Abstract:

This report presents an extension of earlier work in using artificial intelligence to predict football match results. An expanded model is described, as well as a broadening of the area of application of the original work. Several classification algorithms including *Logistic Regression*, *Multi-Layered Perceptron Classifier*, *Random Forest Classifier*, *K-Nearest Neighbours Classifier*, and *Gaussian Classifier* from the scikit machine learning library are used to create a model which forecasts the outcomes of a certain football match. The model is presented with several features that attempt to capture the quality of various sporting teams. Out of all the algorithms used, the most efficient were ***Logistic Regression*** and ***Random Forest Classifier***.

Keywords: [Scikit, Logistic Regression, MLP, Random Forest, KNN, Gaussian Classifier**]**

Introduction:

Technology is inevitably becoming more and more involved in every industry as the world develops. Applications of artificial intelligence can be found in everything ranging from complex problems to everyday jobs. Machine learning and AI have shown promising results in classifying and predicting complex cases. The model will try to simulate human behavior or thinking and can be trained to solve specific problems. During this study, football will be the sport representing this complex area. Football is a decent representation of a complex area because of its dependency on many features, such as home team, away team, match form, goals conceded, goals conceded, red cards, and luck. Large and detailed collections of historical data, per season, per league, and per match, offer great possibilities for investigating the modeling of football results. Now with the help of ML and AI algorithms, we can give an overview of which impact is likely to fall towards.

- **Motivation:**

Football is the most popular sport in the World, with an estimated global following of 4.0 billion fans worldwide. Football draws attention from people of various age groups. As with any business, sports have also embraced the new era of artificial intelligence. In football, artificial intelligence (AI) takes on the job of an assistant referee while acting as a teammate to the players. As with any industry, sports have also embraced the new era of artificial intelligence. In football, artificial intelligence (AI) takes on the job of an assistant referee while acting as a teammate to the players. The FIFA World Cup 2022 features technology like Goal Line Technology, Semi-Automated Offside Technology, and Video Assistant Referee (VAR) to collect data. The market for sports betting is now 500 billion dollars. The world's most-watched domestic league is the English Premier League. For our predictive analysis, we examine data from the previous 20 years.

- **Problem Statement:**

In this project, the Premier League matches are our primary emphasis (England Football league). We worked on this project because have a strong interest in statistical modeling and prediction and are enthusiastic football fans. This topic provided an excellent opportunity to combine these two interests. We will forecast two probabilities for each game: the chances that the home team will win and that they will lose. These anticipated probabilities may be helpful in assessing the accuracy of our models by predicting the outcome of each match, which will be correlated with the largest anticipated probability. To produce a predicted result that might take either of the two options—a home win or an away win—we employed logistic regression, a random forest classifier, an MLP classifier, and a KNN classifier.

Related Work:

Three earlier pieces of research that have all used machine learning to predict football results have been analyzed in our project.

The first study by Arabzad et al. (2014) incorporates the team, its performance throughout the league up to that point, its performance in its four most recent matches, the caliber of its most recent opponents, the league, and the week. The difference between this study and others is that the anticipated output is the number of goals scored by the home side and the away team.

The second study, by McCabe and Trevathan (2008), attempts to forecast whether a football game will end in a home, away, or draw result. They discovered an accuracy of 54% using the collection of features that included goals for, goals against, and home and away performance.

The third study, by Tax and Joustra (2015) predicts the results of football matches in the Dutch Eredivisie League. They were able to forecast the result with a 55% accuracy by using factors like goals for, goals against, outcomes from prior games, and the team's performance (win, draw, lose) in percentages.

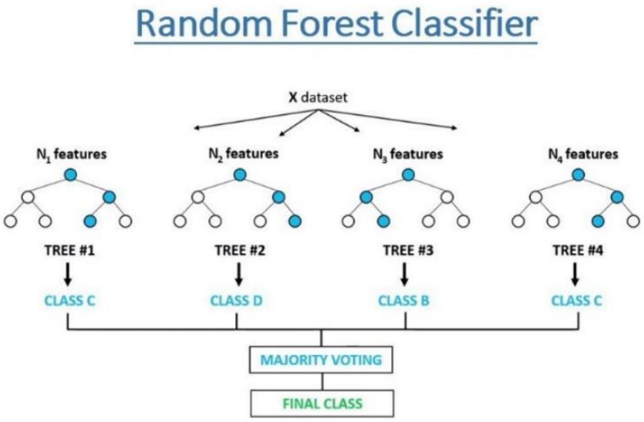
Proposed methodology:

- **Logistic Regression:** It is one of the main classification algorithms that we used. It is used with categorical variables to measure the probability of occurrence. In this form of supervised learning of Machine Learning, we find the odds of an event happening or not (win or loss for a team).

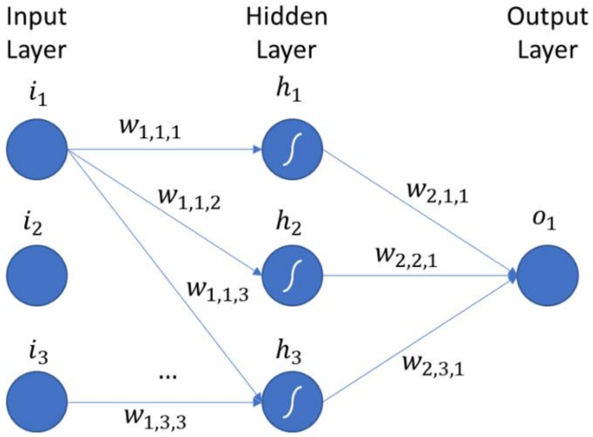
  $$\theta = \frac{p}{1-p}, \qquad \textit{where } p \textit{ is probability of an event happening}$$

  We will use a confusion matrix to see the results.
- **Random Forest Classifier:** The Next Algorithm we used was RFC from the scikit-learn library which operates by constructing multiple decision trees during the training phase. The majority class from all the trees is selected as our result class. Important terms like entropy, information gain, leaf node, decision node, and root node to focus on while

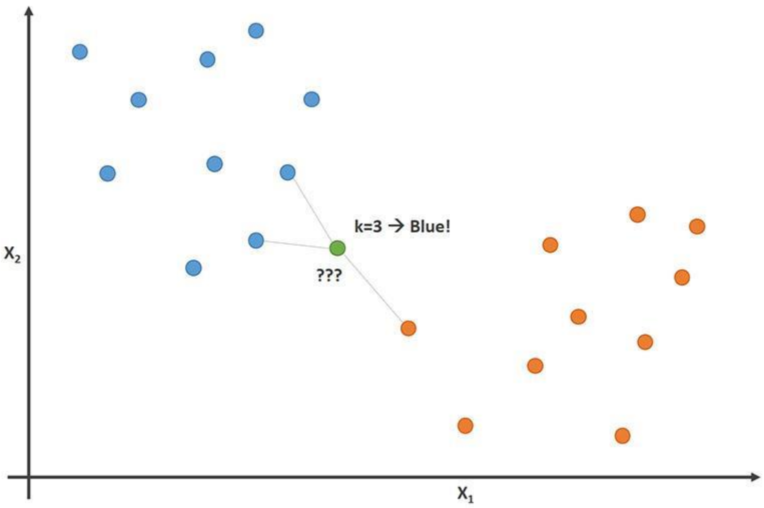
choosing the division criterion. In this algorithm, we use bootstrapped dataset and finally use out-of-bag values to check for the testing of our model.



- **MLP Classifier:** This uses a simple neural network approach that we have learned. It generates random weights, and biases for our hidden layers and produces a result. Then this result is used in backpropagation to nudge the weights and biases accordingly to increase our accuracy and the model learns through it.



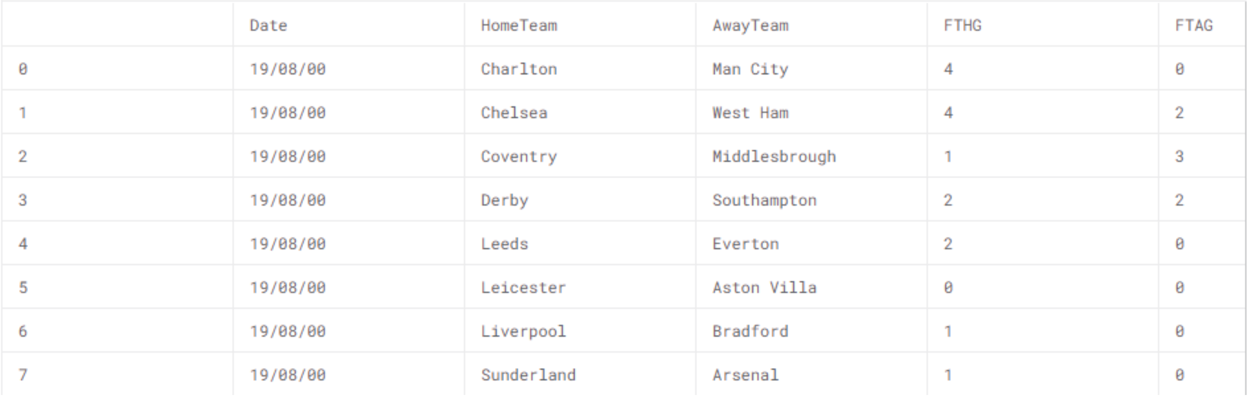
- **KNN Classifier:** The k-nearest neighbor’s algorithm is a supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point meaning it chooses the closest k numbers of known class points to predict the output of which class it might belong to. The distance between can be measured using Euclidean distance or Manhattan Distance.



Dataset details:

https://www.kaggle.com/code/saife245/football-match-prediction/data

We have analyzed data from different sources corresponding to Premier League match results from the 2000/2001 season to the 2018/2019 season. As mentioned, we will be working with data from the last 20 seasons of the Premier League which means we are currently looking at 6840 rows and 40 columns. Some of the key features of the dataset include Home Team, Away Team, FTR, HTP, ATP, Home Team, and Away Team Forms (Previous 5 matches), etc.



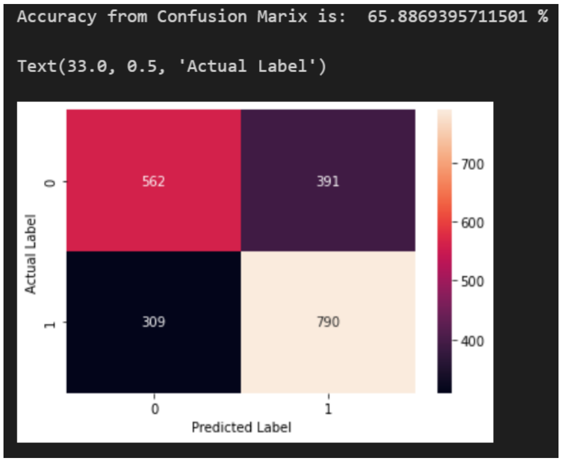
	Date	HomeTeam	AwayTeam	FTHG	FTAG
0	19/08/00	Charlton	Man City	4	0
1	19/08/00	Chelsea	West Ham	4	2
2	19/08/00	Coventry	Middlesbrough	1	3
3	19/08/00	Derby	Southampton	2	2
4	19/08/00	Leeds	Everton	2	0
5	19/08/00	Leicester	Aston Villa	0	0
6	19/08/00	Liverpool	Bradford	1	0
7	19/08/00	Sunderland	Arsenal	1	0

Figure 1 Dataset Image

Experiments and Results:

Dataset: The original data set had quite a few numbers of unnecessary features in it which we later, dropped to make our model efficient. We also tried to train our model using more features and categorically assigned codes to team names to include them in our training. But that decreased our accuracy, so we eliminated that part again. We split our dataset into training data and test data by a ratio of 70% to 30%.

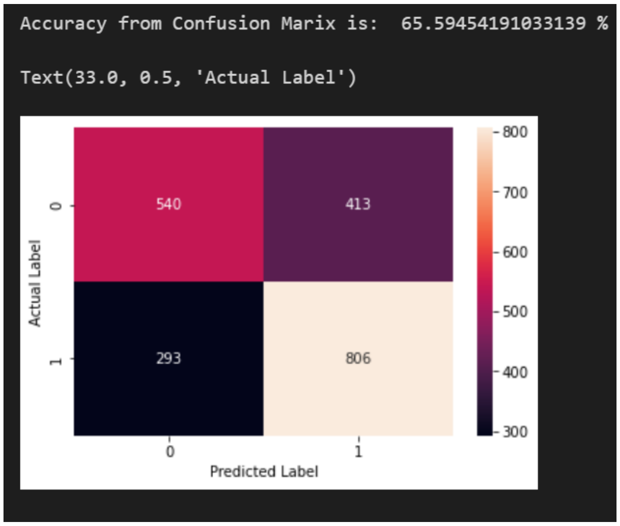
Logistic Regression: This algorithm proved to be efficient for us. Firstly, we used this classifier without passing any argument and it gave us an accuracy of 64% which was performing well keeping in mind our model predicts two possibilities only i.e., win and loss. After that, we introduced a few arguments in our classifier including solver and max iterations. The default solver *lbfgs* is used here. Max Iterations were set to 1000 and the random state was 1. We attained an accuracy of 65.88%. Now, we again tried to increase max iterations to 2000 but it decreased our accuracy to 64.8148% so we dropped it. Tweaking with random state values did not create any big difference in our output.



```
Accuracy from Confusion Marix is:  65.8869395711501 %

Text(33.0, 0.5, 'Actual Label')
```

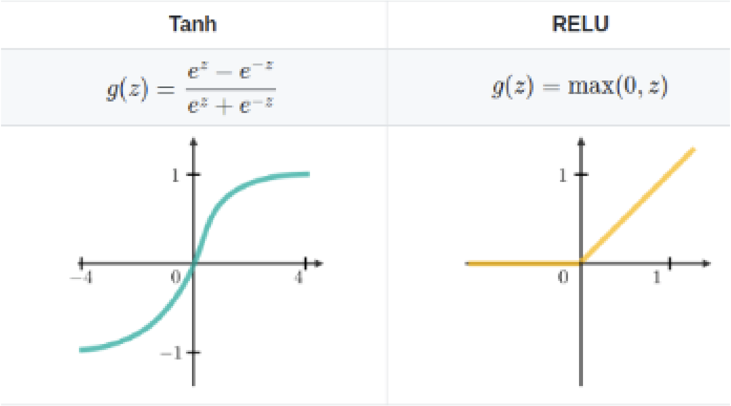
Random Forest Classifier: The second detailed algorithm that we worked on was Random Forest Classifier for which tweaking did not create any large increase in percentage value. Although we did try to change its criterion between *gini* and *entropy*. We also nudged up the value of bootstrapped trees it would create using the *n_estimators'* argument, setting it to 2000. Any more than that, and our training became very slow. We also included some of the other key arguments. The final accuracy of this algorithm was 65.59%.



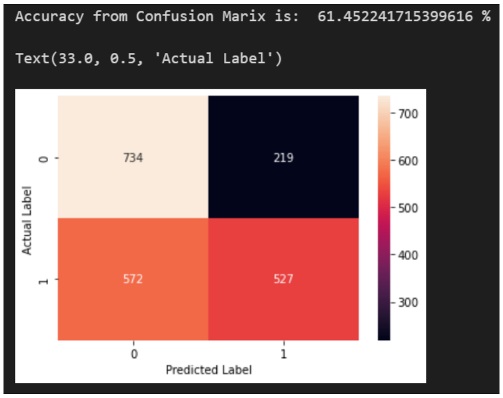
```
Accuracy from Confusion Marix is:  65.59454191033139 %

Text(33.0, 0.5, 'Actual Label')
```

MLP Classifier: The most improvement we attained was in MLP where we jumped from 49.1% to 61.45% by changing its activation function to *tanh* instead of *relu.*



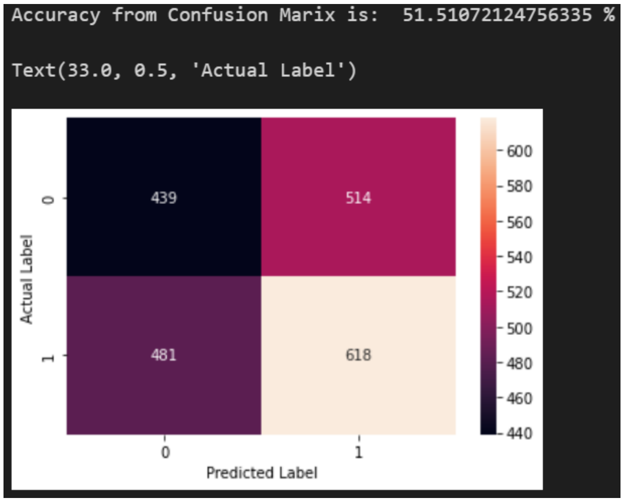
We also tweaked the number of hidden layers and random state values to increase its accuracy. The final accuracy is as follows:



```
Accuracy from Confusion Marix is:  61.452241715399616 %

Text(33.0, 0.5, 'Actual Label')
```

KNN: This model too did not create any big difference either by changing the *n_neighbours*. It kept producing the same value of accuracy around 52%.



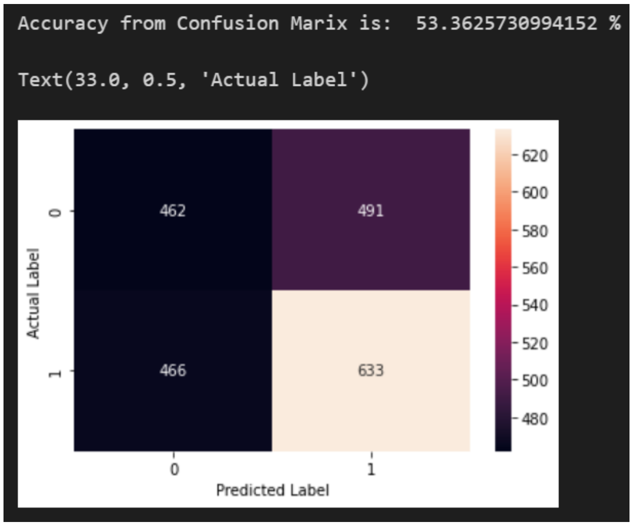
```
Accuracy from Confusion Marix is:  51.51072124756335 %

Text(33.0, 0.5, 'Actual Label')
```

Gaussian Process: Another algorithm that we used was the Gaussian Process Classifier which used the help of RBF (Radial Basis Function) to make our data classes linearly separable and then GaussianProcessClassifier is used. The kernel is used with function 1 * RBF(1.0) and a random state value set to 1. RBF formula can be:

$$k\left(x,y\right)=\exp\left(-\frac{\left\|x-y\right\|^2}{2\sigma^2}\right)$$

Gaussian Process deals with calculating conditional probabilities of an event based on an event that has happened before. The accuracy that we got was 53.36%.



```
Accuracy from Confusion Marix is:  53.3625730994152 %

Text(33.0, 0.5, 'Actual Label')
```

Conclusion:

The accuracy value of each of the algorithms is as follows:

Classification Approach	**Accuracy**
Logistic Regression	65.88%.
Random Forest Classifier	65.59%.
MLP Classifier	61.45%
KNN Classifier	51.51%.
Gaussian Classifier	53.36%

After all the results and observations, we can assume that logistic regression will work the best for us in terms of correctly predicting the outcome considering only the two possibilities of win and loss. The focus of our project was Logistic Regression and Random Forest Classifiers as they were performing efficiently.

References

Ahmed Amr Awadallah, Raghav Khandelwal. (n.d.). *Football Match Prediction using Deep Learning.*

Alan Mccabe, Jarrod Trevathan. (2008, April). *Artificial Intelligence in Sports Prediction*. Retrieved from researchgate.net: https://www.researchgate.net/publication/220841301_Artificial_Intelligence_in_Sports_Prediction

Bhardwaj, A. (2022, June 13). *The Application of Artificial Intelligence in Football Risk Prediction*. Retrieved from hindawi.com: https://www.hindawi.com/journals/cin/2022/6996134/

Samba, S. (2019, May). *Football Result Prediction by Deep Learning Algorithms*. Retrieved from researchgate.net: https://www.researchgate.net/publication/334415630_Football_Result_Prediction_by_Deep_Learning_Algorithms

Aslan, B., & Inceoglu, M. (2007). *A Comparative Study on Neural Network Based Soccer Result Prediction*. Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007)

Mccabe, A., & Trevathan, J. (2008). *Artificial Intelligence in Sports Prediction*. Fifth International Conference on Information Technology: New Generations (itng 2008).

Tax, N., Joustra, Y.P., *Predicting the Dutch football competition using public data: A machine learning approach*, Trans. Knowl. Data Eng. 10 (10) (2015) 1– 13.

udin, s. (n.d.). *Football Match Prediction*. Retrieved from kaggle: https://www.kaggle.com/code/saife245/football-match-prediction/data