal scored Away Match}}{\textit{total number of Away Match}} \div \text{Defence Strength of NPFL} \quad (8.5)$$

Defence Strength of **Team "A"-Home =**

$$\frac{\text{total goal Home conceded}}{\textit{total number of Away Match}} \div \text{Defence Strength of NPFL} \quad (8.6)$$

The above equations can be used to determine the possible goal that **Team "B" - Away** is likely to score in NPFL which is given as

Attack Strength of **Team "B** × Defence Strength of **Team "A" ×** Defence Strength of NPFL (8.7).

The Attack/Defence Strength value is the "λ"for the Poisson equation as in equation 2.5 Pr(X = k) = $e^{-\lambda}\frac{\lambda^k}{k!}$ $\lambda > 0$

The value of k is the goal and it is a positive integer where k = 0,1,2,3… The distribution gives the probability of number of goals scored by individual team. Since both goal are not a dependent variable i.e. they do not depends on one another in the match.

Table 2: Probability Distribution of Goal Scored (K = 0 To 5)

Goals	0	1`	2	3	4	5
Team A	$P_A\{0\}$	$P_A\{1\}$	$P_A\{2\}$	$P_A\{3\}$	$P_A\{4\}$	$P_A\{5\}$
Team B	$P_B\{0\}$	$P_B\{1\}$	$P_B\{2\}$	$P_B\{3\}$	$P_B\{4\}$	$P_B\{5\}$

Source: https://www.pinnacle.com/en/betting-articles/Soccer/how-to-calculate-poisson-distribution/MD62MLXUMKMXZ6A8.

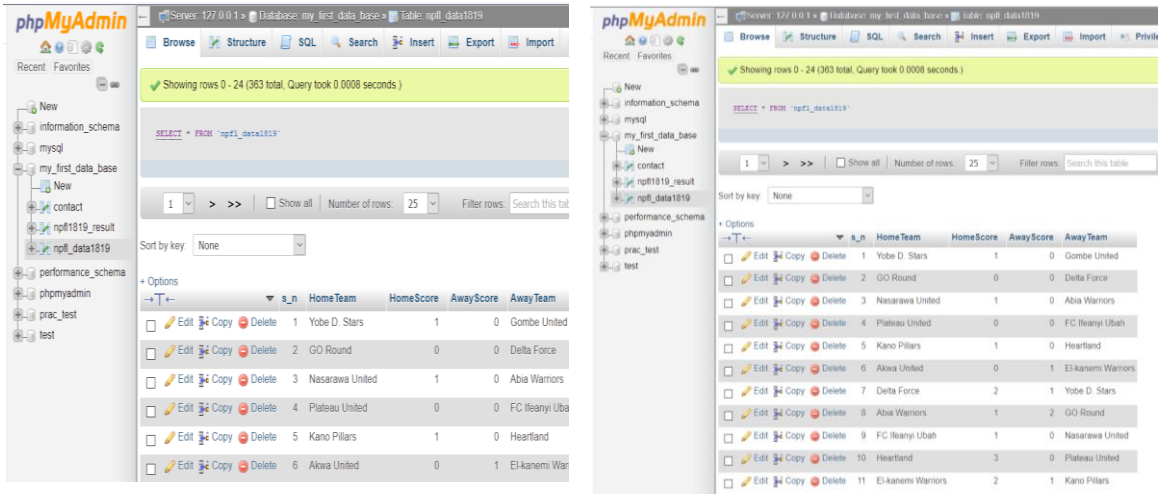
By implication, if we are to determination of a draw outcome between Team A and B, will be deduced from the summation of Joint probability of equal goal i.e. $\sum(A \text{ n } B)$ which is $P_A\{0\}$. $P_B\{0\}$ + $P_A\{1\}$. $P_B\{1\}$ + $P_A\{2\}$. $P_B\{2\}$……..+ $P_A\{5\}$ $P_B\{5\}$. This can be computed using the Poisson regression for various outcome of event that we are interested in.

$$\Pr(X = k) = \frac{\exp(-\exp(k'\beta))\exp k\prime\beta^{x}}{x!}. \quad (8.8)$$

4. IMPLEMENTATION

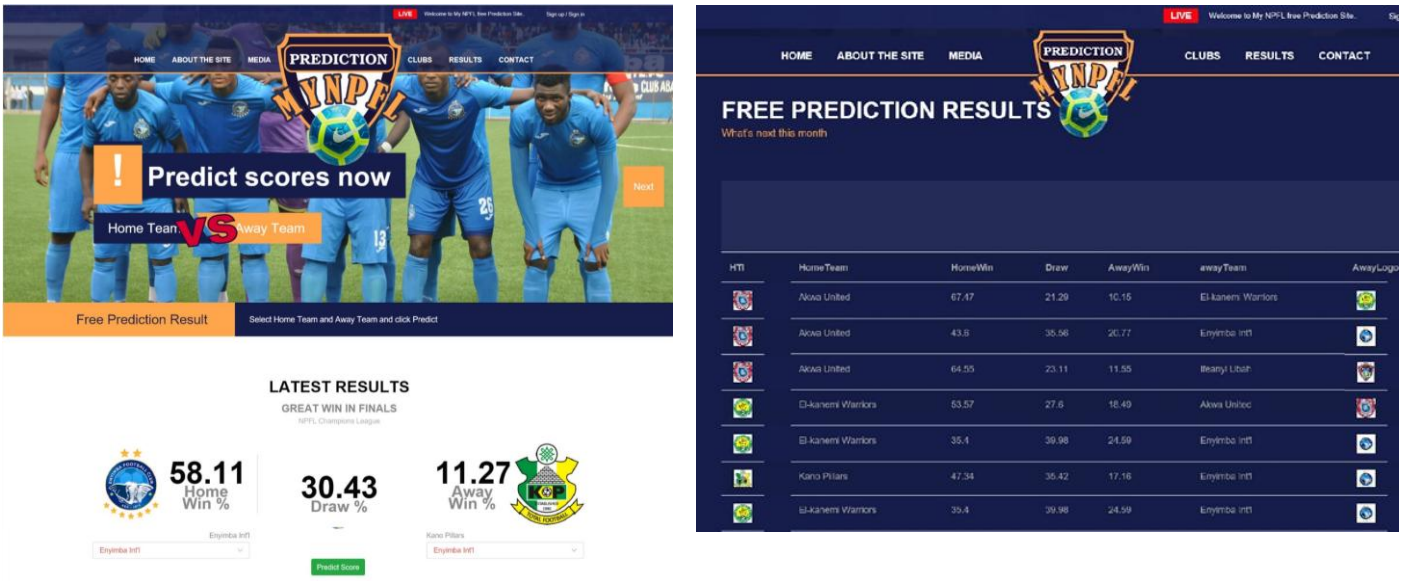
a) Tools and Database System Used

Popular data science tools were considered for this research i.e. Anaconda with python libraries. Python programming language Data was migrated into the database from the raw environment after which the model was used to build the classification. Various inbuilt libraries such as the sklearn, panda was used to build the model. The result of the model was stored in a csv/xlsx file and later displayed on the web through the database (MySQL) used for web analysis of result. A database driven interactive webpage was implemented as shown in Figures 2 and 3, and data stored in the database are displayed on the platform using the Django/Flash framework. These frameworks performed optimally with the python language and delivered a dynamic website. The interactive and dynamic nature of this web application make this work better and unique compared to the related work in this research.



(a) **(b)**

FIGURE 2: MySQL table of showing the database structure for experimentation dataset.



(a) **(b)**

FIGURE 3: Web application landing pager and the result prediction Interface.

4.1 Result Analysis of NPFL

Data of NPFL 2018/2019 contains 380 matches in the league season. The data frame in the columns are the team names and the number of scored goals for the home and away team as shown respectively. From the bar chart analysis in figure 10, the home teams occupies majority of the graph which means that home teams have more advantage of winning probability. Usually in football, away teams travel and may be more fatigue than their opponent in home teams. Also, the away team may not be familiar with the football pitch. This is generated from the implementation code. As a result, the highest frequency is home team winning, they tend to score more goals on average which means that home team scores more goals in the league which was because of the home advantage as stated earlier. It is emphasized the use of Poisson distribution since the goals scored describe the outcome of the match and thus the number of goals scored during the match is independent of the duration the match has commenced. This also gives the clue that home team secure more goals than the away team. Goals scored are better expressed independently by finding the goal distributions. Equation 6.0 and the Poisson model was used to generate the distribution. Figure 4; this shows predictions of the number of goals per match in NPFL 2018/2019 season. Furthermore, the actual data sample compared with the Poisson distribution model is useful for comparison match prediction.

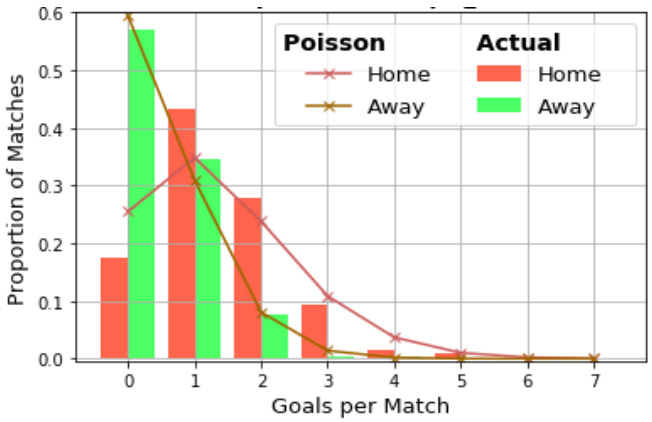


FIGURE 4: Poisson distribution of goals per match (NPFL 2018/2019 season).

4.2 Model Output Result

The test data consist of 24 teams each team plays 23 games for both the home and the away. The accuracy of the model mostly depend on the previous games. In this case, all the teams were treated and each team modeled with the Poisson regression and distribution. Table 3.0 show the model output and observation for each value generated by the model. The attack and defense strength were automatically calculated by the model with respect to the equations stated in the methodology vis-à-vis implementation codes.

TABLE 3: Model information output.

Generalized Linear Model Regression Results			
Dep. Variable:	goals	**No. Observations:**	726
Model:	GLM	**Df Residuals:**	678
Model Family:	Poisson	**Df Model:**	47
Link Function:	log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-814.66
Date:	Thur, 14 Nov 2019	**Deviance:**	591.40
Time:	09:33:38	**Pearson chi2:**	528.
No. Iterations:	5	**Covariance Type:**	nonrobust

Figure 5 shows the Poisson model for the number of goal per match between Akwa United and Eyimba International for both the Home and Away match.

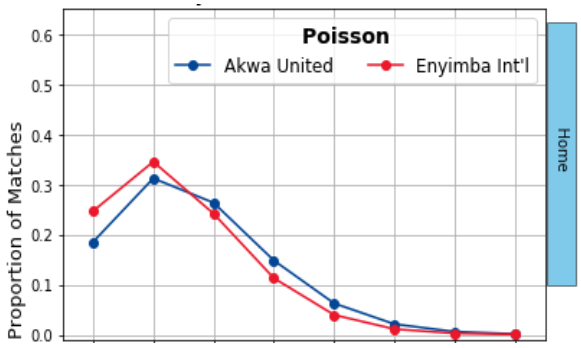


FIGURE 5: Poisson model for two (2) football teams.

4.3 Model Evaluation and Comparison with Bookmarkers Odds

The dataset was shared into training and testing; the testing dataset was set aside for further evaluation of the model. The NPFL super six fixtures were used to carry out the evaluation. The dataset of fixture used and the date the matches were played, the actual win, and the predicted percentage Out of the 15 matches, the model correctly predicted 11 matches correctly which is a percentage of 73.33%. Despite other external factors such as signing of new players, changing of team coach and other social sentiment which was not included in this model. In Table 4.0, the model compared with an online bookmarker's odd. This will help to evaluate the consistency of the model result with the bookmarker's odd and the theory of the mathematical equations. The online bookmarker odds were retrieved from https://hintwise.com/league/Nigeria-NPFL. Table 4.2 shows the comparison between the model and the online bookmarker's odd which was retrieved from www.hintwise.com/league/Nigeria-NPFL as at November 1st 2019.

TABLE 4: Shows the evaluation between the Model Result and the Bookmarker's Odd Using Dataset of NPFL Super Six Fixture 2018/2019.

Srn	Home vs Away Team	Model Predicted Result (%)Home Win(H), Draw(D),Away Win(A)			Bookmarker's Odd in (%) Home Win(H),Draw(D), Away Win(A)			Model Accuracy
1	Enyimba Int'l FC vs Rangers Int'l FC	52.71	35.81	11.49	62.30	25.60	13.10	***True***
2	Kano Pillars Int'l vs Akwa United FC	68.57	20.74	10.69	65.40	24.10	10.50	***True***
3	FC Ifeanyi Ubah vs Lobi Stars FC	49.89	31.45	18.53	54.34	35.23	24.32	***True***
4	Kano Pillars Int'l vs Enyimba Int'l	47.89	35.42	17.17	73.53	28.43	11.65.	***True***
5	Rangers Int'l FC vs Lobi Stars FC	57.93	29.41	12.44	61.00	22.50	16.50	***True***
6	Akwa United FC vs FC Ifeanyi Ubah	64.55	23.12	11.56	59.50	27.60	18.76	***True***
7	Akwa United FC vs Rangers Int'l FC	51.93	29.94	17.94	59.5	27.6	12.9	***True***
8	Lobi Stars FC vs Enyimba FC	43.30	38.43	18.24	59.80	26.41	14.18	***True***
9	Kano Pillars FC vs FC Ifeanyi Ubah	68.27	21.69	9.11	32.4	43.4	23.4.	***True***
10	Enyimba Int'l FC vs FC Ifeanyi Ubah	64.52	27.54	7.65	52.10	26.09	21.00	***True***
11	Akwa United vs Lobi Stars FC	57.23	26.54	15.82	37.00	39.00	24.00	***True***
12	Rangers Int'l vs Kano Pillars Int'l	57.58	28.47	13.69	67.60	22.50	9.90	***True***
13	Lobi Stars FC vs Kano Pillars Int'l	56.44	27.70	15.55	37.00	40.00	23.00	***True***
14	Enyimba Int'l FC vs Akwa United	64.69	25.95	8.96	40.90	31.00	27.70	***True***
15	FC Ifeanyi Ubah vs Rangers Int'l	44.84	34.71	20.40	8.50	27.10	64.4	***False***

Comparing the online bookmaker probabilities with the model output, 14 matches out of 15 were similar which gives 93.33% similarity with other online bookmakers. Using kNN to categorize Home win, Draw, and Away win, the difference in the probabilities of the model and the online book-maker is as a result of some unforeseen contingencies.

5. CONCLUSION

Overall, Football prediction has becoming captivating, as seen in the literatures of different authors from year 1997 to 2021. However, there were difficulties in predicting 100% accurately of outcome of a match especially with the peculiarities and other sentimental factors that come to play in the Nigerian football league. The practical implication of this work, include but not limited to the numerous benefit it pose to make Nigerian football league catch up with their international counterparts, albeit the target audience are football fans and various football club lovers but the predicting the results correctly, larger audience will be reached with this league therefore making it more popular in the international community. The difficulties encountered by specialist in getting hired for international league will also be minimized with the result of this work. Recall, various researchers have used various methods and algorithms to predict the outcome of sport matches. Some algorithms look complex for sport like chess, javelin or even tennis game as it is only one major factor that determines the winner or loser which is the ability of the player. In this research, only team games were considered.

Data mining tool have been maximized for this purpose based on its peculiarities for event prediction. Several predicting models like ML, ANN, BN, LR, SVM, and lazy techniques have been adopted in this work. Several literatures were reviewed vis-à-vis their major drawbacks especially in data availability and evaluation results. This research method allows high prediction accuracy meaning using comprehensive statistics which that gives room for comparison of results with previous studies and available data.

Further study would focus more on sentiment and other factors such as fatigue, change of player, weather conditions, and coaches for a better and reliable model.

6. ACKNOWLEDGEMENT

The authors of this work would like to appreciate NPFL for making available historical data for the public, sport newspapers and magazines among other public sport database used for testing and evaluation purpose in this research.

7. REFERENCES

[1] Adebisi, J. A., Abdulsalam, K. A. and Fawaz O. (2021): IOT Smart Home: Implementation of a real-time Energy Monitoring Pressing Iron. Being a paper International Conference of the Nigeria Computer Society.

[2] Bunker, R. P., and Thabtah, F. (2019). A machine learning framework for sport result prediction. Applied computing and informatics, 15(1), 27-33.

[3] Angelini, G., and De Angelis, L. (2017). PARX model for football match predictions: PARX model for football matches predictions. Journal of Forecasting, 36(7), 795–807.

[4] Baio, G., & Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. Journal of Applied Statistics, 37(2), 253–264.

[5] Buursma, D. (2010). Predicting sports events from past results. 14th Twente Student Conference on IT, 21. Holland.

[6] Constantinou, A. C., Fenton, N. E., and Neil, M. (2013). Profiting from an inefficient Association Football gambling market: Prediction, Risk and Uncertainty using Bayesian networks. Knowledge-Based Systems, 50, 60–86.

[7] Dixon, M. J., and Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. Journal of the Royal Statistical Society: Series C (Applied Statistics), 46(2), 265–280.

[8] Doyle, P. G., Grinstead, C. M., and Snell, J. L. (2006). Grinstead and Snell's Introduction to Probability.

[9] Esme, E., and Kiran, M. S. (2018). Prediction of Football Match Outcomes Based on Bookmaker Odds by Using k-Nearest Neighbor Algorithm. International Journal of Machine Learning and Computing, 8(1)

[10] Forrest, D., & Simmons, R. (2002). Outcome uncertainty and attendance demand in sport: The case of English soccer. Journal of the Royal Statistical Society: Series D (The Statistician), 51(2), 229–241.

[11] Graham, I., & Stott, H. (2008). Predicting bookmaker odds and efficiency for UK football. Applied Economics, 40(1), 99–109.

[12] Haghighat, M., Rastegari, H., and Nourafza, N. (2013). A review of data mining techniques for result prediction in sports. Advances in Computer Science: An International Journal, 2(5), 7–12.

[13] Haight, F. A. (1967). Handbook of the Poisson distribution.

[14] Home Page—Nigeria Professional Football League. (n.d.). Retrieved October 1, 2019, from https://npfl.ng/

[15] Igiri, C. P., and Nwachukwu, E. O. (2014). An improved prediction system for football a match result. IOSR Journal of Engineering (IOSRJEN), 4(12), 12–20.

[16] Joseph, A., Fenton, N. E., and Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. Knowledge-Based Systems, 19(7), 544–553.

[17] Karlis, D., & Ntzoufras, I. (2011). Robust fitting of football prediction models. IMA Journal of Management Mathematics, 22(2), 171–182.

[18] Kuhlman, D. (2009). A python book: Beginning python, advanced python, and python exercises. Dave Kuhlman Lutz.

[19] Lessmann, S., Sung, M.-C., and Johnson, J. E. V. (2010). Alternative methods of predicting competitive events: An application in horserace betting markets. International Journal of Forecasting, 26(3), 518–536.

[20] Leung, C. K., and Joseph, K. W. (2014). Sports data mining: Predicting results for the college football games. Procedia Computer Science, 35, 710–719.

[21] Mathworks (n.d.). machinelearning_supervisedunsupervised.png. Retrieved from <https://de.mathworks.com/help/stats/machinelearning_supervisedunsupervised.png>.

[22] Miljkovic, D., Gajic, L., Kovacevic, A., and Konjovic, Z. (2010). The use of data mining for basketball matches outcomes prediction. IEEE 8th International Symposium on Intelligent Systems and Informatics, 309–312.

[23] Müller, A. C., and Guido, S. (2016). Introduction to machine learning with Python: A guide for data scientists. O'Reilly Media, Inc.

[24] Murphy, K. P. (2012). Machine learning: A probabilistic perspective. MIT press.

[25] NPFL Data for League 2017. (n.d.). Retrieved from https://npfl.ng/league-table/.

[26] Owramipur, F., Eskandarian, P., & Mozneb, F. S. (2013). Football result prediction with Bayesian network in Spanish League-Barcelona team. International Journal of Computer Theory and Engineering, 5(5), 812.

[27] Passi, K., and Pandey, N. (2018). Increased Prediction Accuracy in the Game of Cricket using Machine Learning. ArXiv Preprint ArXiv:1804.04226.

[28] Purucker, M. C. (1996). Neural network quarterbacking. IEEE Potentials, 15(3), 9–15.

[29] Razali, N., Mustapha, A., Utama, S., and Din, R. (2018). A Review on Football Match Outcome Prediction using Bayesian Networks. Journal of Physics: Conference Series, 1020, 012004.

[30] Reed, D., and O'Donoghue, P. (2005). Development and application of computer-based prediction methods. International Journal of Performance Analysis in Sport, 5(3), 12–28.

[31] Şahin, M., and Uçar, M. (2020). Prediction of sports attendance: A comparative analysis. Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology, 1754337120983135.

[32] Sharon Andrews & Mark Sheppard (2020):Software Architecture Erosion: Impacts, Causes, and Management. International Journal of Computer Science and Security (IJCSS), Volume (14) : Issue (2) : 2020

[33] Tavares, A. T. (2018). Predicting Results of Brazilian Soccer League Matches. University of Wisconsin-Madison.

[34] Tina T. (2020): Social Big Data: Techniques and Recent Applications. International Journal of Computer Science and Security (IJCSS), Volume (14): Issue(5): 2020.

[35] Ujwal, U. J., Antony, P. J., and Sachin, D. N. (2018). Predictive Analysis of Sports Data using Google Prediction API. International Journal of Applied Engineering Research, 13(5), 2814–2816.

[36] Van Eetvelde, H., Mendonça, L. D., Ley, C., Seil, R., and Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: a systematic review. Journal of experimental orthopaedics, 8(1), 1-15.

[37] Wilkens, S. (2020). Sports prediction and betting models in the machine learning age: The case of tennis. Journal of Sports Analytics, (Preprint), 1-19.

[38] Yang, S., Luo, L., and Tan, B. (2021). Research on Sports Performance Prediction Based on BP Neural Network. Mobile Information Systems, 2021.

[39] Zaveri, N., Tiwari, S., Shinde, P., Shah, U., and Teli, L. K. (2018). Prediction of Football Match Score and Decision Making Process. International Journal on Recent and Innovation Trends in Computing and Communication, 6(2), 162–165.

[40] Zdravevski, E., and Kulakov, A. (2009). System for Prediction of the Winner in a Sports Game. International Conference on ICT Innovations, 55–63. Springer.