Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/13280
DOI URL: http://dx.doi.org/10.21474/IJAR01/13280

RESEARCH ARTICLE

A COMPARATIVE STUDY OF THE DIFFERENT CLASSIFICATION ALGORITHMS ON FOOTBALL ANALYTICS

Karan Bhowmick and Vivek Sarvaiya

Manuscript Info

Manuscript History
Received: 15 June 2021
Final Accepted: 19 July 2021
Published: August 2021

Key words:-
SVM, Multicollinearity, KNN

Abstract

Sports analytics is on the rise, with many teams looking to use data science and machine learning algorithms to augment their team's research and boost team performance. This is especially true in the case of Football Clubs. In this work, we have taken the statistics of matches for each team from five major football leagues. These include the English Premier League, La Liga, Serie A, Bundesliga, and Ligue 1. We use this data for two kinds of classification to predict a team's win, loss, or draw. First, we implement Multiclass Classification using Naive Bayes classification, Decision Tree classification, and K-Nearest Neighbours classification. We use f1-score, recall, and precision to evaluate the model. Next, we use Binary Classification to predict if a team wins or does not win, i.e., a loss or a draw. We achieve this by using Support Vector Machines, Logistics Regression, K-Nearest Neighbours classification, Decision Tree classification, and Naive Bayes classification. We evaluate the results using the evaluation metrics mentioned above. Now, we compare the accuracy and efficacy of these algorithms based on the evaluation metrics. This will help standardize the means of classification in sports and football analytics in the future.

Introduction:-

In today's world, prediction systems are used in a variety of fields such as Stock Market, Online Shopping, etc. In Sports, these systems could be used to predict the match outcomes, improve the performance of the squad, enhance the game plan. This is a topic that is gaining a lot of interest recently as more teams are trying to implement such systems to improve the performance of the teams. The future of team planning and augmentation of strategies seems to be data-driven. So our work will be in line with this goal.

The paper tackles the issue of match outcome prediction and compares the different classification techniques to find the most optimal one for football analytics. Earlier research has tried tackling the problem of match outcome prediction, but they have not done a comprehensive study of the multifaceted classification algorithms. We aim to add this as a contribution to the existing research. Further, this research will help shed light on the most significant variables affecting the match outcome and the classification algorithm that is most suitable for the prediction.

This goal will be achieved through supervised learning algorithms like, Naive Bayes, K-Nearest Neighbours, Decision Tree, Logistics Regression and Support Vector Machines. In addition, we also plot statistical inferences using data visualization on dependent and independent variables, and we use data visualization to assess the prediction of the models. We compare the Multiclass classification algorithms with the Binary classification

Corresponding Author:- Karan Bhowmick

algorithms on the basis of the evaluation metrics for classification. We will compile the results and insights obtained in tabular form.

Objectives:-

1. Developing classification models to predict the football match result.
2. Comparing Multiclass classification algorithms with Binary classification algorithms.

Literature Survey

Rui Freitas et al. (2020), *Associations of situational and performance variables with defensive transition outcomes in FIFA World Cup* presents the statistical findings of the defensive parameters on the outcome of the match. The paper uses Pearson's chi-square correlation to evaluate the impact of each variable on the result. It also uses log-linear modelling for model development. The most statistically insignificant factor was found to be ball loss. The paper emphasized the importance of recovering possession. This paper could be further elaborated using a more comprehensive statistical and predictive analysis.

Konstantinos Apostolou and Christos Tjortjis (2019), *Sports Analytics algorithms for performance prediction* reviews the literature of sports analytics on the prediction of market value, player injury, and team's performance. The paper uses data visualization and statistical techniques to present trends between the probability of a player scoring and factors such as past record of the player, number of shots taken and the position of the player in the match. However, the paper does not explore different statistical avenues and machine learning techniques that can be used for producing statistically viable insights.

Luca Pappalardo et al. (2019), *A public data set of spatiotemporal match events in soccer competitions* delineates the description of an open dataset, which has a large number of spatio-temporal data. The paper promulgates support for more open datasets to said sports analytics and prediction on the basis of these factors. They use data visualization to encapsulate the insights produced by the data set. Clustering analysis and heatmaps of passes are produced on teams like Juventus and Napoli. Overall, the paper produces a rich analysis of the variables affecting match outcomes. But, the paper fails to broaden its analysis to prediction or classification which could be accommodated by using Machine Learning.

G. Vinué and I. Epifanio (2017), *Archetypoid analysis for sports analytics* emphasizes the use of archetypoid analysis to establish a ranking. The paper uses three scenarios to demonstrate the utility of archetypal analysis. It also presents important insights into an athlete's career, player, team or league performance. The paper also produces a multitude of tables and charts delineating the efficacy of the ADA algorithm.

Vangelis Sarlis and Christos Tjortjis (2020), *Sports Analytics – Evaluation of Basketball Players and Team Performance* presents the findings of the analysis of independent variables on player forecasting. The objective of the paper is to establish a benchmark in predictive analytics for the evaluation of teams and players. The enormous amount of data allows them to accentuate the parameters affecting performance evaluation the most. It evaluates the variable quality through the use of a case study and forecasting scenario.

Anand Ganesan , Harini M (2020), *English Football Prediction Using Machine Learning Classifiers* predicts the match outcomes and the statistically significant variable that affects the target label. This paper compares three algorithms which are Support Vector Machines, XGBoost and Logistic Regression to identify the most suitable algorithm to predict the task at hand. This paper gives fairly accurate predictions for the model they have developed. However, features such as team performance metrics and sentiment analysis can be introduced.

Anurag Gangal et al. (2015), *Analysis and Prediction of Football Statistics using Data Mining Techniques* predicts the success of point system based which makes it interactive for the players over the current FPL system. The GP(Geometric Progression) function is used to predict the outcome of the matches and award points are accordingly awarded to the players.This paper uses interactive prediction models to strengthen the Fantasy Squad.

Maral Haghighat et al. (2013), *A Review of Data Mining Techniques for Result Prediction in Sports* reviews previous data mining techniques used to predict result prediction, player performance assessment, player injury prediction, sports talent identification, and game strategy evaluation. This paper also covers the advantages and

disadvantages of each system. The paper compares the efficacy of different classification techniques such as : ANN, SVM, Logistic Regression, Fuzzy System, Bayesian Model. The paper found that the data quality was sub-standard which led to low accuracy.

Ragini Singla, Dr.Amardeep Singh (2020), *Sports Prediction Using Machine Learning* predicts the match outcome by using classification techniques such as : Naïve Bayes, SVM, Logistic Regression. The algorithms are also compared before and after normalization. The independent variables used are offensive and defensive parameters like Goals scored, corner kicks,red, yellow cards,etc. SVM and Naive Bayes produce the best results without normalization and the Logistic Regression Model produces the best results with normalization. The range of accuracy is around 53-61%.

Darwin Prasetio, Dra. Harlili (2016), *Predicting Football Match Results with Logistic Regression* predicts the match outcomes for a home win or away win and to determine the significant parameters for prediction. The classification technique used is Logistic Regression to develop the model. The independent variables used are Home Offense , Home Defense, Away Offense and Away Defense. The prediction accuracy of the model developed is 69.5% and the most statistically significant variables were found out to be Home Defense and Away Defense.

Research Methodology:-

In this section, we delineate the methodology used in the scope of this wor. Including Data Description and Preprocessing (4.1), Exploratory Data Analysis (4.2), Model Development using Multiclass Classification (4.3), and Model Development using Binary Classification (4.4).

Data Description and Preprocessing

The data set was obtained from datahub, a website with the logistics data of all Football Leagues. Our data set consists of 1827 rows and 61 columns. The dataset includes variables ranging from the number of Fouls, Shots, Goals scored, Bookmaker odds, Division, HomeTeam, AwayTeam among others. We first check for null values, then we proceed to feature selection. From the 61 columns, we bring the number of independent variables to 23 columns. We look at independent variables such as Division, HomeTeam, AwayTeam, Corners, Shots taken, Shots on Target, Yellow Cards, Red Cards, Offsides, Betting odds on win, loss or draw for both Home and Away Teams. We use the result of the game as the target label for classification, denoting 0 for loss, 1 for tie, and 2 for win in Multiclass Classification. For Binary Classification, we use 0 to denote loss or draw, and 1 to denote a win.

Exploratory Data Analysis

We use exploratory data analysis to visualize the relationship between the independent and dependent variables. Furthermore, we see how categorical data influences the different variables and to highlight the important inferences produced. We plot boxplots, line plots and regression plots to analyze the correlations between key components of the model. This is demonstrated below.

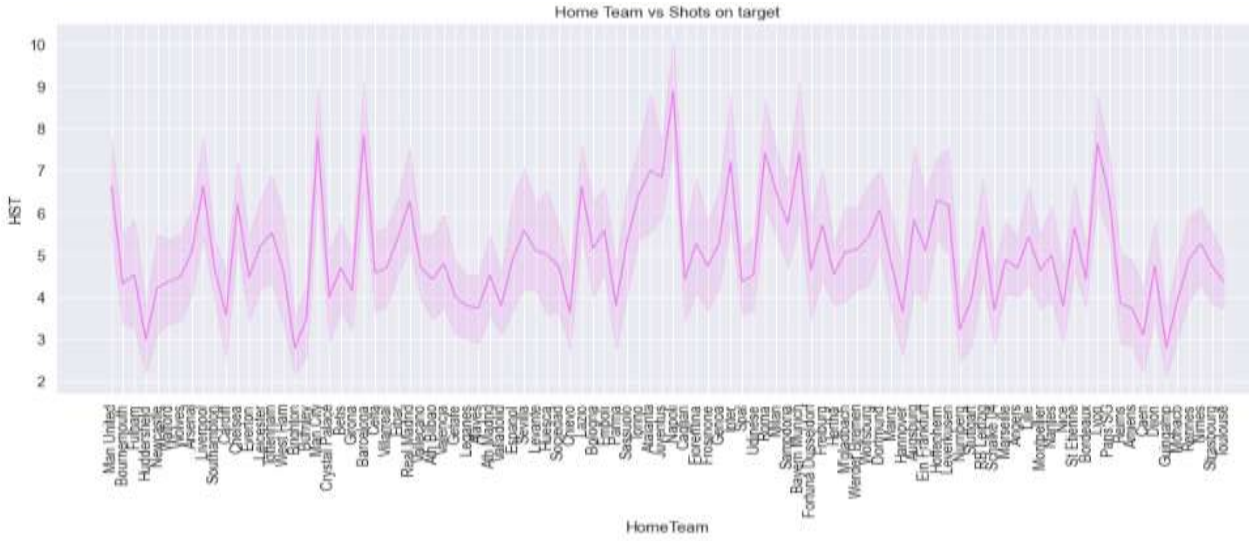

Figure 1:- Home Teams vs Shots on Target.

Here, this plot displays the number of shots on target as a line plot with the names of the teams on the x-axis. We can see that Manchester Utd, Man City, Barcelona, Inter Milan and Paris Saint Germain present the peaks in this graph. Whereas Caen, Guingamp, Brighton and Huddersfield show the fewest shots on target for the season 2018-19.

Next, we plot the correlation plots among the variables selected for the classification model. To see the interplay between these variables and how they predict one another.

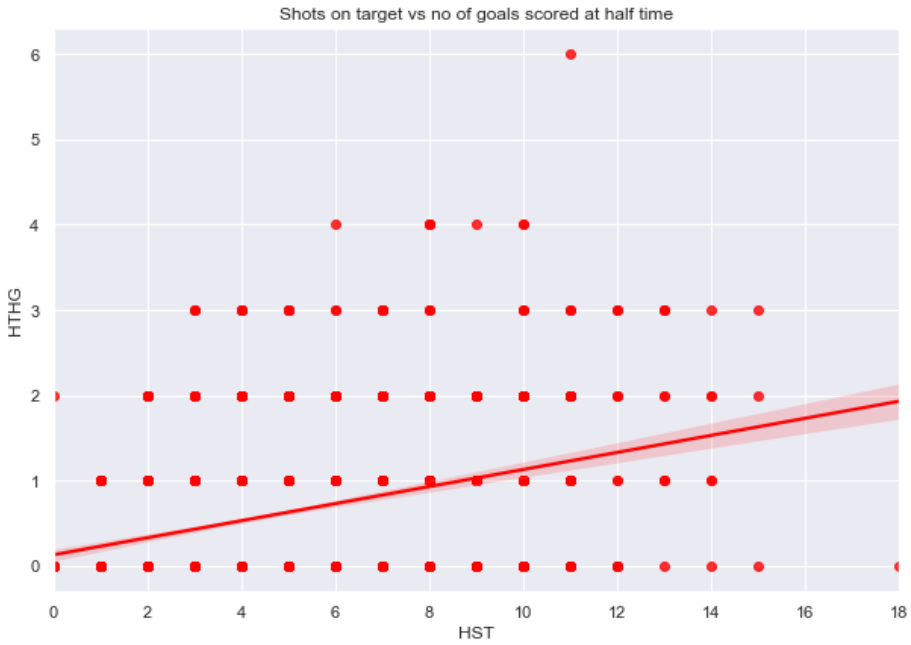

Figure 2:- Shots on Target vs Goals scored at Half-Time.

Fig. 2 shows the regression plot, the slope of the line above indicates a positive correlation between the number of shots on target and the number of goals scored at half-time, whis is to indicate that as the number of shots increase, so do the number of goals at half-time.

Figure 3:- Number of fouls vs Number of Yellow cards issued

This is a line plot demonstrating the relation between the number of fouls committed by the home team and the number of yellow cards issued to the players. We can see that as the number of fouls increase, so do the yellow cards. This is to say, the data is logically coherent.

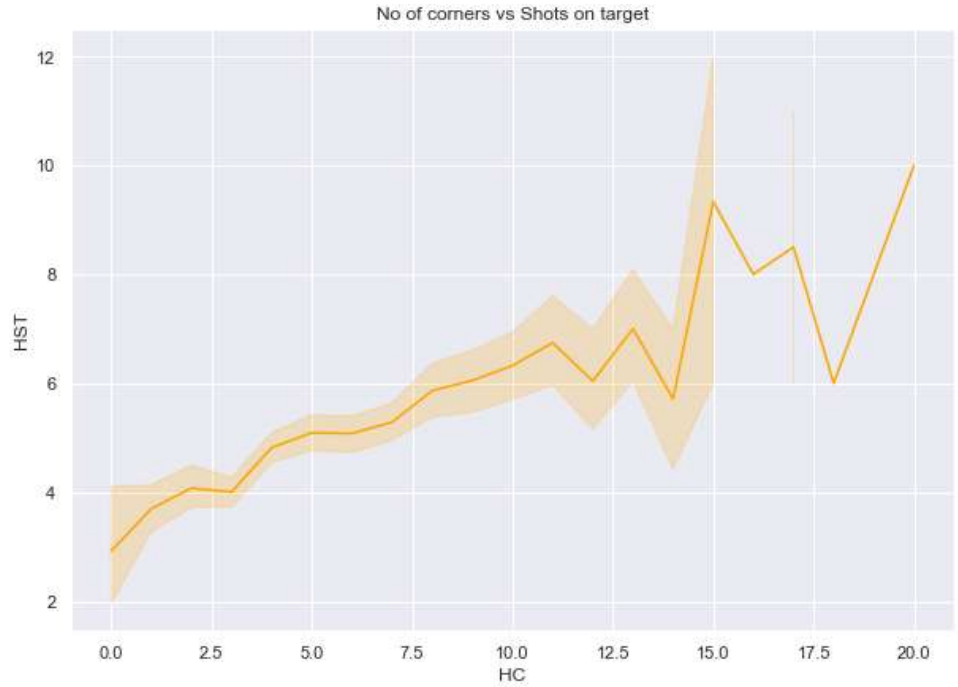

Figure 4:- The number of corners taken vs the number of shots on target.

The figure above presents the line plot of corners vs shots on target. The graph shows steady increase and then an aberrant deviation. This shows that shots on target are more likely if the team has been awarded a corner.

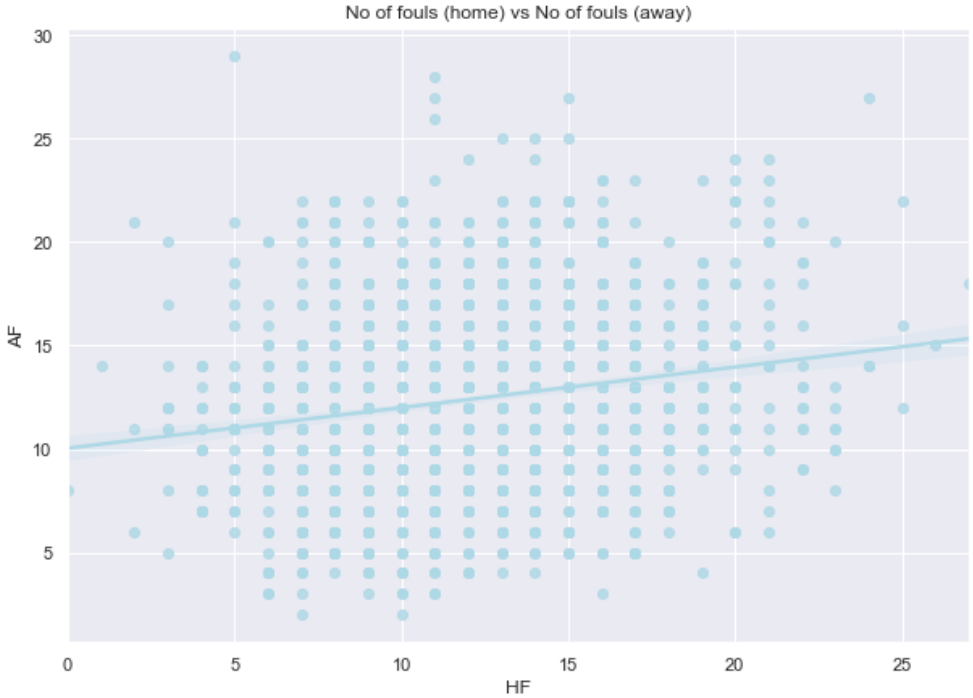

Figure 5:- Number of fouls by Away Team vs Number of fouls by Home Team.

The above graph shows a weak but positive correlation between the fouls committed by the Away Team and the fouls committed by the Home Team. This shows that the number of Home Team fouls increase as the number of fouls against them increase.

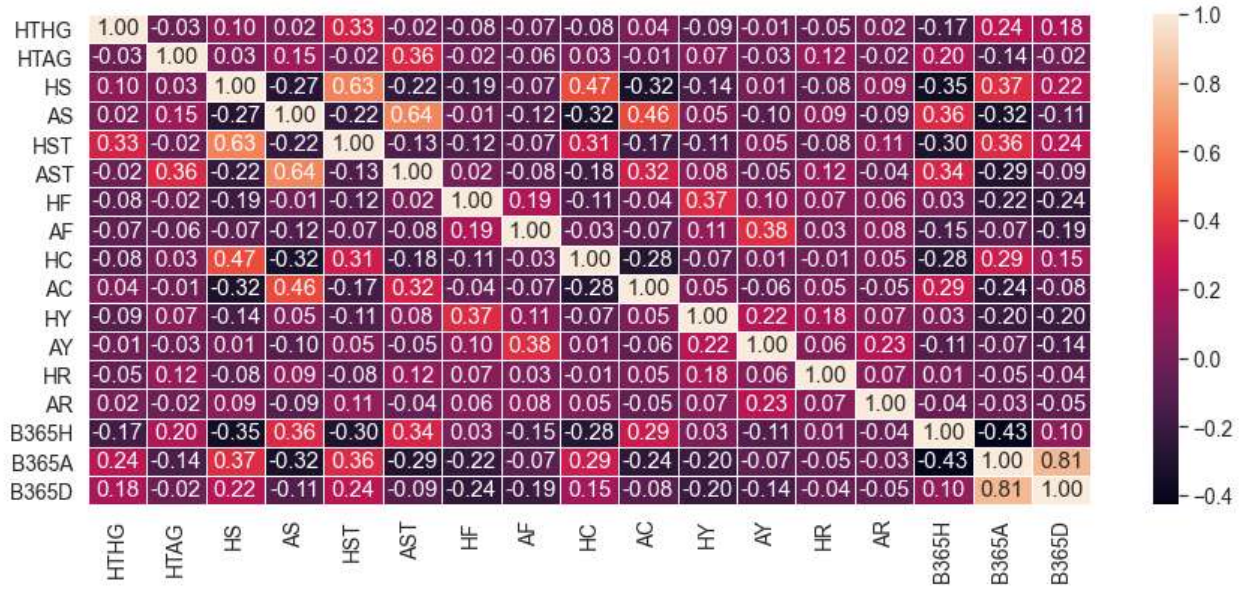

Figure 6:- The Correlation heatmap of the independent variables.

This shows the correlational relationship between the independent variables is shown above. A correlation of 0.5 or greater is considered to be high, since the variables in question are strongly linked to one another.

Model Evaluation Metrics Used

In this section, we go over the metrics used to estimate the accuracy of the classification models being used for analysis. We first take a look at the concepts involved in the evaluation.

		Predicted class	
		Class = Yes	Class = No
Actual Class	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Figure 7:- Confusion Matrix.

In Fig. 7, we have displayed the confusion matrix. As we can see here, there is the actual value and the predicted value. A value is called True Positive when both the predicted and actual labels are positive. Similarly, False Positive indicates that the model predicts a label as positive even though the atual label shows negative. Similarly, we can extrapolate these definitions to True Negative and False Negative. The evaluation metrics we discuss are going to use these definitions. The first evaluation metric we take a look at is the precision.

Precision is used to check the false positive rate. A high value of precision shows that the model has a low false positive rate. It can be defined as the ratio of true positive values to the total positive values reported by the model.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

Next, we take a look at Recall, another metric used in Classification models. Recall is also known as sensitivity and a good score is considered to be above 0.5. It is the ratio of the predicted positive values to the actual predicted values. It tells us how many of the positive results the model correctly predicted.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

F1-score is the weighted average of precision and recall values, it is more significant than accuracy as it differentiates between false positives and false negatives. Accuracy is used for even class distributions, whereas, f1-score is more useful for uneven class distribution.

$$F1-score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad (3)$$

Accuracy is the ratio of the accurate predictions to the total observations.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad (4)$$

Multiclass Classification

We first develop the classification to predict all three forms of result in a football game, i.e. a win, a loss or a draw. We use three Multiclass classification models, namely, KNN Classification, Decision Tree Classifier, and Naive Bayes Classifier. We use a 70-30 split for training and testing data for the model data. We use this to calculate the evaluation metrics as f1-score, recall and precision. First, we take a look at the KNN Classification algorithm.

KNN Multiclass Classification

K-Nearest Neighbours algorithm is a classification algorithm that can be used for both multiclass and binary classification. KNN can be used for classification, regression and outlier detection. It is modelled around variability, depending on the K selected. Here, K denotes the number of classes for the data points to be segregated in. It uses the distance between data points to classify its 'neighbours' i.e. the distance between a random point and its neighbours is lesser than the distance between that datapoint to the other classes.

The KNN Classification used for the prediction of a win, team or loss is demonstrated below by the Confusion Matrix.

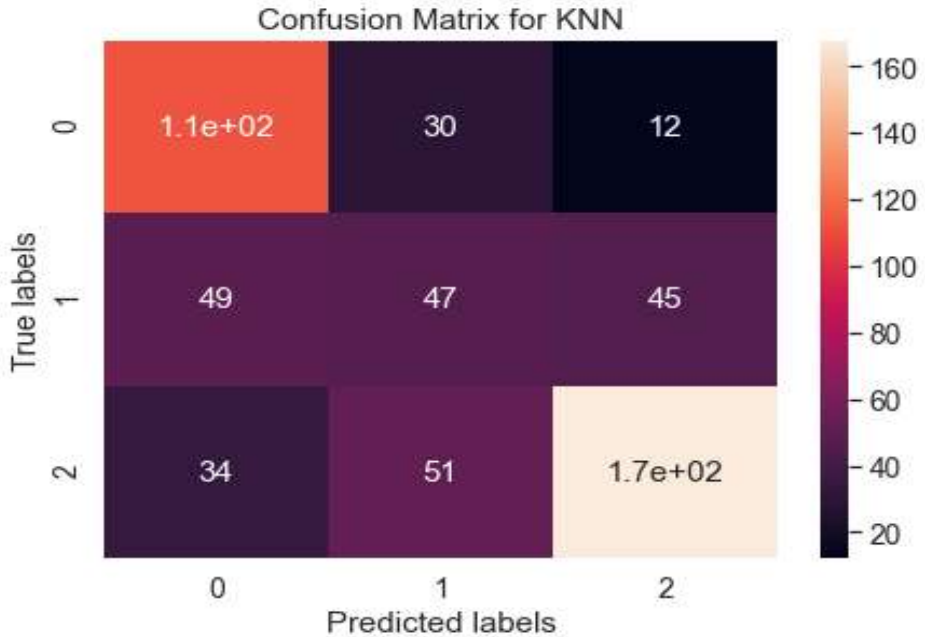

Figure 8:- The Confusion Matrix for the Multiclass KNN model

Here, we can see the efficiency of the model in predicting a win denoted by 2, a draw denoted by 1 and a loss denoted by 0. We can see what mistakes the model makes, i.e. the KNN model confuses 30 values supposed to be a loss and assigns it to the draw class, and 12 values supposed to be a loss to a win. We can see the deficiencies in the KNN model and how it misinterprets them.

Next, we take a look at the weighted average metrics shown by the model in accordance with the metrics mentioned in section 4.3.

	precision	recall	f1-score	support
0	0.57	0.73	0.64	154
1	0.37	0.33	0.35	141
2	0.75	0.66	0.70	253
accuracy			0.60	548
macro avg	0.56	0.57	0.56	548
weighted avg	0.60	0.60	0.59	548

Figure 9:- Classification report produced for the KNN model.

We can see here that the weighted average of precision and recall produce 0.6, whereas f1-score has a weighted average of 0.59 and an accuracy of 0.6. From the figure, we can see that the model shows poor performance in terms of the classification of matches that end in a draw.

Decision Tree Multiclass Classification

The Decision Tree Classifier builds classification models using a tree structure. It divides the data into smaller subsets on the basis of unique variables. The tree has internal nodes which denote tests on particular attributes. Furthermore, each leaf node consists of a class label. In the sklearn module we can set the depth of the tree used in the model, i.e. the complexity of attributes being tested. The model developed for the Multiclass Classification is shown in Fig. 8 below.

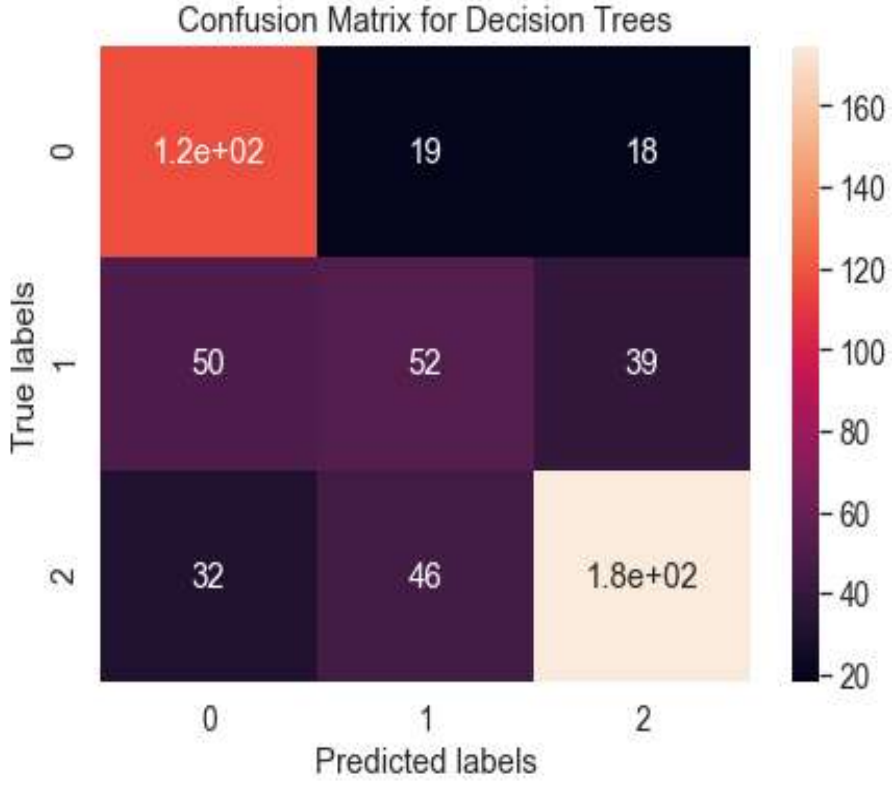

Figure 10:- The Confusion Matrix for the Multiclass Decision Tree Classifier.

We can see here that the Decision Tree Model, just like the KNN classification model, has the most ambiguity in the prediction of a draw. We can see that the prediction of a loss or win are quite streamlined, and it shows lesser confusion than the KNN model.

	precision	recall	f1-score	support
0	0.59	0.76	0.66	154
1	0.44	0.37	0.40	141
2	0.75	0.69	0.72	253
accuracy			0.63	548
macro avg	0.60	0.61	0.60	548
weighted avg	0.63	0.63	0.62	548

Figure 11:- Classification report produced for the Decision Tree model.

We can see from Fig. 10 that the decision tree model produces better overall accuracy than the KNN model. A weighted average of 0.63 for precision and recall, and a 0.62 for f1-score and an accuracy of 0.63.

Naive Bayes Multiclass Classification

The Naive Bayes Classification is an amalgamation of multiple classification algorithms using Naive Bayes principles for classification. They work effectively on smaller datasets and work with surprising speed for the task at hand. This is the last model created for the Multiclass Classification analysis.

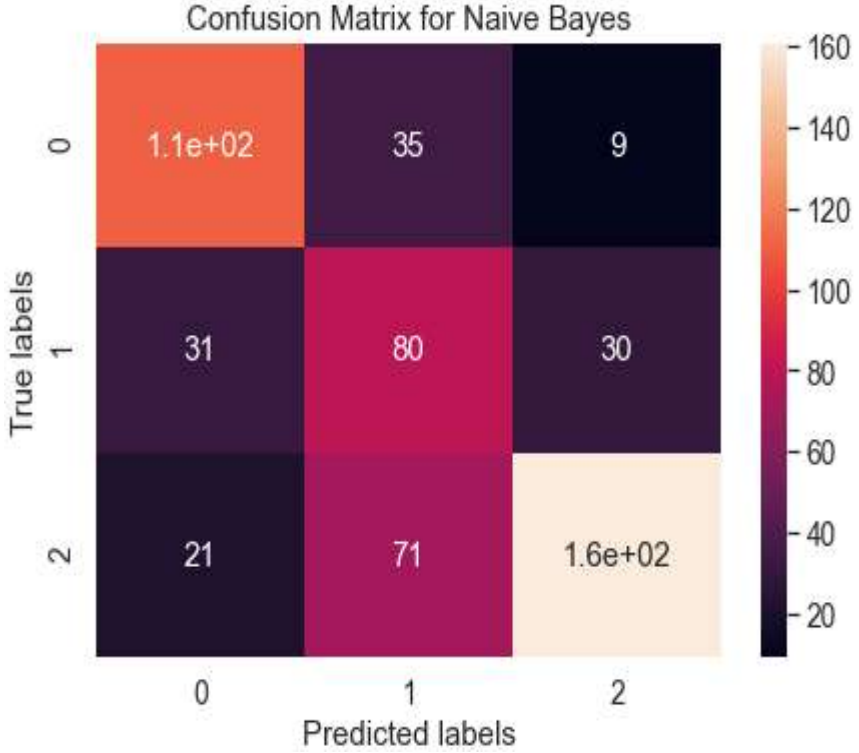

Figure 12:- The Confusion Matrix for the Multiclass Naive Bayes Classifier.

From Fig. 11, it is clearly demonstrated that Naive Bayes Classifier produces the best results for classifying draw matches. The Gaussian Naive Bayes Classifier shows the least ambiguity among the other Multiclass classification algorithms used. This is presented in the classification report produced below.

	precision	recall	f1-score	support
0	0.68	0.71	0.70	154
1	0.43	0.57	0.49	141
2	0.81	0.64	0.71	253
accuracy			0.64	548
macro avg	0.64	0.64	0.63	548
weighted avg	0.67	0.64	0.65	548

Figure 13:- Classification report produced for the Decision Tree model.

As shown, the Gaussian Naive Bayes Classifier produces the highest accuracy results, with a weighted average of 0.67 in precision, 0.64 in recall, f1-score of 0.65 and an accuracy of 0.64. From the individual accuracy metrics shown, we can see that the prediction of draw matches rane from 43-57%, the highest values among the other classifiers.

Binary Classification

We use five Binary classification models, namely, KNN Classification, Decision Tree Classifier, Naive Bayes Classifier, Support Vector Machines (SVM) and Logistic Regression. We use the Multiclass classification for binary prediction in this section to compare the difference in efficiency from Multiclass to Binary. Here, 1 denotes a win and 0 denotes a draw or a loss.

Logistic Regression

Logistic regression uses the sigmoid function to classify categorical data (binary output). It is the first algorithm used for Binary classification analysis in the scope of this paper. The sigmoid function is an S shaped curve when plotted on a curve. It only takes values between 0 and 1 and classifies them to 0 or 1 on the basis of where the data points lie, i.e. which value they are closest to.

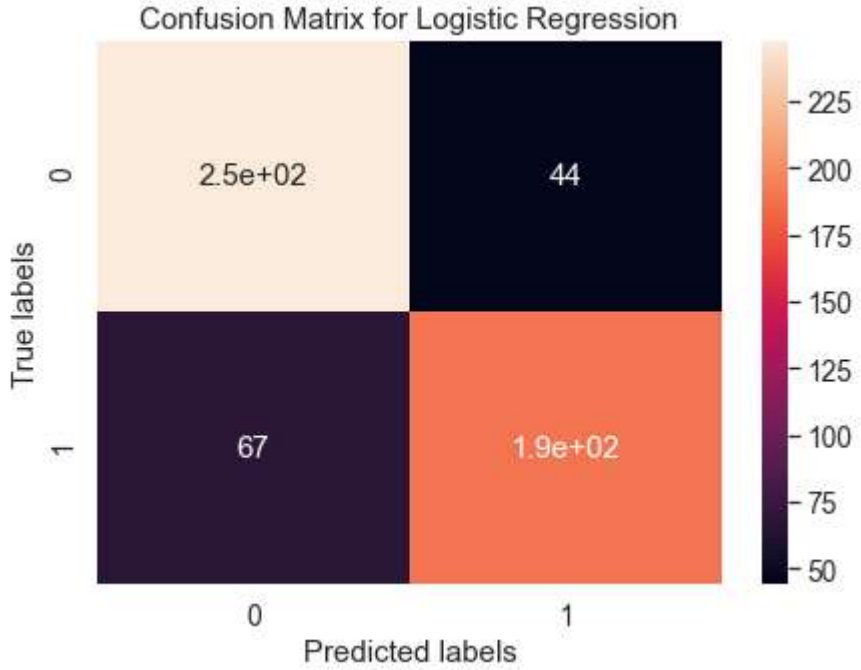

Figure 14:- The Confusion Matrix for Logistic Regression.

The Logistic Regression confusion matrix shows very little ambiguity, the accuracy of predictions are extremely high. Since draws were more ambiguous and they have been clustered into loss and draw, thus, the accuracy results produced are better.

	precision	recall	f1-score	support
0	0.79	0.85	0.82	292
1	0.81	0.74	0.77	256
accuracy			0.80	548
macro avg	0.80	0.79	0.80	548
weighted avg	0.80	0.80	0.80	548

Figure 15:- Classification report produced for Logistic Regression.

The classification report produced shows an extremely ideal weighted average score of 0.8 for precision, recall, f1-score and accuracy. The accuracy for prediction of a loss or a draw ranges from 79% to 85% and the accuracy for predicting a win ranges from 74% to 81%.

Support Vector Machine

Support Vector Machines are used to classify categorical data. It follows the suit of Logistic Regression in the sense that it cannot be used for Multiclass Classification. It can be used for binary output. It uses hyperplane segregation for classification of data points to labels. It selects the best equation of a hyperplane to divide the data into two classes. The model developed for the problem at hand has the confusion matrix shown below.

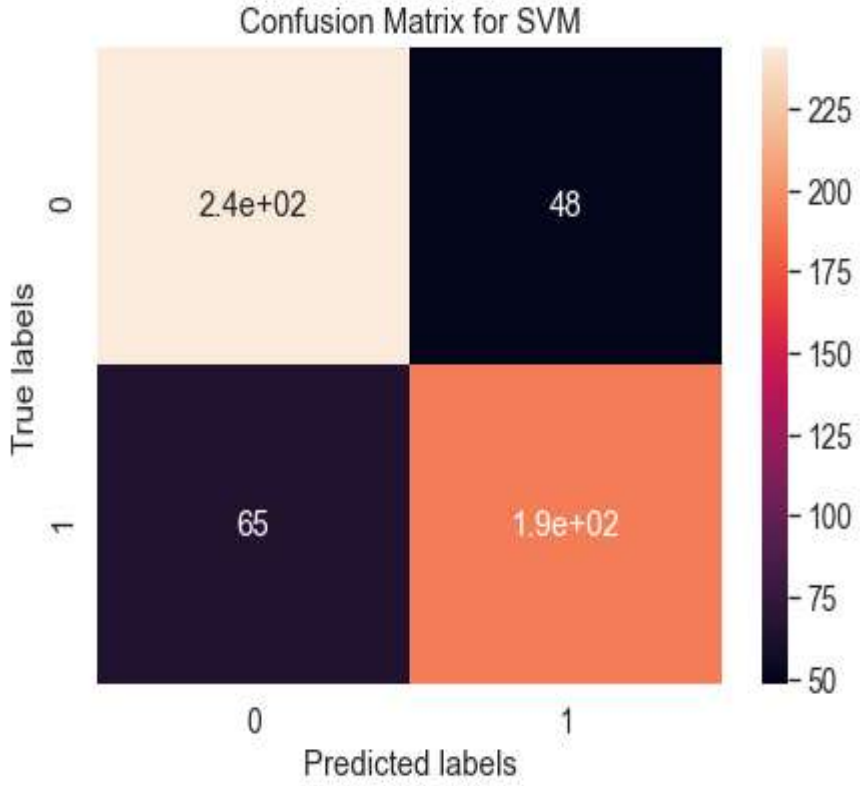

Figure 16:- The Confusion Matrix for SVM.

The SVM model also produces very little ambiguity as shown above. It is similar to the Logistic Regression model in terms of the confusion matrix. We need to quantify the accuracy results for this model to delve into the intricacies of model quality.

	precision	recall	f1-score	support
0	0.79	0.84	0.81	292
1	0.80	0.75	0.77	256
accuracy			0.79	548
macro avg	0.79	0.79	0.79	548
weighted avg	0.79	0.79	0.79	548

Figure 17:- Classification report produced for SVM.

The classification report shows that the SVM model shows a weighted average score for recall, precision, f1-score and accuracy of 0.79. A difference of 1% in accuracy from the Logistic Regression model. Nonetheless, a high accuracy score is produced by SVM. With the prediction of a win ranging from 75% to 80% and prediction of a loss or a draw ranging from 75% to 80%.

KNN Binary Classification

The KNN classification model is the same as that mentioned in section 4.4.1, but here we use it to classify between a win and not a win, i.e. a loss or a draw.

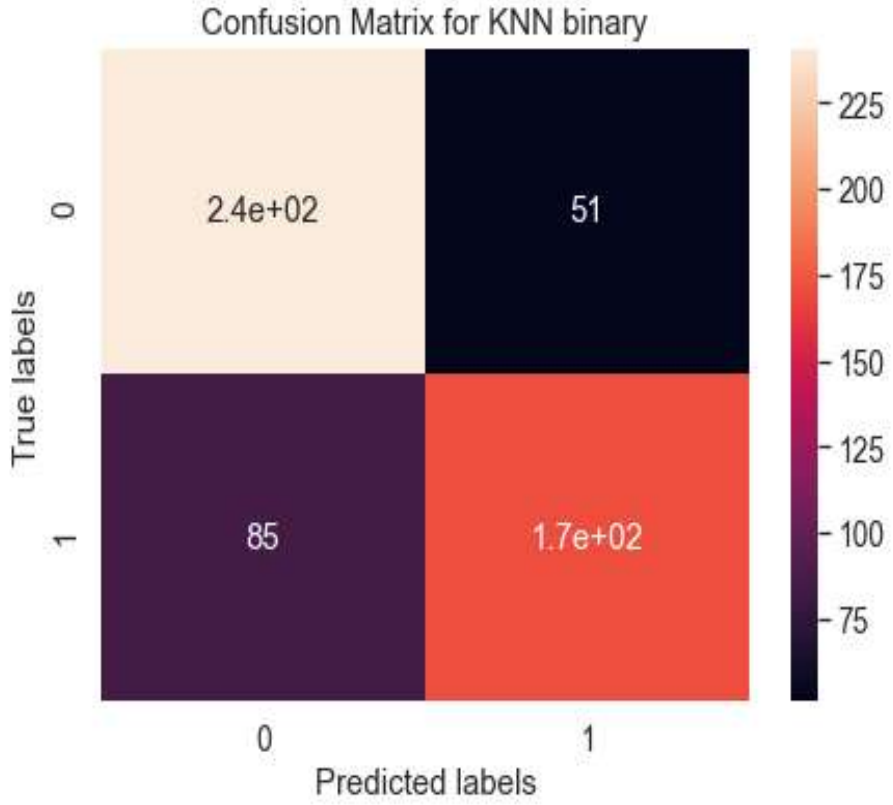

Figure 18:- The Confusion Matrix for Binary KNN.

The KNN Binary Classifier produces better results than the Multiclass Classification since it removes some of the ambivalence witnessed in the prediction of a draw.

	precision	recall	f1-score	support
0	0.74	0.83	0.78	292
1	0.77	0.67	0.72	256
accuracy			0.75	548
macro avg	0.75	0.75	0.75	548
weighted avg	0.75	0.75	0.75	548

Figure 19:- Classification report produced for Binary KNN.

The Binary KNN classification presents better results for the weighted average of recall, precision, f1-score and accuracy. But it shows significantly weaker results when compared to SVM and Logistic Regression.

Decision Tree Binary Classification

As explained in the sub-section above, we use the same algorithm and apply it to the binary problem at hand. The model produced has the following implications.

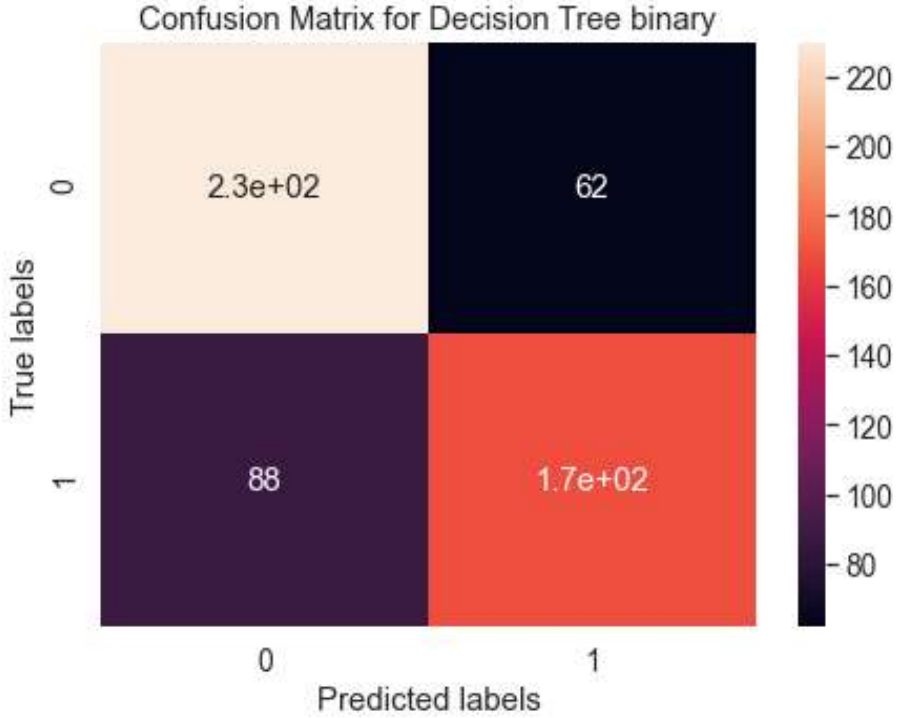

Figure 20:- The Confusion Matrix for Binary Decision Tree.

The confusion portrayed by the model is more than the strictly binary algorithms like SVM and Logistic Regression.

	precision	recall	f1-score	support
0	0.72	0.79	0.75	292
1	0.73	0.66	0.69	256
accuracy			0.73	548
macro avg	0.73	0.72	0.72	548
weighted avg	0.73	0.73	0.72	548

Figure 21:- The Classification report for Binary Decision Tree

Here, we see that the weighted average f1-score is 0.72 and an accuracy is 0.73. The results are an improvement from the 0.62 f1-score and 0.63 accuracy from Multiclass Classification. And, in Multiclass Classification, Decision Tree produced better results than KNN classification. This is not the case here. It produces worse results in Binary Classification compared to KNN Binary Classification.

Naive Bayes Binary Classification

We use the Gaussian Naive Bayes Classifier for Binary Classification. The model's accuracy is demonstrated below.

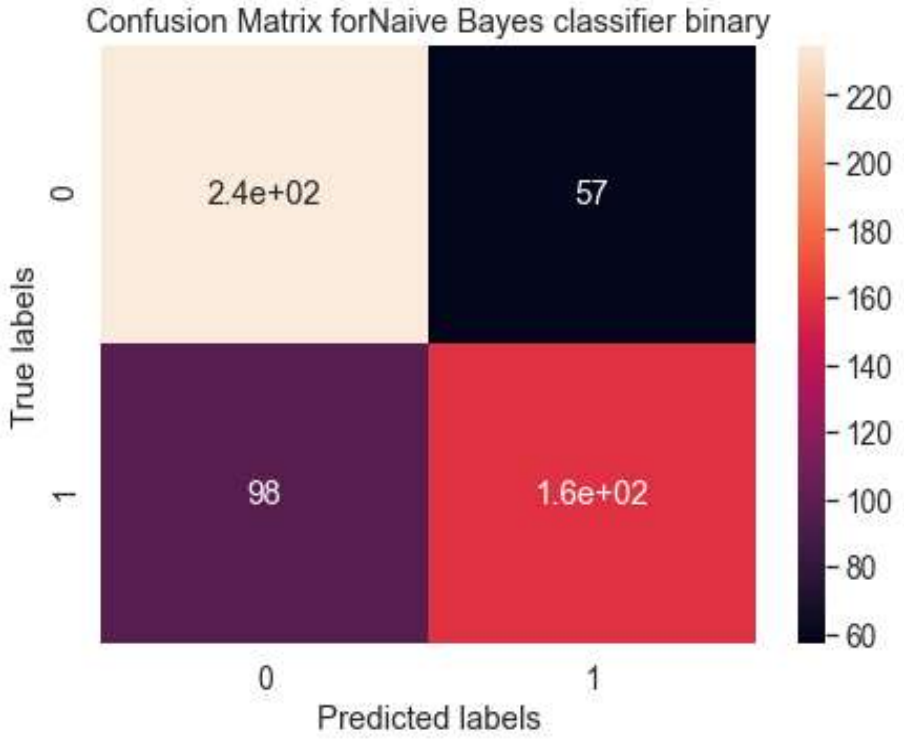

Figure 22:- The Confusion Matrix for Binary Naive Bayes Classifier.

As has become apparent, the results are better than the Multiclass Classification results but wanting in terms of comparison with SVM and Logistic Regression.

	precision	recall	f1-score	support
0	0.77	0.77	0.77	292
1	0.74	0.73	0.74	256
accuracy			0.76	548
macro avg	0.75	0.75	0.75	548
weighted avg	0.76	0.76	0.76	548

Figure 23:- The Classification report for Binary Naive Bayes Classifier.

Naive Bayes Classification model shows the best results out of all the Multiclass Classification algorithms, but the difference in the accuracy of the model is less significant when compared to the prediction of their Binary counterparts. The accuracy results are below SVM and Logistic Regression.

Results and Discussion:-

After creating the models, we now experimentally analyse the findings and insights revealed through the course of this study. In this section, we demonstrate the efficiencies of Multiclass and Binary classification algorithms and the intricacies that divide them in Football Prediction.

Table 1:- Precision, Recall and F1-scores produced for the Classification models.

Algorithm	Precision	Recall	F1-score
KNN Multiclass	0.6	0.6	0.59
Decision Tree Multiclass	0.63	0.63	0.62
Naive Bayes Multiclass	0.67	0.64	0.65
Logistic Regression	0.8	0.8	0.8
SVM	0.79	0.79	0.79
KNN Binary	0.75	0.75	0.75
Decision Tree Binary	0.73	0.73	0.72
Naive Bayes Binary	0.76	0.76	0.76

The Multiclass classification algorithms show lesser values of precision, recall and f1-score. The binary counterparts of multiclass algorithms like KNN, Naive Bayes and Decision Tree produce better results. This can be seen in the classification report as the model finds it difficult to classify draw matches. These metrics are used for evaluating a model with uneven class distribution, they are used in most real-life scenarios. From the table above, we can see that Logistic Regression predicts with the highest efficiency.

Next, we take a look at the accuracy metric for the different models. Accuracy is used when true positives and true negatives are given more importance over false positives and false negatives. Or, for even class distribution.

Table 2:- Accuracy of the Classification models.

Algorithm	Accuracy
KNN Multiclass	0.60
Decision Tree Multiclass	0.63
Naive Bayes Multiclass	0.64
Logistic Regression	0.8
SVM	0.79
KNN Binary	0.75
Decision Tree Binary	0.73
Naive Bayes Binary	0.76

The accuracy metrics give a better estimate than the unequal class distribution metrics suggesting that the data has an even class distribution. As expected, the Binary Classification algorithm outstrip the ambiguity produced by the Multiclass Classification algorithms. With Logistic Regression, SVM and Binary Naive Bayes Classification producing the highest accuracy results.

For a more comprehensive analysis, we take an in-depth look at the differences in Binary and Multiclass classification algorithms of KNN, Decision Tree and Naive Bayes Classifier.

Table 3:- Comparative analysis of Binary and Multiclass Classification.

Algorithm	Precision	Recall	F1-score	Accuracy
KNN Multiclass	0.6	0.6	0.59	0.60
KNN Binary	0.75	0.75	0.75	0.75
Decision Tree Multiclass	0.63	0.63	0.62	0.63
Decision Tree Binary	0.73	0.73	0.72	0.73
Naive Bayes Multiclass	0.67	0.64	0.65	0.64

Naive Bayes Binary	0.76	0.76	0.76	0.76

The difference in accuracy between KNN Multiclass and Binary is 15% for Accuracy and 16% for F1-score. For Decision Tree Classifiers there is a 10% increase in both F1-score and Accuracy. And finally, for Naive Bayes Classification there is an increase of 11% for f1-sore and an increase of 12% for accuracy results.

Conclusion:-

In the scope of this paper, we use a wide variety of Classification algorithms. There are intrinsic differences when Multiclass classification themselves are used in Multiclass or Binary problems. Binary classification algorithms produce significantly better results in accuracy, especially SVM and Logistic regression. Furthermore, we find that the best model for multiclass football prediction is the Naive Bayes Multiclass Classifier. The difference produced in accuracy and f1-score when Multiclass algorithms range from 10% to 16%. Thus, this paper neatly ties all the insights regarding Football Classification for sports analytics and presents them in a streamlined fashion in this paper.

References:-

[1] Rui Freitas, Anna Volossovitch, Carlos H Almeida (2020), Associations of situational and performance variables with defensive transition outcomes in FIFA World Cup, *SAGE journals*.

[2] Konstantinos Apostolou and Christos Tjortjis (2019), Sports Analytics algorithms for performance prediction, *Conference: 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA*)

[3] Pappalardo, L., Cintia, P., Rossi, A. *et al.* A public data set of spatio-temporal match events in soccer competitions. *Sci Data* 6, 236 (2019). https://doi.org/10.1038/s41597-019-0247-7.

[4] G. Vinué and I. Epifanio (2017), Archetypoid analysis for sports analytics ,Data Mining and Knowledge Discovery 31(1):1-35.

[5] Vangelis Sarlis, Christos Tjortjis, Sports analytics — Evaluation of basketball players and team performance, *Information Systems*, Volume 93, 2020, 101562, ISSN 0306-4379, https://doi.org/10.1016/j.is.2020.101562.

[6] Anand Ganesan , Harini M, English Football Prediction Using Machine Learning Classifiers, *International Journal of Pure and Applied Mathematics*, Volume 118 No. 22 2018, 533-536, ISSN: 1314-3395.

[7] Gangal, Anurag et al. "Analysis and Prediction of Football Statistics using Data Mining Techniques." *International Journal of Computer Applications* 132 (2015): 8-11.

[8]Maral Haghighat, Hamid Rastegari and Nasim Nourafza, A Review of Data Mining Techniques for Result Prediction in Sports, *ACSIJ Advances in Computer Science: an International Journal*, Vol. 2, Issue 5, No.6 , November 2013 ,ISSN : 2322-5157

[9] Ragini Singla, Dr. Amardeep Singh, Sports Prediction Using Machine Learning , International Journal of Emerging Technologies and Innovative Research, ISSN:2349-5162, Vol.7, Issue 10, page no. pp2759-2765.

[10] D. Prasetio and D. Harlili, "Predicting football match results with logistic regression," 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), 2016, pp. 1-5, doi: 10.1109/ICAICTA.2016.7803111.