Predicting English Premium League (Epl) Game Results Using Machine Learning

[1]Shrikant Burje, [2]Sandeep Bhad

[1]Asso.Professor, [2]Asst. Professor

Rungta College of Engineering and Technology, Bhilai (CG), India.

Abstract— Nowadays, most people just refer to their own intuition or the opinions of a few experts when they need to predict the outcome of a football game. However, because artificial intelligence is so good at analysing vast amounts of data, it is increasingly utilised to predict outcomes rather than use personal experiences to develop toward accuracy. Common Machine Learning applications in sports research include predicting player injuries and taking appropriate action, evaluating potential talent or market value, and predicting individual or team performance. In this research, we use fake information and AI algorithms to predict the outcomes of soccer matches in the English Premier League (EPL). The outcomes of each game will be predicted using historical data and game results. The data is collected between 1993 and 2021. Kaggle.com was consulted for the data. The information includes a few distinct components, such as the dates, the two teams, the half- and full-time scores, and various game-related highlights. To get the highest level of precision on the test information, the highlights will be evaluated using direct relapse models and information cleaning techniques. This project's main goal is to examine various Machine Learning techniques for predicting the score and outcome of football games using in-game match events rather than the number of goals scored by each team. We will research novel model plan theories and evaluate the performance of our models in comparison to benchmarks.

Keywords— Machine Learning, EPL, Python, Regression, Neural Network.

I. INTRODUCTION

During the 2010 World Cup, there were numerous displays of virtuosity, from Thomas Muller to Andrew Iniesta, but none were quite as astonishing as Paul the Octopus. This ocean dweller correctly predicted the winner of a contest each of the eight times he was tested. This accuracy sets itself clearly apart from one of our colleagues' predictions for the World Cup, who was accurate just sometimes. Due to our love of the sport and to spare ourselves the embarrassment of being outperformed by an octopus, we have decided to try and predict soccer match results. This has verifiable uses in betting, improving training, and reporting. This has verifiable uses in betting, improving training, and reporting. We chose the English Premier League (EPL), the league with the highest TV viewers in the world (4.7 billion people), out of all the leagues we could have chosen. In our project, we will discuss past research that came before our investigation, including selection, an analysis of how various models were implemented, and analysis of our results. Sports analysis is the use of reliable data and advanced metrics to assess performance, make decisions about execution and results, and gain an advantage over rivals.

As a result, several plans and computations have been sent. Clubs use sophisticated equipment and software (such as GPS tracking systems) to gather and analyse data produced by players during practise sessions and competitions. They analyse this information to be used for both long-term association improvement and independent short-term direction. For betting companies, a thorough analysis of all available information is also crucial. Last but not least, the latest discoveries and what they signify for football greatly stimulate supporters. Numerous people have always followed football closely as it is one of the most well-known sports in the world. Recent times have seen the gathering of new types of data for specific games in various countries, such as detailed data collecting for each shot or pass made in. New types of data have recently been obtained for several games in various countries, such as detailed data recalling data for each shot or pass performed during a match. With a wide range of possible uses and applications, the diversity of this information has placed data science at the forefront of the football industry: [1]

- Match system, tactics, and research
- Differentiating players' playing techniques A player's acquisition, evaluation, and group expenditure putting plans and focus together
- Injury anticipation and prevention using test findings and employment
- Board execution and expectations
- Association table expectation and match result

- Planning and organising competitions
- Estimating wagering chances [2]

Inspiration

The ability to evaluate a team's performance in games and use that data to try to predict the outcome of future games based on that information is a particularly important aspect of data science in football. Sports game outcomes can be unpredictable, with surprises frequently occurring. Football in particular offers an intriguing paradigm because games are predetermined in length (unlike racket sports like tennis, where play continues until one player wins). Additionally, it only features one type of scoring event—goals—which can occur countless times throughout a game and are typically worth one point, as opposed to a sport like rugby where different scoring occasions result in different numbers of points.

Problem Statement

The impacts of the game are more eagerly anticipated by Head League fans. They are more motivated to know the result before to the game. The main challenge, in any event, is to determine the precise outcome of the expectation between the groups.

Goals

The main objective of Premier League Game Result Prediction is to:

- Predicting the outcome of each round of the group of the head association.
- To perform Back Propagation calculations in order to create and test the model.

Scope

- In a sense, the English Premier League will be associated with this prognosis [3].

II. RELATED WORK

Barcelona, a Spanish team, used a Bayesian network to predict its score. They studied the link and model likelihood within the specified area. Data was acquired from reliable websites that provided analysis of football. They also made use of two non-mental and mental factors for the match forecast in order to predict the final result [1]. [10] used Rapid Miner as an information mining tool in conjunction with Artificial Neural Network (ANN) and calculated relapse (LR) algorithms, with results of forecast precision of 85 and 93 percent, respectively. They made a comparison between the present framework and their framework, finding that their framework was twice as precise as the current framework. They considered particular variables that affect how the team performed after the game, such as the effect of home advantage on the team's performance, the effect of critical participant injuries on the team's performance, and the effect of the outer cup on association performance. After observing these components, they concluded that the group will also experience the effects of the particular components. 2 used a three multi-facet perception idea to predict the outcome, adding home advantage and positioning contributions along with factors such as each team's most recent three encounters with the other team and its last five matches, with the idea that the more factors considered, the more likely the expectation will be accurate [2] [3]. [11] handled and made fictitious predictions about Spurs' results using the Bayesian organisational strategy. The MC4 Learner, Nave Bayesian Network, Hugin Bayesian Network, Expert Bayesian Organization, and KNN (K Nearest Neighbor) calculation were among the specific Bayesian organization characteristics they expected the dataset to have. They utilized the forecasting and comprehending capacities of AI, which were two unique advantages. The student in MC4 is aware of the factors that most significantly affect the game's outcome. Their interrelation and impact on the result of the game are shown. This is a significant upgrade that is on par with the genuine game. For the uninformed Bayesian factor, a model is predefined; no model is constructed in this sense.

In [12] reviewed previous and ongoing soccer expectation research projects, organizing their techniques and conclusions. It assumes that a game would deliver black boxes afterward, overlooking the raucous but incredibly intricate cycles that go into each shot and goal. More precise models of in-game cycles may be put together as XY information develops under explorer's supervision. As a result, new questions and lines of inquiry will become available, hopefully leading to a deeper understanding of the game itself [8]. Additionally, it has begun the process of determining the general value of these types of information in forecasting and research. chosen the crucial characteristics for a component, they are Include a concept from the opposing perspective on the problem. Be aware of a group's new kind and demonstrate the advantage to the home. They used the following three tactics: First Approach: Their main strategy consisted on applying Multinomial Logistic Regression. Instead of using the average over the prior k matches, they took into account the exhibition measures obtained from the current match. They appeared at the component vector using KPP during testing, where they expected the match result of group A versus group B [11-13].

III. MATERIALS AND METHODS

We will introduce a summary of well-known controlled machine learning processes in this section for its order and relapse subsets. The assignment of learning a skill that directs input information to yield information based on model contribution to-yield sets is known as administered learning. Relapse occurs when the yield is a stable number, but characterisation occurs when the output is a categorization. We are only interested in the machine learning administered learning scene since in our circumstance we need to predict the result classification (home win/draw/away win) or the amount of objectives scored by a group (constant number). In Fig. 1, we summarised a few well-known administered learning strategies that we will now discuss in more depth.

A. Decision tree

A well-known machine learning technique is decision trees, which connect input considerations (which are handled in the branches and hubs of the tree) with yield esteem (which is addressed in the tree's leaves). Trees can be used to solve grouping problems by producing a class name or solving relapse problems by producing a real number. Numerous computations, including the most well-known CART or ID3 choice tree calculations, can be used to fit choice trees. These algorithms combine voracious searching and pruning to create a tree that matches the data and adds it to new input/yield sets.

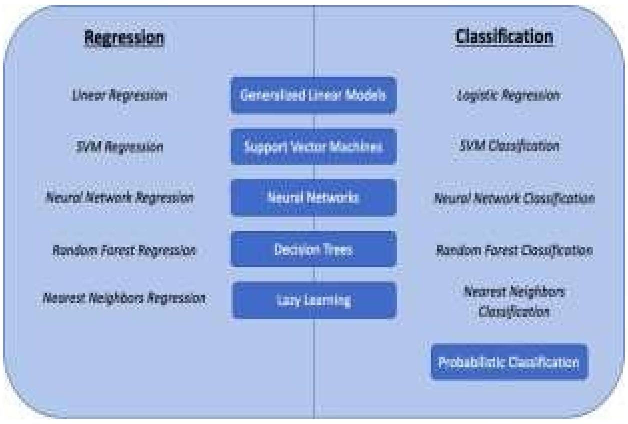

Regression
Classification
Linear Regression
Generalized Linear Models
Logistic Regression
SVM Regression
Support Vector Machines
SVM Classification
Neural Network Regression
Neural Networks
Neural Network Classification
Random Forest Regression
Decision Trees
Random Forest Classification
Nearest Neighbors Regression
Lazy Learning
Nearest Neighbors Classification
Probabilistic Classification

Figure 1: Supervised Algorithm

B. Support Vector Machines

Machine learning models for both categorization and relapse include Support Vector Machines (SVMs). An SVM model addresses the preparation information as focuses in space, resulting in a hyperplane (see

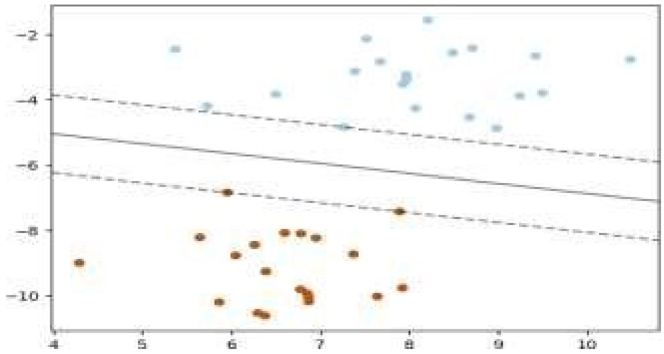


Figure 2: SVM

Fig. 2) that partitions models into different classes beyond what many people would think is feasible from the nearest information point. Similar to the preparation information, new information sources are planned and given classification names (which side of the hyperplane). When the information isn't clearly distinguishable, the part method can be used to organise the data into high-dimensional component spaces and locate an appropriate high-dimensional hyperplane by using several potential portion capacities like Radial Basis Functions (RBF) or polynomial capacities.

C. Neural Network models

Artificial neural networks (ANNs), also known as neural networks, are frameworks that rely on a variety of hubs (neurons) that simulate the connections between neurons in the human brain on an algorithmic level. Each neuron has the ability to receive a signal from other neurons and pass it on to other neurons. An edge connecting two neurons has a weight assigned to it that models the importance of this neuron's contribution to the output of other neurons. Information layers, which contain one neuron for each input variable in the model, a yield layer, which is made up of one neuron and provides the grouping or relapse outcome, and numerous secret layers between the two, which contain a varied number of neurons, make up a neural organisation [10].

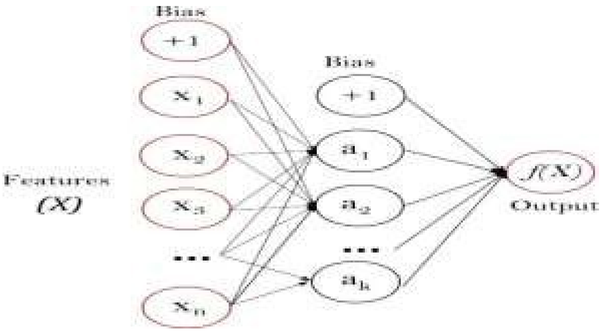


Figure 3: Neural network architecture

IV. IMPLEMENTATION

Regression techniques and Artificial Neural Networks (ANN) are two data-mining methods that can be applied. I'll start by compiling the previous match results along with each group's and each team's performance records. Then, the functions such as Home and Away Goal Difference, Points, Attack and Defense Skills, which aren't always needed for the techniques, will be extracted from the accumulated information. Once all the relevant information has been gathered, a collective database can be created and stored in an MS Excel spreadsheet. The technique that must be used in conjunction with creating the forecast is shown in Figure 4. The statistics are first accumulated and then normalised with the sigmoid feature. The normalised data is then trained using a neural network, and the educated data is assessed for prediction accuracy. The forecast can then be carried out following the testing.

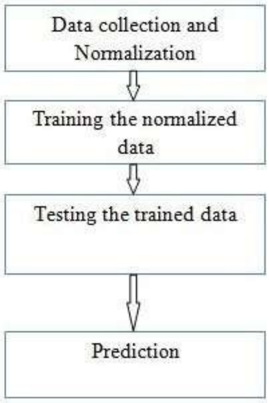

Data collection and Normalization
Training the normalized data
Testing the trained data
Prediction

Figure 4: Process flow

The information is gathered from the official Premier League website, www.Premierleague.Com, and http://www.Soccer-records.Co.United Kingdom/england. The optimal league began in 1993, and there are typically 9000 records of the league dating back to that year. The statistics that can be gathered include pointless parameters that should be removed. Only the relevant data has been accumulated. The accumulated records are then normalised. The procedure used to normalise the data is described below: [17].

$$\text{Normalized cost} = (\text{xi} - \min(\text{x})) / (\max(\text{x}) - \min(\text{x})) \quad (1)$$

The following method changed into used to generate the attitudelon—23features

A simplified diagram of the database structure and functions is supplied in Fig5 . We will now present the exceptional tables and capabilities that we've in our database and that we are able to use in our models: Matches desk, ID, League ID, Season, Date, Home crew ID, Away team IDL,

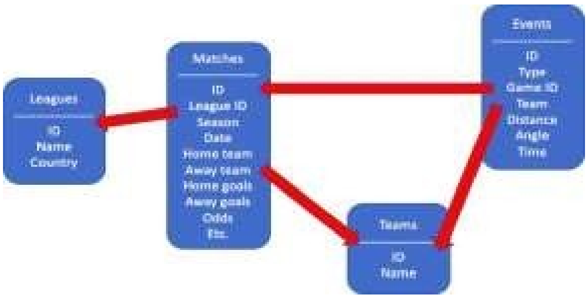


Figure 5: pre-processing

Training: The data was gathered from a reliable best league website, kaggle.com, and 80 percent of the information was chosen for this program's educational purposes. As an input and output parameter, the normalised statistics have been split into two parts[18].

Testing: 20 percent of the 9000 available fact units have been chosen for this application's checking-out purpose out of the total. Two items were created from the normalised records as an entry parameter and an output parameter. Because it is supervised learning, the result and input are previously known.

.

Statistics and Data Set Analysis: Despite Cooper receiving an early dismissal in Game Week 31, Leeds United finished Man City's 24 match unbeaten streak in the League. With 74 points, Man City now holds the top spot in the standings. They are in second place, 11 points ahead of their local rivals Manchester United. Spurs were pushed to seventh place in the association after Man United crushed them by 13 points. The top-of-the-table matchup between Westham United and Leicester City was predicted to be the most fiercely contested game of the week. The Hammers prevailed in the game, handing the Foxes their second straight defeat. The major four remaining components, however, are unaffected, and just one point of the hole is affected by the Hammers.

Using feature engineering, handle missing values and pre-process data

Clarification of the Read CSV Information Interface Utilize Pandas to browse and create URLs for information API summary Pandas.read csv() is typically used to read csv data: pandas.read csv(filepath or buffer, sep = ",", delimiter = None) Filepath or buffer: document direction Sep: chooses the separator; a comma is used by default. Delimitation: delimitation, optional delimitation (on the off chance that the boundary is determined, sep is invalid).

Feature Engineering for Imbalanced Data Handling: When preparing the Pandas dataframe for machine learning calculations, we use a few comfort tasks. Prep dataframe is meant to give the designer a single location to decide which highlights (segments) to include in the ML calculation. split data divides an organised data frame into creating and testing informational indexes. Additionally, I shall perform lopsided dataset operations here even though I am mostly using unbalanced learning computations. The fundamental ability to empower expectation on a predictive dataset is split pred data. Common Information capabilities are for some future period. Using Connection Matrix, it is possible to separate the highlights.

Defined as win ratio head2head(train df,test df), this function calculates a team's overall win percentage against each opponent. Internal for circle repeats over every foe of that group while external for circle repeats over a list of groups. Every opponent's win percentage is calculated and recorded in a dict. The dataframe chunk comprising the games for the first group is attached to a rundown and Dict is planned to another part. All of the dataframe's component parts are joined back together once the two circles have closed.

def add league position(df1,df2): This function adds each opponent's league position from the previous season as an element from two dataframes, the matches DF and the association table DF. Since we are interested in a group's performance from prior years, two rundown appreciations are used to create a season.

V. RESULT

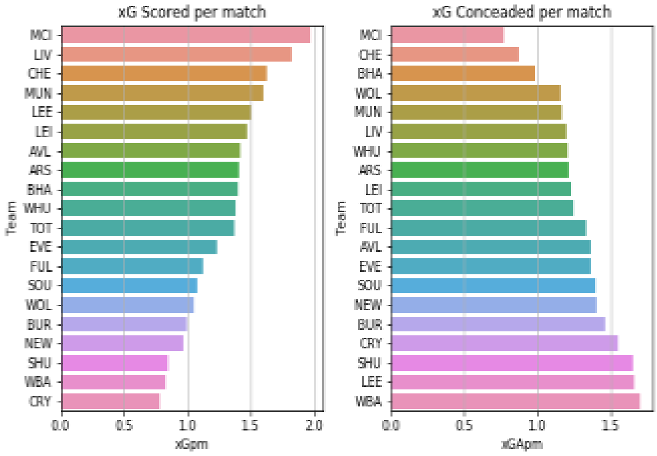


Figure 6. xG scored and conceded per game

Figure 6 shows As seen in the above graphic, groups can be divided into 4 quadrants based on xG Scored and xG Conceded. The average xG scored every game is displayed on the even spotted line. Groups above the evenly spaced line are strong attacking teams, while those below are weak in attack. The upward dabbed line displays the average number of xG given up per game. The groups on the left have a strong defence, while the groupings to the right have a weak ones [20].

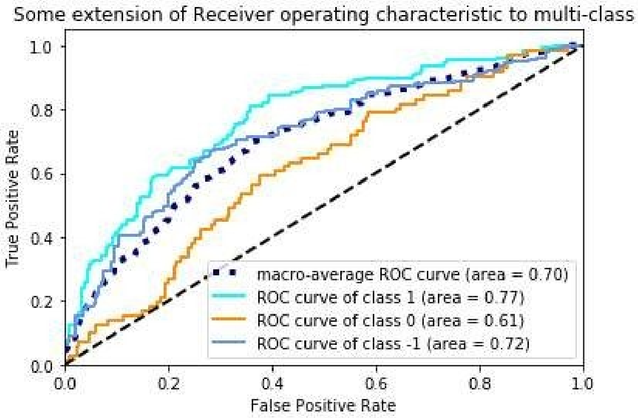
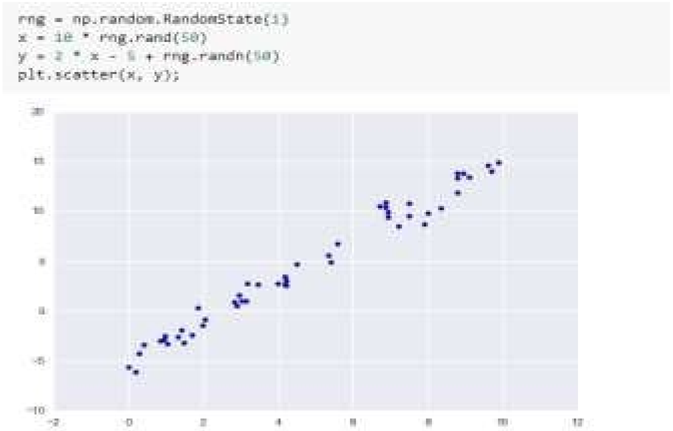


Figure 7 : ROC curve correctly predicted area

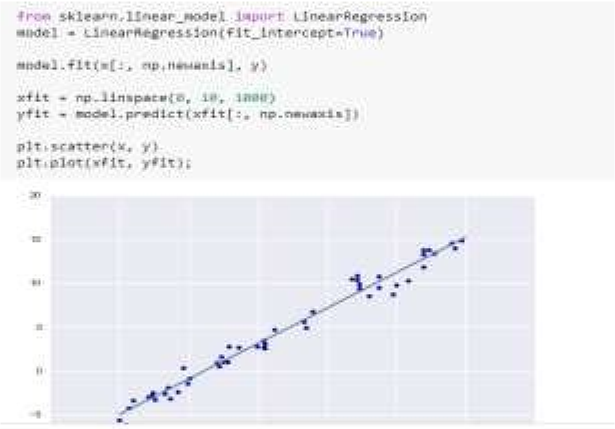
Although ROC Curves were initially developed for paired order, we may still use them for multiclass problems for certain alterations. By obtaining ROC Curves for each of the classes, we can then obtain a'signify' ROC bend.

Regularisation

The addition of basic functions to our linear regression model greatly increases the model's flexibility, but it can also quickly result in over-fitting (refer back to Hyperparameters and Model Validation for a discussion of this). For instance, if we select an excessive number of Gaussian basis functions, the outcomes are less than desirable: The larger C is analogous to the penalty slack variable, and it is assumed that the leeway variable will typically be fully combined for the practise set and be close to 0, meaning that the punishment for misclassification increments. As a result, the preparation set's precision is excellent, although there is speculation. fragile ability. Resistance is allowed, the C worth is low, the punishment for misclassification is reduced, and they are,



```
rng = np.random.RandomState(1)
x = 10 * rng.rand(50)
y = 2 * x - 5 + rng.randn(50)
plt.scatter(x, y);
```



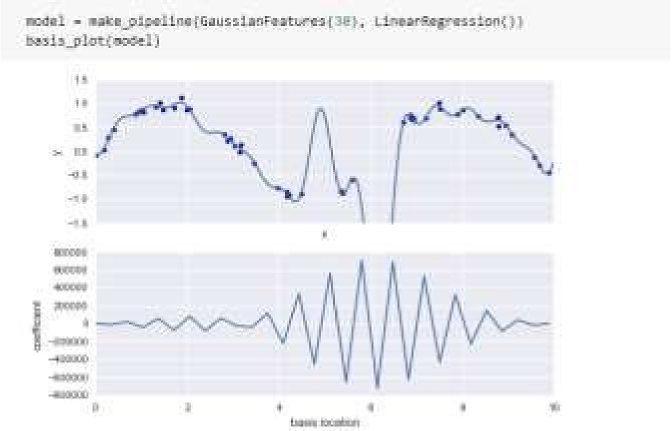
```
from sklearn.linear_model import LinearRegression
model = LinearRegression(fit_intercept=True)

model.fit(x[:, np.newaxis], y)

xfit = np.linspace(0, 10, 1000)
yfit = model.predict(xfit[:, np.newaxis])

plt.scatter(x, y)
plt.plot(xfit, yfit);
```

Figure 8: Prediction Of Variable Result Of Linear Regression



```
model = make_pipeline(GaussianFeatures(30), LinearRegression())
basis_plot(model)
```

Figure 9: Hyper Parameter Tuning Result Of Linear Regression

TABLE II CLASSIFIER COMPARISON

Method		
Measure	**Value**	**Derivations**
Sensitivity	0.4247	TPR = TP / (TP + FN)
Specificity	0.5510	SPC = TN / (FP + TN)
Precision	0.4133	PPV = TP / (TP + FP)
Negative Predictive Value	0.5625	NPV = TN / (TN + FN)

False Positive Rate	0.4490	FPR = FP / (FP + TN)

IN Table 2, different parameter of confusion matrix is calculated fir SVM and liner regression model. In any case, it should be clear that this approach is unreasonable for calculating irregular yields. The reason is that each tree will decide whether the result for each of the three focuses is 0 or 1, and the yield might be either (0, 0, 0) or (1, 1, 1), which is absurd. We chose group size to be 80, learning rate base to be 0.30, learning rate rot to be 0.999, regulization rate to be 0.0020, preparing ventures to be 400, and moving normal rot to be 0.85 for convolution neural organisation calculations. The best result will result from this, and the appropriate rate is fluctuating between 0.533 and 0.574, indicating that the projection isn't very steady. Resistance is allowed, misclassification punishment is scaled back, and they

VI. CONCLUSIONS

The ROC bends reveal that "Draw" was the class that was hardest to order. Despite achieving high characterizing scores in each class, we were unable to acquire a truly exceptional increase in the initial wager (our best was an increase of 7.155263157894765 percent). This may be the result of various causes, including: [24] - Betting establishments are incredibly good at determining probabilities, making it extremely challenging to defeat them. Since the goal of this activity was to determine whether one could judge high probabilities for matched results by selecting a small number of factors they may perceive to be relevant, we did not use the numbers they have as elements. For models, groups are differently persuaded for different matches, which may affect match outcome. For instance, if a group believes it can win the championship by dominating the next match, they are likely to give their all in that match and succeed. Then, a component could be created that associates with inspiration. We used bad wagering strategies: We ultimately settled on the most straightforward strategy, which involves simply betting on what the classifier says. This approach probably won't be the only one used; perhaps some others will be suggested that use the likelihood of each class to try to make a better wager.

REFERENCES

[1] G. Fialho, A. Manhães, and J. P. Teixeira, “ScienceDirect ScienceDirect Predicting Sports Results with Artificial Intelligence – A Proposal Predicting Sports Results with Artificial Intelligence – A Proposal Framework for Soccer Games Framework for Soccer Games CENTERIS - International Conference on ENTERprise Information Systems /,” Procedia Comput. Sci., vol. 164, pp. 131–136, 2019, doi: 10.1016/j.procs.2019.12.164.

[2] I. P. L. Matches and U. Machine, “Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning.”

[3] A. P. Report, “Premier league game result prediction,” 2016.

[4] “12445633 (4).pdf.” .

[5] H. Q. Tran, “applied sciences Improved Visible Light-Based Indoor Positioning System Using Machine Learning Classification and Regression,” 2019, doi: 10.3390/app9061048.

[6] S. Lee, “applied sciences The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction,” 2019, doi: 10.3390/app9153093.

[7] B. Ulmer and M. Fernandez, “Predicting Soccer Match Results in the English Premier League,” 2013.

[8] S. Cell, C. A. Preliminary, N. Fujima, Y. Shimizu, D. Yoshida, and S. Kano, “Machine-Learning-Based Prediction of Treatment Outcomes Using MR Imaging-Derived Quantitative Tumor Information in Patients with Sinonasal,” pp. 1–12, 2019.

[9] N. Campanelli, “Can I Beat the Bookies ? Betting on the English Premier League with Logistic Regression,” 2018.

[10] J. Sharma, “Predicting Results of Indian Premier League T-20 Matches using Machine Learning,” no. November 2018, 2020, doi: 10.1109/CSNT.2018.8820235.

[11] O. Of, “Predicting Football Results Using Machine Learning Techniques,” 2018.

[12] W. Chang, Y. Liu, Y. Xiao, X. Yuan, X. Xu, and S. Zhang, “A Machine-Learning-Based Prediction Method for Hypertension Outcomes Based on Medical Data,” 2019.

[13] P. Sources, “Machine Learning-Based Approach to Predict Energy,” 2020, doi: 10.3390/en13184870.

[14] A. Giełczyk, “Who Will Score? A Machine Learning Approach to Supporting Football Team Building and Transfers ´,” 2021.

[15] D. Junior, V. Lopes, and G. W. Burgreen, “North American Hardwoods Identification Using Machine-Learning,” 2020.

[16] Y. Huang, S. Li, M. Chen, T. Lee, and Y. Chien, “Machine-Learning Techniques for Feature Selection and Prediction of Mortality in Elderly CABG Patients,” pp. 1–11, 2021.

[17] A. Agibetov et al., “Machine Learning Enables Prediction of Cardiac Amyloidosis by Routine Laboratory Parameters : A Proof-of-Concept Study,” no. 4, 2020.

[18] G. Battineni, G. G. Sagaro, and N. Chinatalapudi, “Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis,” 2020.

[19] M. Hung et al., “Using Machine Learning to Predict 30-Day Hospital Readmissions in Patients with Atrial Fibrillation Undergoing Catheter Ablation,” pp. 1–10, 2020.

[20] P. M. Welsing and R. De Jonge, “Complex Machine-Learning Algorithms and Multivariable Logistic Regression on Par in the Prediction of Insufficient Clinical Response to Methotrexate in Rheumatoid Arthritis,” 2021.