Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/16481
DOI URL: http://dx.doi.org/10.21474/IJAR01/16481

RESEARCH ARTICLE

A NOVEL APPROACH FOR PREDICTING FOOTBALL MATCH RESULTS: AN EVALUATION OF CLASSIFICATION ALGORITHMS

Skanda Aithal[1] and Dr. S.K. Manju Bargavi[2]

1. MCA, Department of Computer Science and Information Technology, Jain (Deemed-to-be-University) Bengaluru, India.
2. Professor, Department of Computer Science and Information Technology, Jain (Deemed-to-be-University) Bengaluru, India.

Manuscript Info

Manuscript History
Received: 19 January 2023
Final Accepted: 24 February 2023
Published: March 2023

Key words:-
Data Science, Feature Engineering, Football Analytics, Machine Learning, SVM

Abstract

The task of predicting the outcomes of football matches is rendered increasingly complex by the intricate nature of the game and the variety of variables that could affect how things turn out. In the recent past, machine learning algorithms have been applied to this challenge, with varying degrees of success. In this particular research paper, we have meticulously evaluated the performance of several classification algorithms with the objective of predicting the outcomes of football matches in a tournament setting. The algorithms that were thoroughly tested encompassed a diverse range of classification models, including logistic regression, support vector machines and random forests. The study employed a dataset of historical match data drawn from the FIFA World Cup, historical team ranking data and team strength data from FIFA games. In order to accurately assess the efficacy of the algorithms tested, the evaluation metrics used were accuracy, precision and recall. The results of the study highlight the fact that machine learning algorithms can indeed prove to be effective tools for predicting the outcomes of football matches.

Introduction:-

Football, a globally recognized sport also known as soccer in some regions, has a massive following of over 3.5 billion enthusiasts, bettors, and analysts. This has fueled discussions on predicting football match outcomes, which is a daunting task due to the game's intricate nature and numerous variables that can influence the final result. These variables could encompass a wide range of factors from individual player performance, team tactics, weather conditions, pitch conditions to the crowd atmosphere. The high unpredictability of football has made it an enthralling sport, but a challenge for those seeking accurate match predictions.

In recent times, machine learning algorithms have emerged as a promising tool for addressing this challenge. These algorithms use mathematical techniques to analyze data and learn from it, aiming to make accurate predictions or classifications. Historical football match data can be used to train these algorithms to identify trends and patterns that could be help in forecasting future match outcomes.

Corresponding Author:- Skanda Aithal
Address**:-** MCA, Department of Computer Science and Information Technology, Jain (Deemed-to-be-University) Bengaluru, India.

Machine learning algorithms applied to predicting football match outcomes have received increasing attention in the research community. Studies have investigated various machine learning techniques such as decision trees, support vector machines, and neural networks, to predict football match outcomes. These studies have revealed that machine learning algorithms provide valuable insights into the factors that impact the outcomes of football matches and can improve prediction accuracy.

The objective of this research paper is to evaluate the performance of different classification algorithms for predicting football match outcomes in a tournament setting. The range of classification models tested encompasses logistic regression, svm, and random forests. A dataset of historical match data from the FIFA World Cup, as well as historical team ranking data and team strength data from FIFA games, was employed for the study. Accuracy, precision, and recall were the evaluation measures used to gauge how well the evaluated algorithms performed. We have predicted the 2022 FIFA World cup knockout stages and verified them with the real-world results.

The paper is set up as follows. In Section 2, we present a comprehensive literature review of previous research on predicting football match outcomes using machine learning algorithms. Section 3 outlines the methodology utilized in this study, including data preprocessing, feature engineering, classification algorithms, and evaluation metrics. Section 4 displays the results of the experiments, followed by a discussion of the findings in Section 5. Finally, Section 6 concludes the paper with a summary of the primary contributions and suggestions for future research.

Related works:-

With billions of supporters and a long history of professional and amateur competition, football is one of the most widely watched sports in all parts of the globe. Through the years, many researchers have used data science and machine learning techniques to analyze and understand various aspects of football, such as team performance, player performance, and tactics and predict outcome of the matches.

1. Predicting the outcomes of football matches using team and player ratings.

One area of particular interest is the analysis of team performance. Researchers have used a variety of techniques to analyze team performance, including descriptive statistics, machine learning, and visualization. For example, some studies have used regression models to predict the outcome of football matches based on various features such as team rankings, home field advantage, and past performance. This article [1] examines and contrasts the effectiveness of team grading and individual player grading for predicting outcomes of football matches. The Elo rating system, a popular method for computing team ratings, is employed, while a modified version of plus-minus ratings is utilized to rate individual players. Two statistical models were used to generate forecasts. The models included an ordered logit regression model and a competing risk model. While both models were capable of generating pre-game forecasts, only the latter could be used for predictions when the game is taking place. Results portrayed that there was no significant deviation in prediction quality between the OLR and competing risk models for forecasting prior to the game starting. Also, both team and player ratings were effective at forecasting match results. However, combining team and player ratings led to significantly better forecasts compared to using only one of the ratings. While a more basic simplistic model that simply takes goals rates into account is sufficient for pre-game projections, using an advanced complex model that calculates rates of both scoring and cards for fouls leads to more accurate in-game predictions.

2. Classification Algorithms on Football Analytics

In the article [2], the authors use sports analytics to predict the outcomes of football matches in five major leagues. They use two types of classification: multiclass classification, which predicts a team's win, loss, or draw, and binary classification, which predicts a team's win or loss/draw. The study evaluates different machine learning algorithms including Naive Bayes, Decision Tree, K-Nearest Neighbours, Support Vector Machines, and Logistic Regression, and compares their accuracy and efficacy based on f1-score, recall, and precision. The results of the study show that binary classification algorithms like SVM and logistic regression produce the best accuracy, while Naive Bayes Multiclass Classifier is the best multiclass classifier. The study also provides insights into the variables affecting match outcomes and the most suitable classification algorithm for prediction. The authors use supervised learning algorithms and data visualization to present their results in a clear and concise manner.

This study demonstrates the potential of using sports analytics in forecasting the outcome of football matches. The use of advanced algorithms and machine learning models has allowed for a more sophisticated analysis of the sport, helping to uncover new insights and understandings. The results of this study highlight the importance of incorporating technology and data-driven analysis into the decision-making process in sports, and pave the way for

further research in the field of sports analytics. By continuing to refine and develop these methods, we can gain a deeper understanding of the sport and help teams to perform at their highest level.

In the study [3], the process of using machine learning algorithms to predict the results of football matches was described with the aim of supporting sports betting. The study collected data from two sources, one for previous game statistics and the other for team attributes, and analysed the data to determine which variables to include in the models. the study analyzed various algorithms and utilized data from four different seasons to train the models. The models were then tested using data from the 2017 season of the English first division league. The research thoroughly evaluated the methodology of predicting football match outcomes to support forecasting. To create the most accurate prediction models, the study experimented with different combinations of variables. The findings demonstrated that the best models, with a better success rate than the initial model, were developed by using machine learning algorithms such Support Vector Machine (SVM), Random Forest (RF), Xgboost, and Recurrent Neural Network (RNA). The success percentage of the original model was 61.32%. however, after applying the aforementioned techniques, the success rate dramatically increased, reaching 65.26%.

Testing combinations of the 8 variables and the 7 most crucial variables, for a total of 15 variables, produced the ideal model. This approach demonstrated the importance of considering multiple variables and the interplay between them in making accurate predictions.

Team performance prediction and player performance prediction were the two scenarios of sports analytics that were explored in the research [4]. For the first case, two approaches were followed to predict the performance of teams in four European leagues during the 2018-19 season. The first method, which had a 70% success rate, attempted to categorise teams into those that will do better or worse than the previous campaign. The second method simulated every game of the season and made league table and match predictions, achieving an AccuracyM of 57% for the EPL and a lowest RMSE of 9 for the Spanish league. It also correctly predicted the winner of the league 64% of the time and the European qualification teams 75% of the time. For player performance prediction, multiple linear regression with backward elimination was used to find the 13 features that significantly influence a central defender's match rating, with a R-squared of 0.907 and an adjusted R-squared of 0.88. The characteristics show the difference in playing style for central defenders by including traditional defensive actions, player attributes and also the passing and offensive match actions.

3. Data-driven performance indicators for football teams

By a thorough review of pertinent literature, the study [5] sought to identify important variables with predictive value for match outcome. For this, a dataset was compiled from thirteen seasons of match data in the Eredivisie, obtained from publicly available sources. The training set was then subjected to rigorous testing of a large number of dimensionality reduction techniques and classification algorithms to identify the optimal set of techniques. The results of this study were that the highest prediction accuracy was achieved through the use of PCA (with a 15% variance reduction) in conjunction with either a Naive Bayes or Multilayer Perceptron classifier. The study also created models for gambling odds and a feature set (combining both public data and betting odds), and found that the hybrid feature set had an accuracy similar with the betting odds model. This led to the conclusion that it could be possible to build a system that could outperform bookmakers, using a combination of open data and betting odds. This research shows the power of data science in the world of sports, and demonstrates the potential for using data-driven approaches to gain a competitive advantage in the field of football. Whether it's predicting match outcomes, identifying key performance metrics, or uncovering hidden trends, the use of analytics has the capability to transform the way in which the sport is played, managed, and analyzed.

Sports analytics, especially in football (soccer), have become popular due to the availability of high-quality data from games through various kinds of sensing technologies. Data analysis companies like Prozone and Opta provide this data to all the major entities like clubs and leagues to monitor their performance. Fans also use these statistics for enjoyment and critical analysis. Despite the growth in amount of data, the potential of football analytics has not been fully exploited, with limited use of data science techniques. A data-driven approach can greatly enhance the comprehension of team performance. The authors of the study [6], propose a single value, the "H indicator," representing a team's passing behavior, which shows a strong correlation with team success. They conducted two analyses using event data from four European football leagues and found that wins, losses, and draws can be predicted with high accuracy based on the H indicator. They also conducted a complete simulation of each of the four leagues and found close agreement with actual rankings, with the strongest teams having the highest

performance indicators. This study shows that football analytics has only just started to unveil the trends and behavior of performance.

This study analyzed 1,446 soccer matches from four major European leagues using data science. The results showed that a team's passing activity is related to its victory and different indicators were extracted to measure different characteristics of a team's passing activity. The indicators obtained were used to express a team's performance during a match. The study proposed two methods for predicting game outcomes and simulating games across four leagues, respectively. The study found that the suggested indicators were effective in describing team performance, but acknowledges that adding details on the difficulty of passes, defensive events, and player performance would improve the model.

The evolution of football statistics, thanks to high-fidelity data streams provided by sensing technologies, has led to the use of data by professional analysis firms to help football clubs improve their performance. The standard approach uses history-related factors to evaluate and predict team performance, but this paper proposes a different model that observes players' behavior on the pitch. The authors model a team's game as a network and draw out simple network measures to show their value in predicting the outcome of a tournament like the Italian major league. The authors show that their data-driven approach, which uses network measures to evaluate the performance of a team, has a correlation with the team's success during the competition. The authors ran a simulation using the 2014 FIFA World Cup and the 2013–2014 season of the Italian Serie A, and they found that their method beats other models, such as naive models and models based on historical data. The authors came to the conclusion that their indications are a good gauge of team performance and that a more thorough analysis, taking into account defence tactics and off-the-ball movements, has the potential to uncover hidden patterns and superior conduct.

Proposed Methodology:-

This section describes the sequence of activities that occur before to the experiments as well as the methods used to collect and prepare the data. Fig. 1 shows the block diagram that encapsulates the procedure.

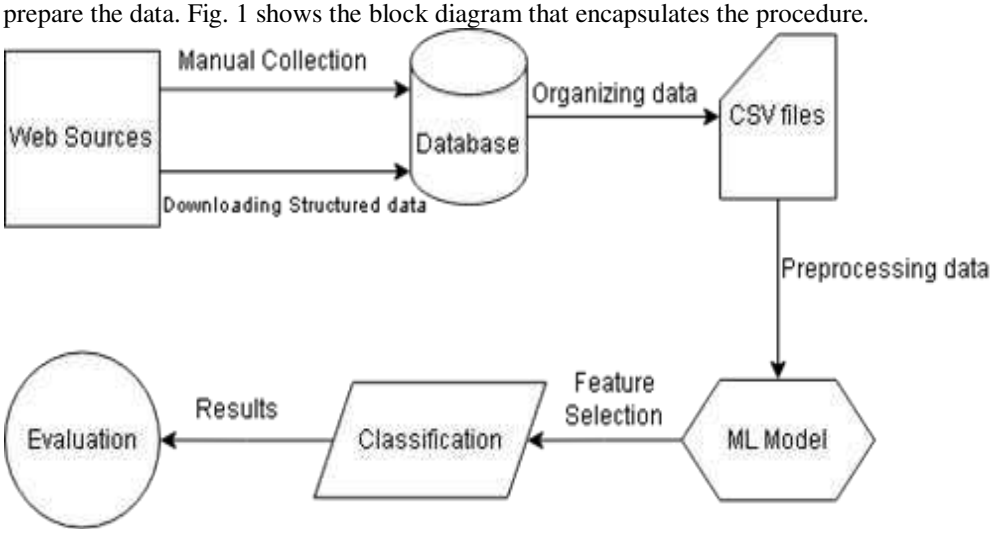

Figure.1:- Process flow diagram.

The methodology employed in this research paper for predicting football match outcomes involves the following steps:

1. Data Preprocessing

The dataset for the experiments was prepared by cleaning the data obtained through scraping and manual collection, which consists of historical data of results of world cup matches, national team rankings for each year and team strength ratings from respective year's FIFA games. The data considered ranges from the year 2005 to 2022. The data was cleaned, and any missing or erroneous values are imputed using appropriate techniques. The data was also standardized to ensure that all features are on a similar scale. The below figure contains the snippet of processed data. We can observe that there are 17 columns.

	att1	def1	mid1	ovr1	rank1	att2	def2	mid2	ovr2	rank2	score	winner	att	def	mid	ovr	rank
0	74	71	75	73	49	90	86	88	88	8	-2	-1	-16	-15	-13	-15	41
1	81	69	74	73	46	65	63	58	62	66	2	1	16	6	16	11	-20
2	74	80	75	76	19	83	75	76	77	13	1	1	-9	5	-1	-1	6
3	85	77	80	80	14	75	81	78	79	18	1	1	10	-4	2	1	-4
4	89	88	88	88	9	74	69	65	68	74	1	1	15	19	23	20	-65
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1324	70	68	71	71	60	87	82	85	85	3	-2	-1	-17	-14	-14	-14	57
1325	81	83	83	82	11	79	74	77	76	37	2	1	2	9	6	6	-26
1326	59	66	66	66	78	79	78	79	79	15	3	1	-20	-12	-13	-13	63
1327	74	69	70	71	50	71	71	73	72	24	-1	-1	3	-2	-3	-1	26
1328	74	73	76	75	21	74	75	72	73	55	-2	-1	0	-2	4	2	-34

1329 rows × 17 columns

Figure.2:- Snapshot of processed data.

2. Feature Engineering

The features used for the experiments were team FIFA ranking, team strength data which included defense, midfield, attacking and overall ratings of each team and their difference. We have labelled our target variable as results which can be classified as 1(Win), 0 (Draw) and -1 (Loss).

3. Classification algorithms and evaluation metrics

The algorithms used were logistic regression, random forests, and support vector machines. The performance of each algorithm is evaluated using accuracy, precision and recall metrics. The proportion of accurately anticipated outcomes is measured by accuracy.

Implementation and Results:-

The results of the experiments are presented in Table 1. The table shows the accuracy, precision, and recall for each of the classification algorithms tested. The results show that support vector machines and logistic regression perform similarly whereas random forest in slightly worse than the other two algorithms.

Table 1:- Results - Evaluationof algorithms.

Classifier	Accuracy	Precision	Recall
Logistic Regression	0.77	0.81	0.78
Random Forest	0.71	0.77	0.72
SVM(Linear)	0.77	0.82	0.77

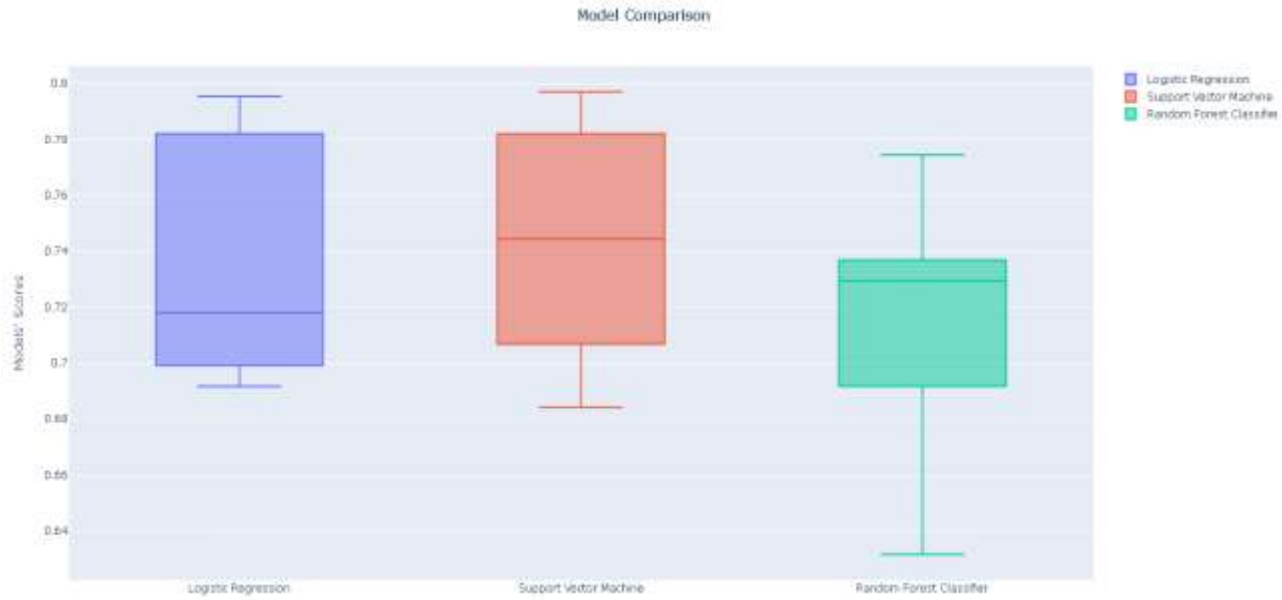

Figure.3:- Model Comparison.

We can observe from fig. 3, SVM and logistic regression have similar scores but SVM is more consistent. Hence SVM was chosen as the algorithm for predicting the matches of FIFA World cup 2022 knockout stages.

1. Simulation

The matches of knockout stages of the Football World cup 2022 was acquired and the entire bracket was simulated using SVM algorithm

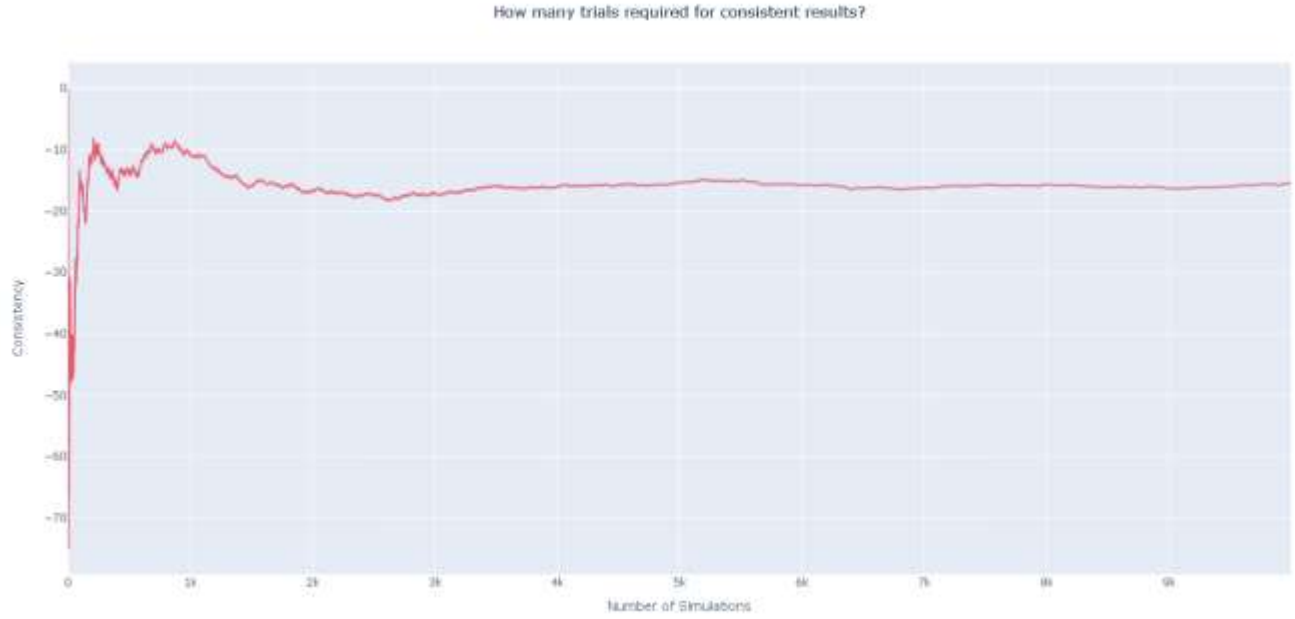

Figure.4:- Graph to show number of trials needed for consistent result.

From the graph in fig.4, we can see that we can get consistent results if we run the simulation for approximately 6500 times. So, we run the simulation for 6500 times and get the probabilities of winning for each team in a match to decide who would win the game. The result obtained from this is given in the below figure.

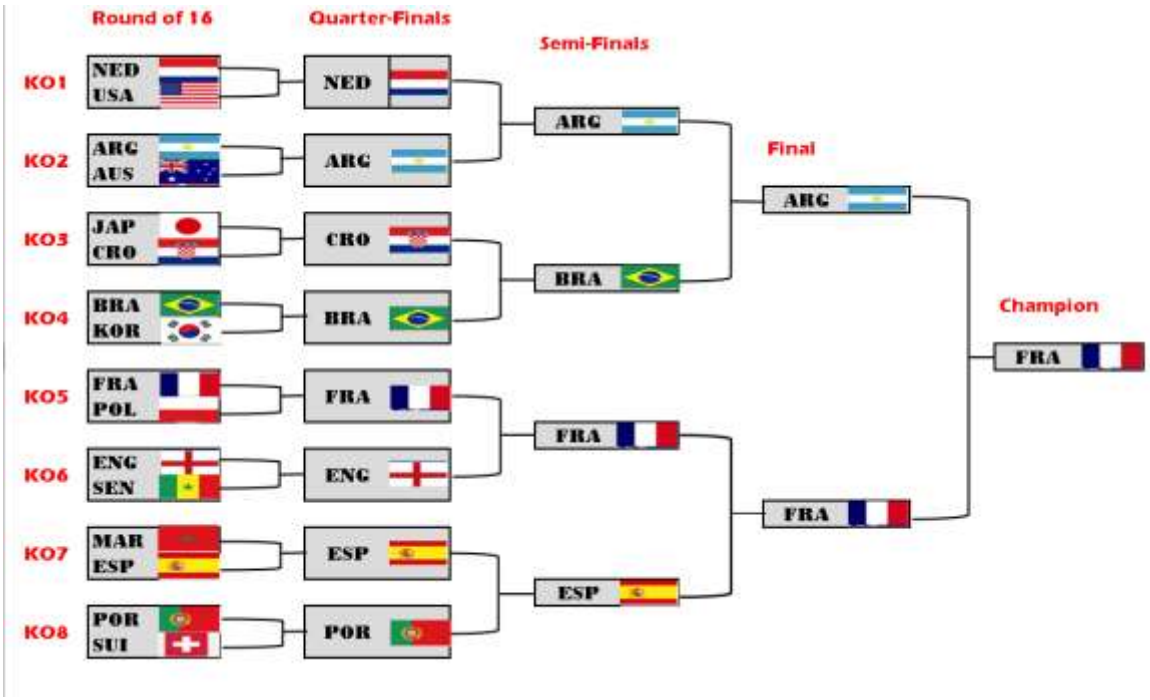

Figure.5:- Knockout stage simulation.

Discussion:-

FIFA World cup 2022 has been concluded and we have the results of the tournament. This helps us in verifying the performance of our simulation with actual results. As we can observe from fig.5, we have some differences from the actual results. This may be because of surprising results in the matches in earlier rounds which changes the results of the entire bracket in the simulation. To get better results we simulated knockout stages match by match individually based on the actual real-world results.

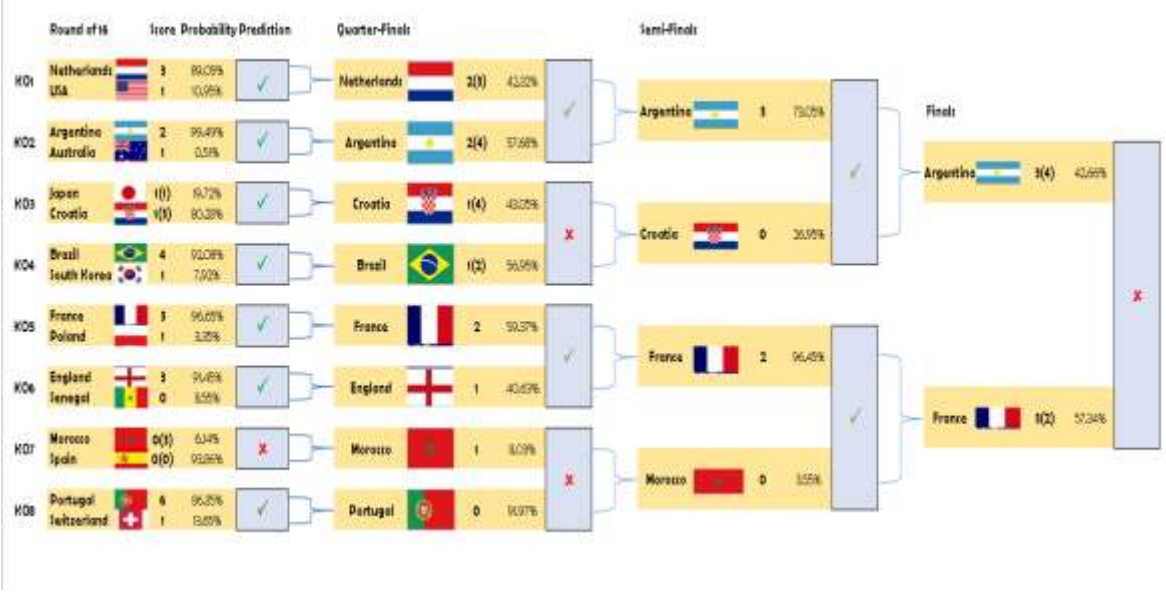

Figure.6:- Individual match result prediction.

We get much better results when we predict each match individually as observed in fig.6.The actual scores of each match and the probabilities of each team winning a particular match gathered from simulating the matches using our model is mentioned in the above figure. We could predict the winner of 11 out of 15 knockout stage matches correctly. Through these results we get to know the teams which overachieved and underachieved. Morocco and

Croatia were the over-achievers of this world cup according to our model and we can say that it aligns with the common belief among the experts of the game.

Conclusion:-

In this research paper we have showcased the potential of machine learning algorithms in forecasting football match outcomes. The research study used a historical dataset of match data from the FIFA World Cup, along with team ratings data from FIFA games. A variety of classification algorithms, including random forests, logistic regression, and support vector machines, were employed and evaluated using a multitude of accuracy, precision, and recall metrics. The support vector machines, with their superior performance and consistency, proved to be the most effective classification algorithm for predicting football match outcomes

However, it is essential to note that this research study has some limitations. The dataset used in the study is restricted to the FIFA World Cup which could limit the applicability of the results to other football leagues and tournaments. To enhance the generalizability of the findings, future research could broaden the dataset to include more football leagues and tournaments. It is also important to factorize that the team strength ratings data obtained from the FIFA games is not intrinsic to the actual football game. Furthermore, future research could explore the use of more complex algorithms such as deep learning models for predicting football match outcomes. Deep learning models have displayed promise in other areas of machine learning and could potentially improve the accuracy of predictions for football match outcomes. In a nutshell, this research paper serves as a testimony to the tremendous potential of machine learning algorithms in predicting football match outcomes.

References:-

[1] Arntzen H, Hvattum LM. Predicting match outcomes in association football using team ratings and player ratings. Statistical Modelling. 2021;21(5):449-470. doi:10.1177/1471082X20929881

[2] Karan Bhowmick and Vivek Sarvaiya (2021); A COMPARATIVE STUDY OF THE DIFFERENT CLASSIFICATION ALGORITHMS ON FOOTBALL ANALYTICS Int. J. of Adv. Res. 9 (Aug). 392-407, (ISSN 2320-5407).

[3] Rodrigues, Fátima & Pinto, Ângelo. (2022). Prediction of football match results with Machine Learning. Procedia Computer Science. 204. 463-470. 10.1016/j.procs.2022.08.057.

[4] Chazan-Pantzalis, Victor &Tjortjis, Christos. (2020). Sports Analytics for Football League Table and Player Performance Prediction.

[5] Tax, Niek &Joustra, Yme. (2015). Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach. 10.13140/RG.2.1.1383.4729

[6] Cintia, Paolo & Pappalardo, Luca & Pedreschi, Dino & Giannotti, Fosca & Malvaldi, Marco. (2015). The harsh rule of the goals: Data-driven performance indicators for football teams. 10.1109/DSAA.2015.7344823.