IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

FOOTBALL WIN PREDICTION

Srimathi E, Ahalya R, Abarna N, Shree Nandhini V N
Dept. Of Artificial Intelligence, Kongu Engineering College, Erode,
Tamil Nadu

Abstract

Soccer, also known as soccer or football, is a game in which two teams of 11 players try to direct the ball into the opponent's goal without using their hands or arms. The team that scores more goals wins the game. Football is the most popular sport in the world in terms of the number of participants and spectators. Understanding the game and predicting the outcome are common interests of fans, coaches, media and players. Although predicting football results is a very difficult task, football betting has grown over time. This is where football predictions play an important role. Various machine learning (ML) algorithms are used to perform such prediction. Various attributes from previous years' records are taken into account, including home goals, full-time goals, half-time goals, away games, etc. Our model classifies the result as win/loss ratio using the above attributes using various machine learning (ML) algorithms with good accuracy.

Keywords

Football, machine learning, algorithms, accuracy, classification, attributes.

I. INTRODUCTION:

Predicting the results of football matches using machine learning algorithms has become quite popular recently. Soccer betting is used by soccer coaches, fans, players, guys and people who bet on soccer matches. Soccer forecasting can have classification and regression problems. Classification predicts which category something belongs to. In this case, something can be a team, a player, a match, etc. Regression predicts how much is. It is a constant amount. Examples: goal, point, yellow cards, etc. There are many potential classification and regression problems in football forecasting. A classification problem may seem like a prediction. Is the player injured? In addition, it is possible to ask a question to predict whether two teams will win or not, who will win, whether a goal will be scored, whether there will be red cards and much more. Regression problems can be just as extensive. Can you predict how many goals or points each side will score? The team gathers during the season. When you watch football, at the end of the day, there's only one thing that matters and that's winning. This means power Predicting the outcome of a football match is very valuable. Predicting the outcome of a soccer match is an interesting two-class classification problem. The two categories are win and loss. Many things can affect or cause the outcome of any football match. When using machine learning to predict this classification problem, the two most important things a data scientist needs are functionality and algorithms. During training and testing, one key function is important, called the target function. It describes the actual class to which the match belongs. i.e. profit or loss. If the algorithm we use predicts the class to which the match belongs, we can talk about a correct prediction. Although there has been some research in this area, there are opportunities to develop a system with higher predictive accuracy without major complications. This study aims to contribute to the existing literature by developing and implementing an improved model using different machine learning classification algorithms such as decision tree, random forest, XGB classifier, logistic regression and KNN algorithm. A feature extraction process is also used to achieve better accuracy.

II. LITERATURE SURVEY:

1. **Predictive analysis and modeling of football results using a machine learning approach for the English Premier League**

Author: RahulBaboota and HarleenKaur, Guru Gobind Singh Indraprastha University, New Delhi, India Department of Computer Science and Engineering, School of Engineering Sciences and Technology, Jamia Hamdard, New Delhi, India.

Available online 28 March 2018, Version of Record 14 March 2019.

2. **Predicting football scores using machine learning techniques**

Author: Josip Hucaljuk and Alen Rakipović, Published in 2011 Proceedings of the

34th International Convention MIPRO

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

Author: Y Joustra - Transactions on knowledge and data engineering, 2015 - academia.edu

4. **Predicting football results using Bayesian nets and other machine learning techniques**

Author: A Joseph, NE Fenton, M Neil - Knowledge-Based Systems, 2006 – Elsevier

5. **Machine learning in football betting: Prediction of match results based on player characteristics**

Author: J Stübinger, B Mangold, J Knoll - Applied Sciences, 2019

6. **Predicting football results using machine learning techniques**

Author: C Herbinet - MEng thesis, Imperial College London, 2018

7. **Using machine learning to improve our understanding of injury risk and prediction in elite male youth football players**

Author: JL Oliver, F Ayala, MBADS Croix, RS Lloyd… - Journal of science and …, 2020

8. **Machine learning in men&#39;s professional football: Current applications and future directions for improving attacking play**

Author: M Herold, F Goes, S Nopp, P Bauer… - … Journal of Sports …, 2019

9. **Who will score? a machine learning approach to supporting football team building and transfers**

Author: B Ćwiklinski, A Giełczyk, M Choraś - Entropy, 2021

10. **The use of machine learning in sports outcome prediction: A review**

Author: T Horvat, J Job - Wiley Interdisciplinary Reviews: Data Mining …, 2020 -

Wiley Online Library

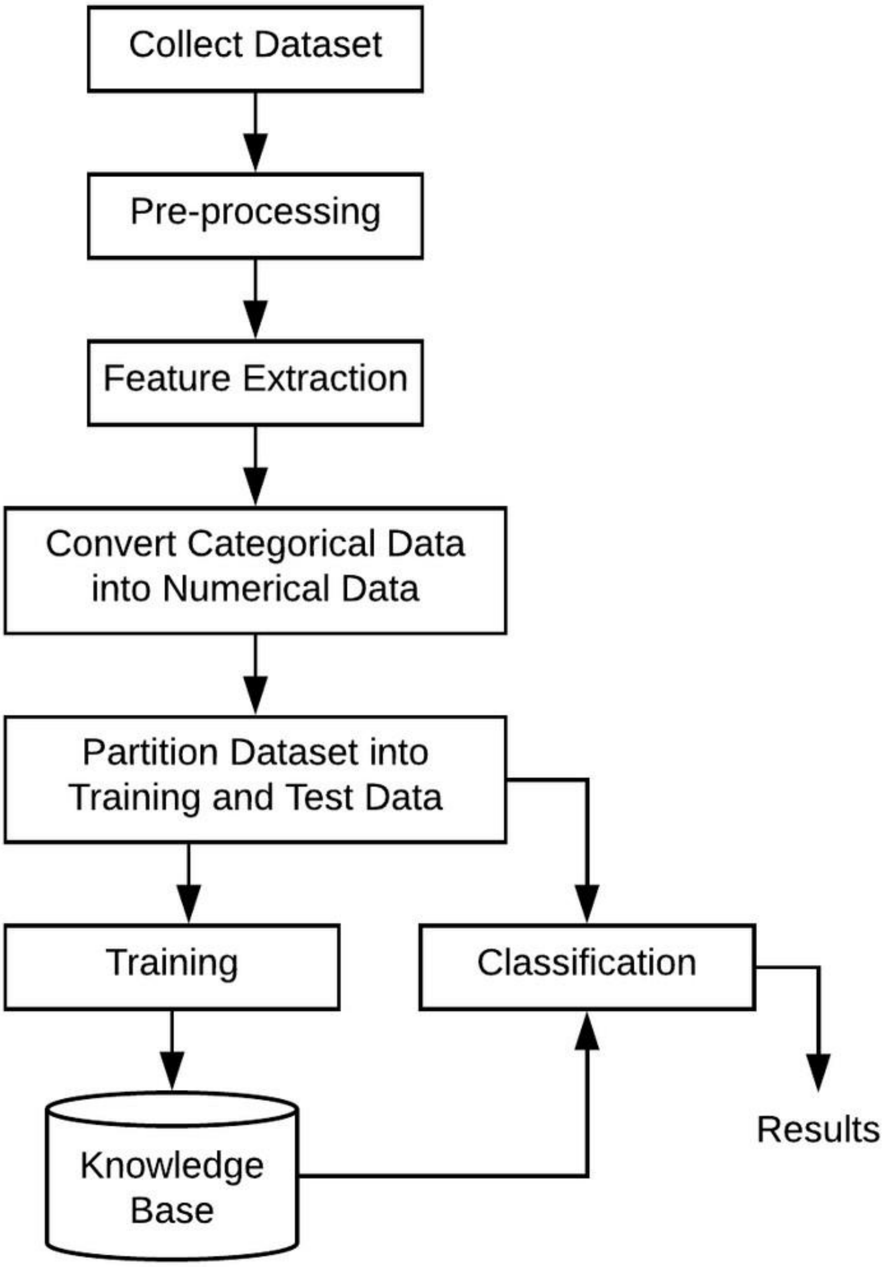
III. METHODOLOGY:

i). Dataset description

Attribute	**Description**
Game id	An in-game id is a unique sequence used to identify your player profile.
2018WL	The result of the football games played in 2018.
HTP	Halftime in a football game refers to the 45th minute.
2019WL	The outcome of 2019 football matches.
FTR	Full Time, or following 90 minutes with added time.

Goals	When the ball completely crosses a goal line, a goal is scored.
Home Goals	We gave the opposition team a goal.
Aways	A ball will go out of the sideline.
2020WL	The outcome of 2020 football matches.
Win/Loss	The outcome of matches.

ii) Proposed model:



The Football data is acquired from Kaggle, and it has 20 attributes. By using feature engineering, 10 of these features were chosen to create the suggested model. After then, preprocessing techniques are used on the dataset to see if it includes any null values. Through this procedure, we divided the dataset into training and testing datasets. Most classification algorithms were employed on the training dataset, and the accuracy of the model was tested using the test dataset.

iii) Correlation matrix:

A table of correlation coefficients between variables is called a correlation matrix.

The correlation between any two variables is shown in each cell of the table. The value

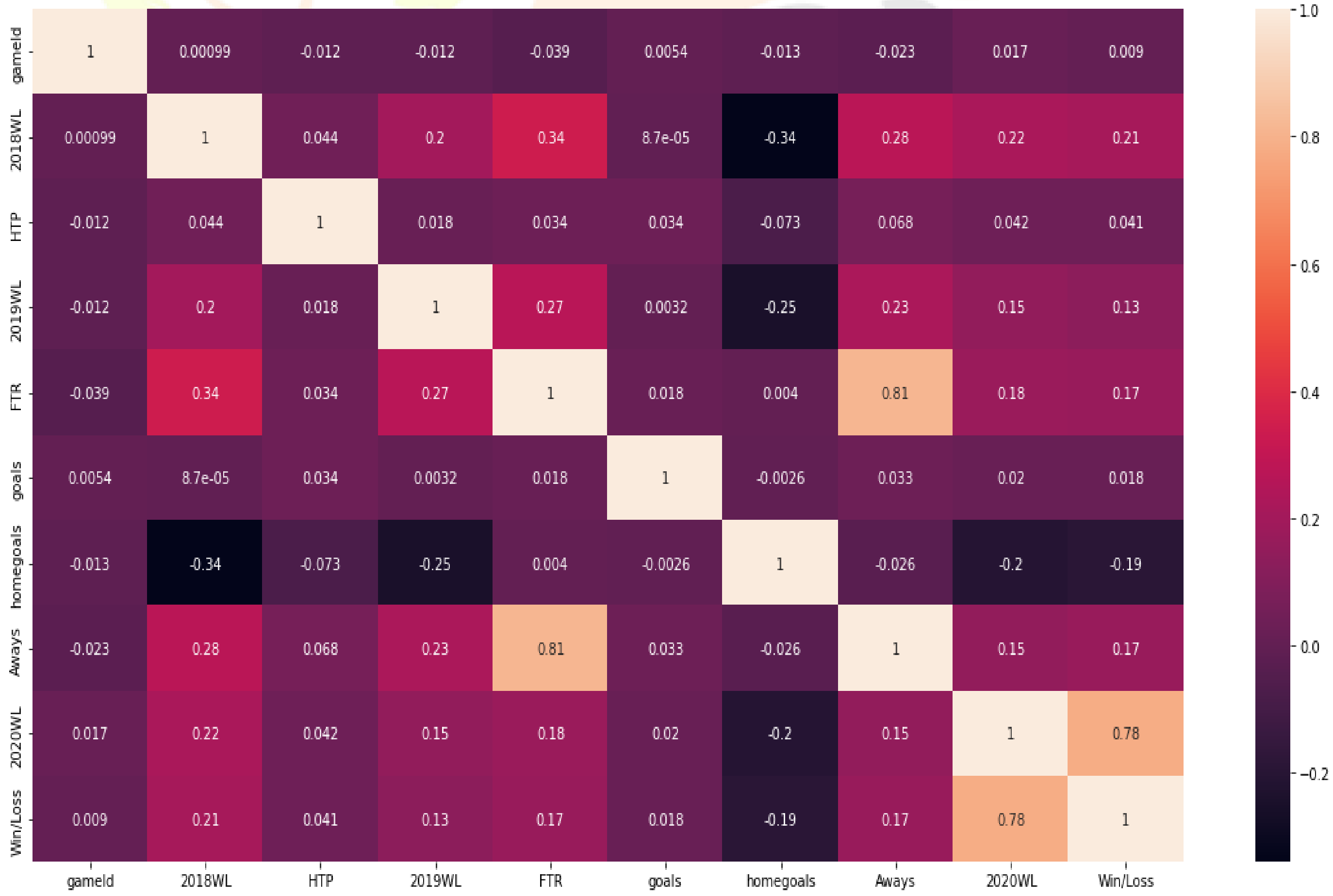
ranges from -1 to 1.

POSITIVE CORRELATION: They are positively connected if an increase in trait A causes a

rise in trait B. A complete positive correlation has a value of 1.

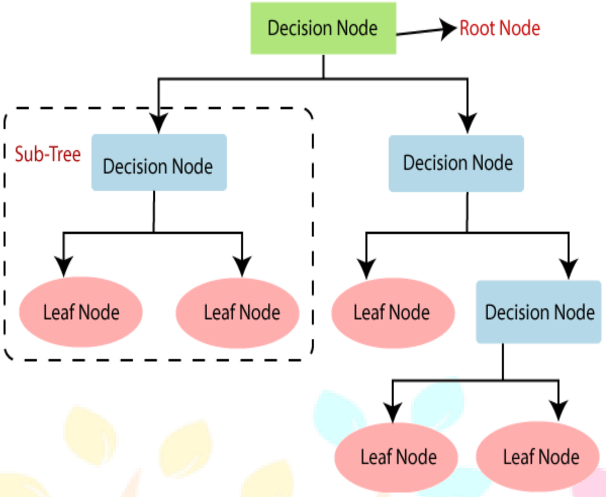
NEGATIVE CORRELATION: They are negatively associated if an increase in trait A causes

a reduction in trait B. A perfect negative correlation has a value of -1.



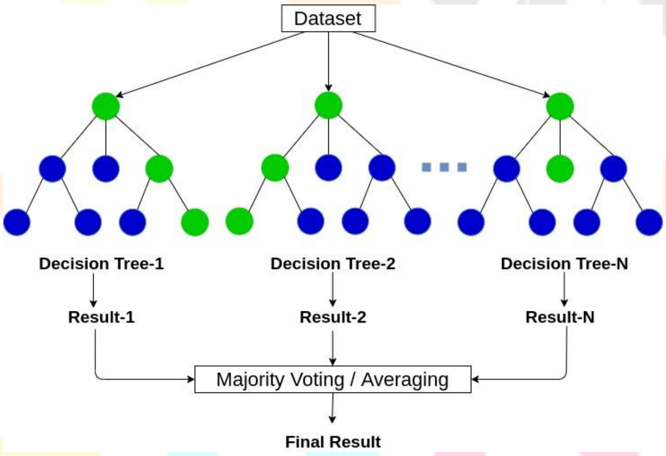
iv) Algorithms used:

1. Decision Tree -Decision trees are a subset of supervised machine learning in which the data is continuously segmented based on a certain parameter. There are two methods

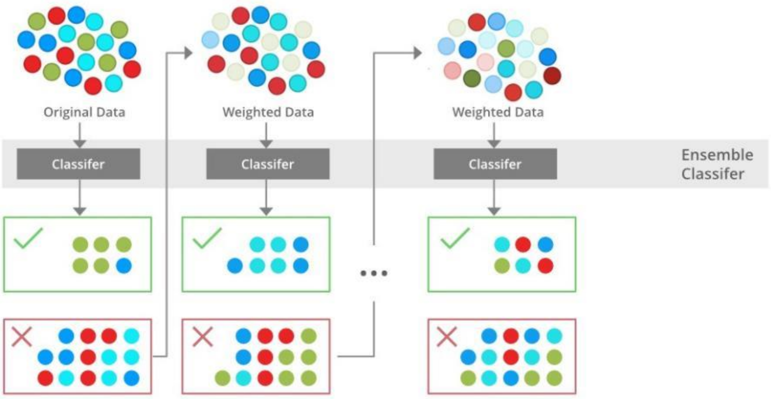


Random forest - Learning is controlled by random forest regression. For regression, the technique employs an ensemble learning approach. The ensemble learning method combines predictions from various machine learning algorithms to provide predictions that are more

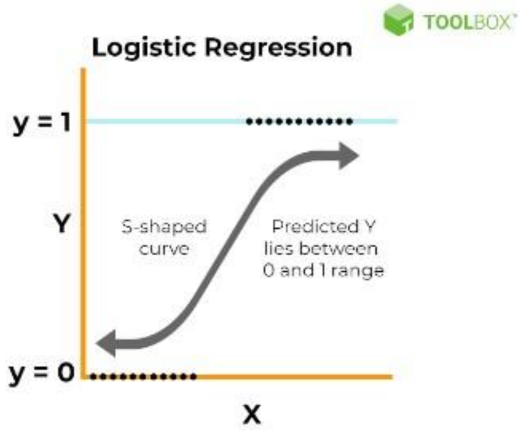
accurate than those from a single model.



XGB classifier – .XG Boost is a scalable Gradient-boosted distributed decision tree (GBDT) machine learning package. XG Boost is an acronym for Extreme Gradient Boosting. is the top machine-learning library for regression, classification, and classification issues and offers parallel tree support.

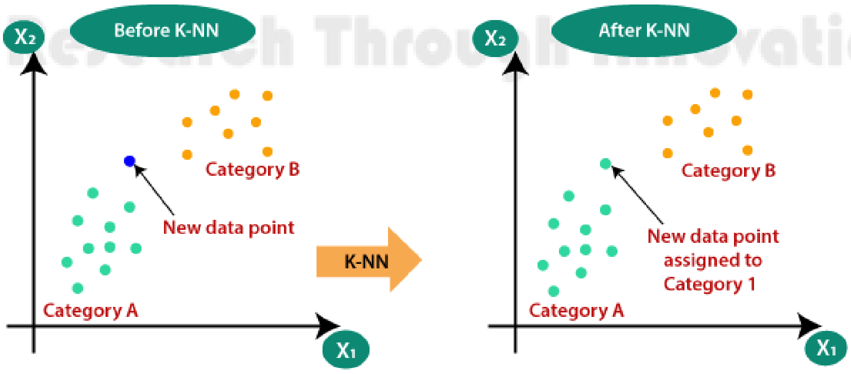


4. Logistic regression - The probability of the target variable is predicted using the supervised learning classification algorithm known as logistic regression. There are only two viable classes if the target or dependent variable is dichotomous in nature.

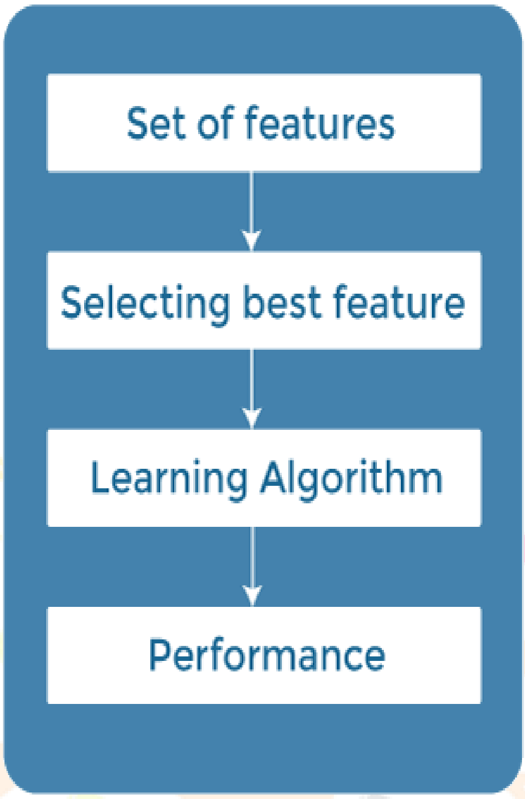


5. KNN Algorithm - One of the simplest supervised machine learning techniques for classification is the K-Nearest Neighbors algorithm. identifies a data point's classification based on that of its neighbors. stores all of the cases that are accessible and groups new

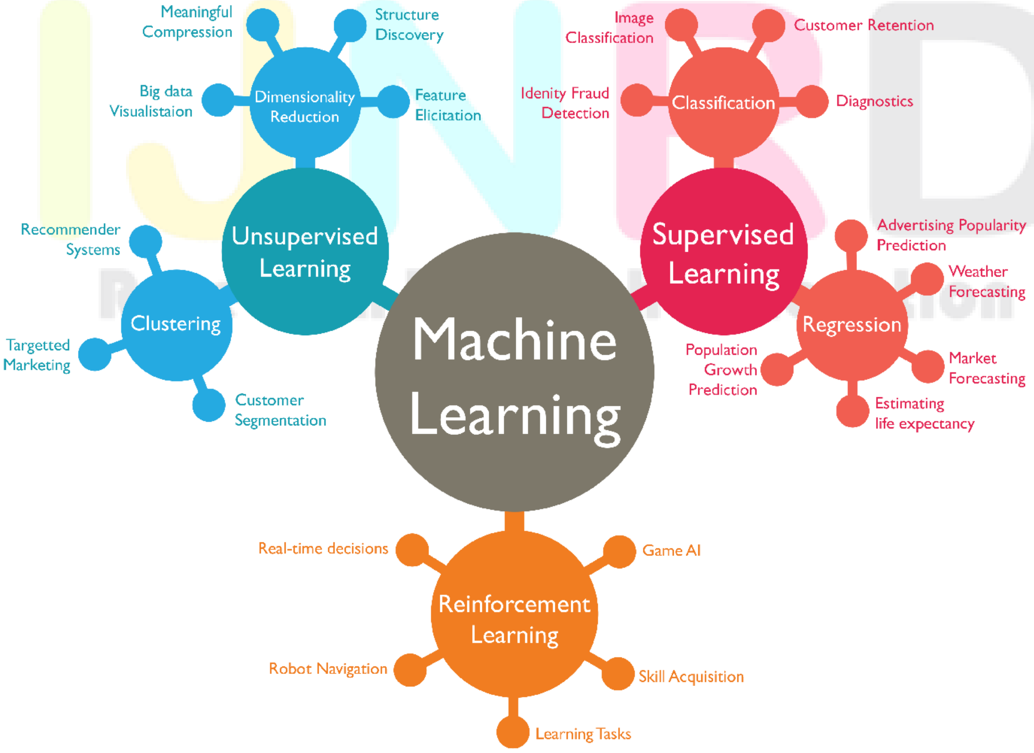
cases according to related functions.



V) Feature Extraction: By generating new features from existing features, feature extraction seeks to reduce the number of features in a dataset (and discarding the original features). Most of the data in the original feature set should be condensed into these new, smaller feature sets.



APPLICATION OF MACHINE LEARNING



IV. PERFORMANCE EVALUATION

The machine learning algorithms that we used are:

1. Decision tree classifier
2. Random Forest Classifier
3. XGB Classifier
4. KNN
5. Logistic Regression

The measure and the metrics that were used are:

1. Precision
2. Recall
3. Accuracy
4. F1-score

1.Precision:

The precision metric gives the ratio of true positives and total positives predicted.

$$\text{Precision} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Positive})}$$

2. Recall:

A Recall provides the ratio of true positives to all the positives in the ground truth

$$\text{Recall} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Negative})}$$

3. Accuracy:

It is defined as the number of correct predictions divided by the total number of predictions, multiplied by 100.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

4.F1-Score:

The F1-score metric results in a combination of precision and recall.

$$\text{F1-Score} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$$

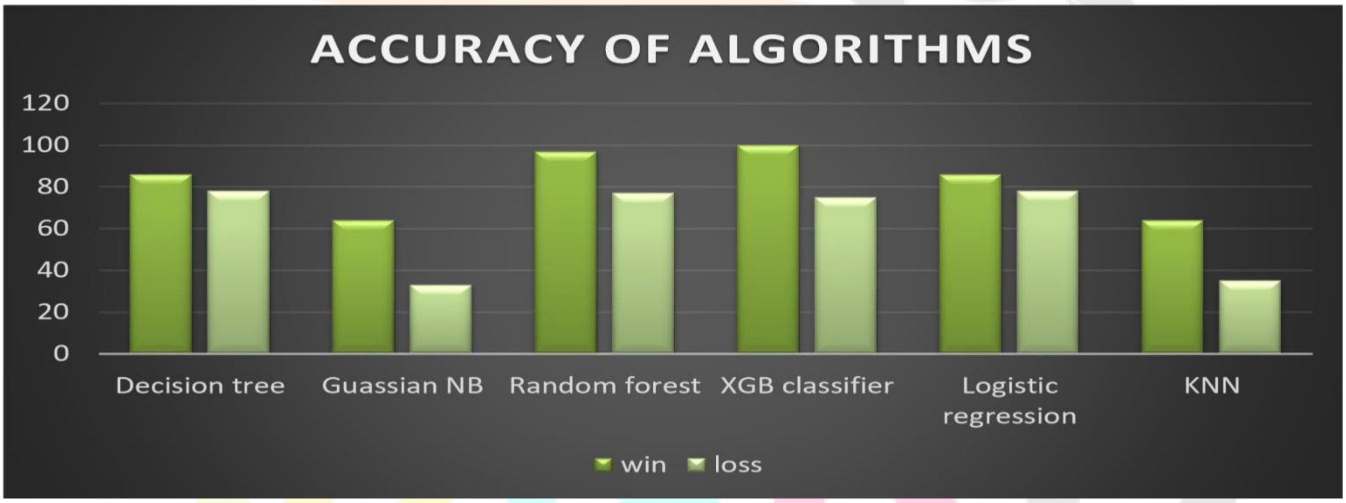
Confusion Matrix:

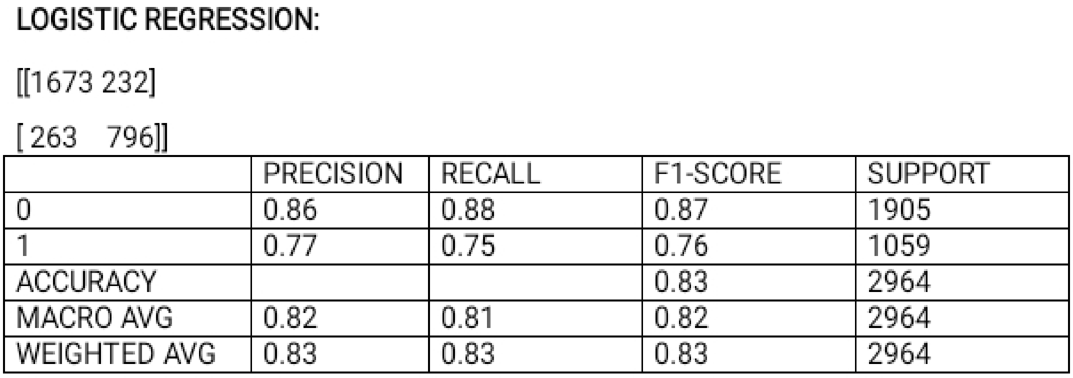
A confusion Matrix is defined as the summary of the model that denotes the predictions

Representation-

TP	**FP**
FN	**TN**

COMPARISON OF MODELS BASED:



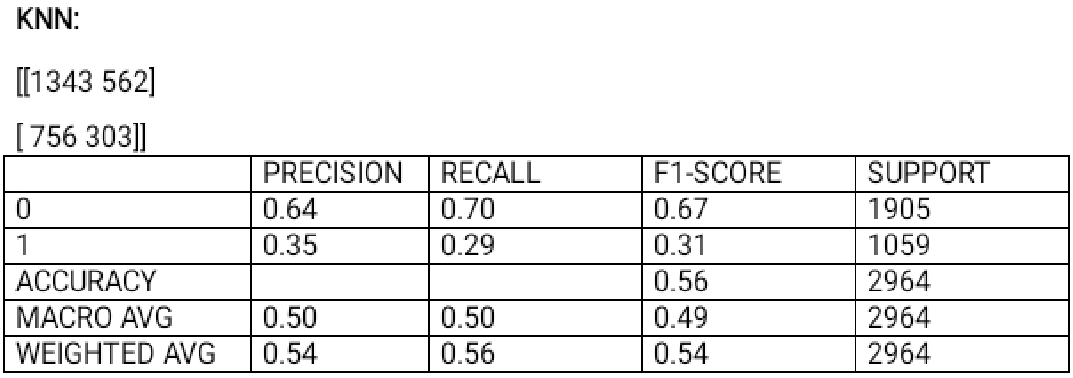


LOGISTIC REGRESSION:

[[1673 232]

[263   796]]

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.86	0.88	0.87	1905
1	0.77	0.75	0.76	1059
ACCURACY			0.83	2964
MACRO AVG	0.82	0.81	0.82	2964
WEIGHTED AVG	0.83	0.83	0.83	2964

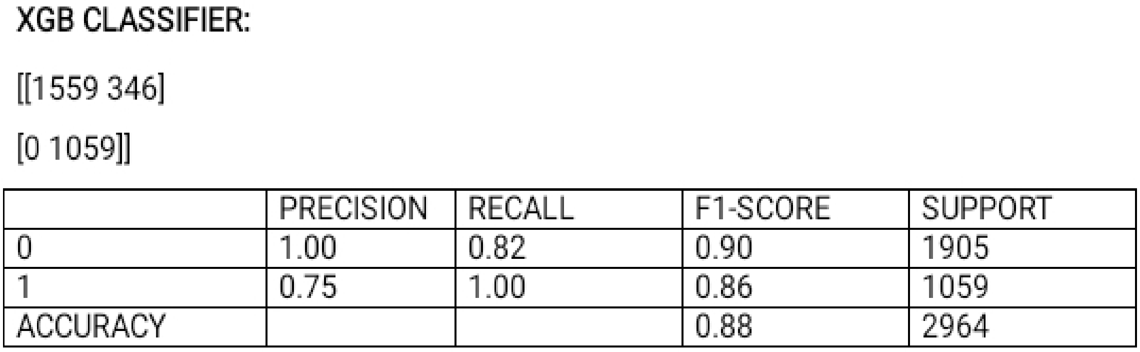


KNN:

[[1343 562]

[756 303]]

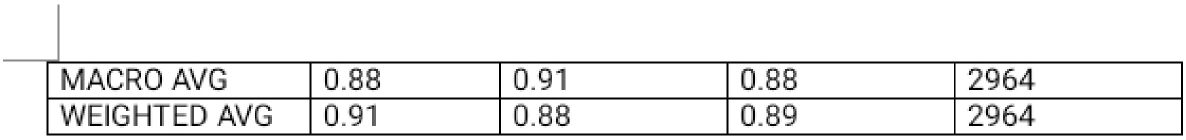
	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.64	0.70	0.67	1905
1	0.35	0.29	0.31	1059
ACCURACY			0.56	2964
MACRO AVG	0.50	0.50	0.49	2964
WEIGHTED AVG	0.54	0.56	0.54	2964



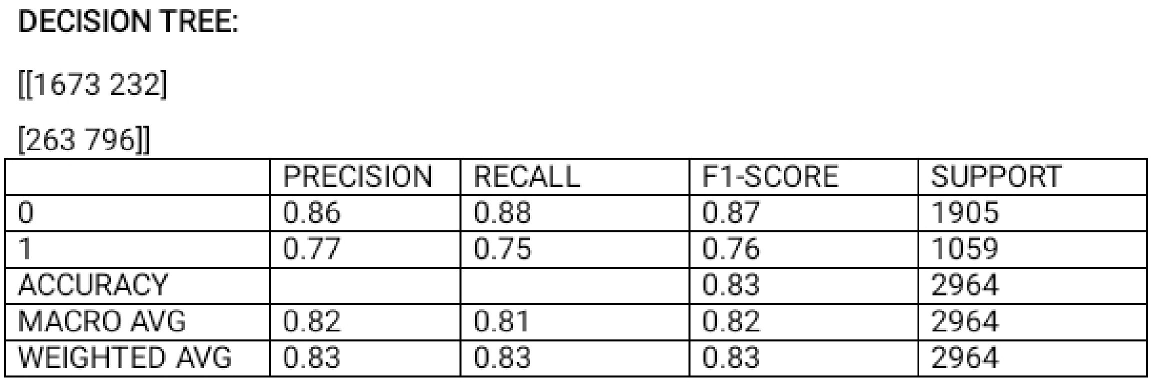
XGB CLASSIFIER:

[[1559 346]

[0 1059]]



	PRECISION	RECALL	F1-SCORE	SUPPORT
0	1.00	0.82	0.90	1905
1	0.75	1.00	0.86	1059
ACCURACY			0.88	2964
MACRO AVG	0.88	0.91	0.88	2964
WEIGHTED AVG	0.91	0.88	0.89	2964

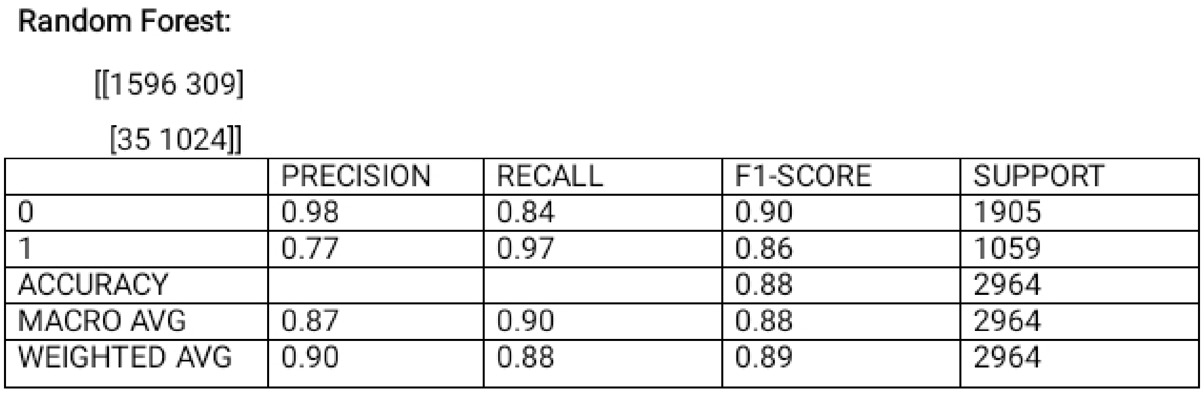


DECISION TREE:

[[1673 232]

[263 796]]

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.86	0.88	0.87	1905
1	0.77	0.75	0.76	1059
ACCURACY			0.83	2964
MACRO AVG	0.82	0.81	0.82	2964
WEIGHTED AVG	0.83	0.83	0.83	2964



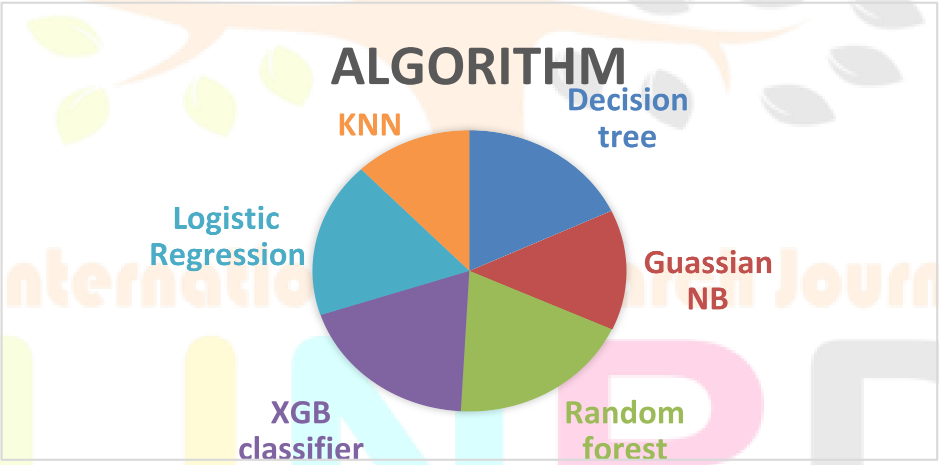
Random Forest:

[[1596 309]

[35 1024]]

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.98	0.84	0.90	1905
1	0.77	0.97	0.86	1059
ACCURACY			0.88	2964
MACRO AVG	0.87	0.90	0.88	2964
WEIGHTED AVG	0.90	0.88	0.89	2964

EVALUATION:



V. CONCLUSION

This paper concludes the prediction of football match using the history of matches played in the previous years. So, the model works on the basis of previous history data. In this model, we implemented around six classifiers to predict the output. Among this six classifiers Random Forest Classifier and XGB Classifier has the highest accuracy of around 90%. It predicted the output through the given dataset “football league”. But the major drawback of the model is the limited number of data fed. When the training data increases the model may give the highest accuracy. The exact model can be used to predict the outcome of other sports too by feeding the model with the required dataset.

To improve the model accuracy, we’ve also used feature extraction and reduced some attributes which do not play a major role in prediction. and we almost got 90% accuracy after feature extraction.

References

1.“Optimization analysis of football match prediction model based on the neural network”-Shuo Guan & Xiaochen Wang, S.I: Cognitive-inspired Computing and Applications-Published: 31 March 2021

2.“Predictive analysis and modeling football results using machine learning approach for English Premier League”-RahulBaboota & HarleenKaur, International Journal of Forecasting , Volume 35, Issue 2-Published: April–June 2019

3.“Predicting The Dutch Football CompetitionUsing Public Data: A Machine Learning Approach-Niek Tax & Yme Joustra, ResearchGate-Published: September 2015