International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 3, ISSN No. 2455-2143, Pages 83-91
Published Online July 2020 in IJEAST (http://www.ijeast.com)

PREDICTING EPL FOOTBALL MATCHES RESULTS USING MACHINE LEARNING ALGORITHMS

Sayed Muhammad Yonus Saiedy
Bakhtar University
Kabul, Afghanistan

Muhammad Aslam HemmatQachmas
Balkh University
Balkh, Afghanistan

Dr. Amanullah Faqiri
Bakhtar University
Kabul, Afghanistan

***Abstract*- Machine learning is a subset of artificial intelligence (AI) in which algorithms learn by example from historical data to predict outcomes and uncover patterns that are not easily spotted by humans. Machine learning evolved from the study of pattern recognition and explores the notion that algorithms can learn from and make predictions on data. And, as they begin to become more 'intelligent', these algorithms can overcome program instructions to make highly accurate, data-driven decisions. Predictive analytics encompasses a variety of statistical techniques (including machine learning, predictive modelling and data mining) and uses statistics (both historical and current) to estimate, or 'predict', future outcomes. These outcomes might be behaviors of a customer likely to exhibit or possible changes in the market, for example. Predictive analytics help us to understand possible future occurrences by analyzing the past. In this research it's intended to combine machine learning algorithms with predictive analytics to do predictions on sports results specially football matches result prediction.**

Keywords: Machine learning, artificial intelligence, pattern recognition, prediction, Predictive analytics, statistical techniques, sports, and football match results.

I. INTRODUCTION

To put it simply, the machine learning algorithm learns by example, and then users apply those self-learning algorithms to uncover insights, determine relationships, and make predictions about future trends (https://www.datarobot.com/wiki/machine-learning, 2020).

Machine learning (ML) is one of the intelligent methodologies that have shown promising results in the domains of classification and prediction. One of the expanding areas necessitating good predictive accuracy is sport prediction, due to the large monetary amounts involved in betting. In addition, club managers and owners are striving for classification models so that they can understand and formulate strategies needed to win matches (Bunker, R. P., and Thabtah, F., 2017).

In the area of sports result prediction we can find many researches such as, in (https://www.coursehero.com/file/23471228/SteffenSmolka-BeatingTheBookies-poster/, 2020) the project demonstrates that it is possible to predict the outcome of soccer games win, tie, or loss with over 50% accuracy automatically, in the research (Esumeh, E. O., 2015) the author proclaims that making prediction in sports, football in particular is tasking and very intricate but this can be made even simpler and easier by the implementation of machine learning rather than mathematical analysis, the author in (Stenerud, S. G., 2015) has given a review of existing literature in Soccer result-prediction, and has expanded on existing models by including data that previously were not considered and developed 6 different models for prediction with increasing complexity and in (Kampakis, S., and Thomas, W., 2015) the researcher says that it is possible to predict the winner of English county twenty twenty cricket games in almost two thirds of instances.

In this research I intend to get an accurate result for the football matches by using machine learning algorithms, In addition I chose English Premier League (EPL) as the test bed for the research due to the competitiveness and popularity of the league rather than other leagues in the world. In this research the research question is as follows:

How to make an accurate *prediction on EPL matches results* using machine learning algorithms?

To reach the above research question there have been considered the following objectives:

- To explore a model for predicting the EPL matches accurately.
- To create a model than can predict the rankings of the teams (the final table of the rankings of the teams in the league).
- Testing the dataset for the machine learning algorithms to find an accurate predicting model.

As it's mentioned above the test bed for the research is EPL the desired machine algorithms are Support Vector Machines and random forest classifier.

II. LITERATURE REVIEW

Most of the work on this task has been done by gambling organizations for the benefit of odds makers. However, because our data source is public, several other groups have taken to predicting games as well. One example of the many that we examined comes from a CS229 Final Project from autumn 2013 by (Timmaraju et al. 2013). Their work focused on building a highly accurate system to be trained with one season and tested with one season of data. Because they were able to use features such as corner kicks and shots in previous games, as well as because their parameters were affected by such small datasets, their accuracies rose to 60% with an RBF-SVM. We chose to focus on a much larger training set with the focus on building a more broadly applicable classifier for the EPL.

Betting in the sports is a global business which lots of billion dollars invested in it, popularly the betting market of United Kingdom which is formatted by the fixed odds that means all odds are determined by the bookmakers a number of days before every match to be played. They do not update the odds by the number of betting or the player status in this case they can make good predictions of the matches.

The author in article (Goddard, J., 2005) has compared the performance of professional British odds-setters to an ordered probability model during five seasons of 1998/1999 to 2002/2003. At the beginning of the model deployment the statistical model was beating the performance but the end processes the odds-setters' predictions performed better, this performance casts doubt prediction of the statistical models performing better than experts predictions. This impression is because of the case that tipsters and independent experts whose predictions broadcasted in daily newspapers, generally perform poorly compared to statistical models (Spann, M., and Skiera, B., 2009). In contrarily, it's not fair to compare the professional odd-setters predictions to these statistical models due to the differences in the forecasting performance

The article (Reep, C., and Benjamin, B., 1968). is one of the first researches done on the football matches prediction. They predicted the matches based on some negative binomial distributions to scores from each football match, in conclusion they couldn't predict the results of the matches reliably.

Another attempt at predicting soccer works was done by Joseph et al. This group used Bayesian Nets to predict the results of Tottenham Hotspur over the period of 1995-1997. Their model falls short in several regards: self-admittedly, it relies upon trends from a specific time period and is not extendable to later seasons, and they report vast variations in accuracy, ranging between 38% and 59%. However, they do provide useful background on model comparison and feature selection (Joseph, A., Fenton, N. E., and Neil, M., 2006) is also result-based but take a very different approach: it compares the performance of an expert constructed Bayesian network to the performance of several models trained by machine learning algorithms. They conclude that the expert BN is generally superior to the automatic techniques.], The authors specifically choose to focus on predicting match outcomes of a single team - Tottenham Hotspur. Also, their model is dependent on the presence of particular players and is thus, by their own admission, not scalable. There is no standard way of reporting results for this problem.

Rue et al., who used a Bayesian linear model to predict soccer results. Notably, they used a time-dependent model that took into account the relative strength of attack and defense of each team (Rue, H., and Salvesen, O., 2000). A Bayesian dynamic model is built, but there are many parameters and the authors do not justify the choice of their values, which may not work well with other testing sets. Also, out of the 48 EPL games they bet on, they won on only 15 games. And indicates that they didn't have data available regarding statistics for each individual player, we did take into account the notion of time dependence in our model.

In the reference (Stenerud, S. G., 2015) the author has intended to develop a six deferent models which can beat the betting strategies, the author has developed a model for result prediction in soccer. The model is based on chances created being modeled as a Poisson process while goals scored is seen as a result of first creating chances and then converting them, here modeled as a Bernoulli trial.

In (Zaveri, N. et al, 2018) the researchers have come up with a solution using machine learning algorithms that can fulfil all the current needs in football match prediction. The implementation only includes teams from Spanish La Liga over the last 5 seasons. We have predicted the outcome of matches between Home Team and Away Team which would include the final score, the starting 11 players, the substitutes and the names of probable goal scorers. For the purpose of analysis, we have provided the stats of players and teams referring to the FIFA 18 game database as well as their actual career stats. We have also provided the analysis of strength, weakness and tactics of players and teams. Finally, for decision making purposes, we would make a system that can analyze the Home and Away team and then suggest the tactics to the user for their team that can maximize their winning chances. They implemented the model using different machine learning algorithms and were able to reach the accuracy of 71.63% with Logistic Regression on the Match History Database of 5 seasons along with the Team Vs Team Database.

The author demonstrates that it is possible to predict the outcome of soccer games (win, tie, or loss) with over 50% accuracy automatically in reference (Steffen. S, 2107). In his project the researcher used the outcome of previous games for prediction, and I expect more sophisticated features such as the fatigue of a team can enable predictions with an accuracy beyond 60%. And also he has discovered that it seems to be much easier to predict the last games of a season. In particular, he achieved over 72% accuracy for predicting the final games of the 2015/16 EPL season.

The reference (Goddard, J., 2005) compares the two approaches and concludes they are of similar predictive power, but suggest that hybrid approaches may perform best.

The work of (Karlis, D., and Ntzoufras, I., 2003) is goal-based, using Poisson distributions to model the number of goals scored by a team. Similar to my work, they use attack and defense parameters; but in contrast to my work, their models are team-dependent. Unfortunately, neither of the papers report how many outcomes they can predict correctly.

In particular, the focus in (Angelini, G., and De Angelis, L., 2017) is mainly on how well the model can fit existing data. The PhD thesis of (Constantinaou, A. C., et al, 2012) is result-based and uses a Bayesian network based on team strength, form, psychological impact, and fatigue. The thesis describes a model based purely on objective data, and a model that incorporates subjective estimates from a human expert. It concludes that the accuracy of the purely objective forecasts is significantly inferior to bookmaker's forecasts, while the subjective model is on par.

The research paper (Liu, F., Shi, Y., and Najjar, L., 2017) explores three easy to grab factors and uses DOE (Design of Experiment) method to determine the effects on sports result prediction. Another goal is to explore the three factors' importance. This research data modeling is based on the NBA (National Basketball Association) game data. All the data is extracted from an official and public data source of 2015 to 2016 regular season.

In (Yezus, A., 2014) the author claims that the machine learning methods can be applied to different fields, including sports, for instance English Premier League it is shown that it is possible to find a classifier that predicts the outcome of soccer matches with the precision of more than 60%. And demands that, there is still a lot of work to be done and the research will be proceeded.

Another study (Blundell, J., 2009) has proved that American Football matches can be accurately modelled using features within a regression model. It was also discovered that simple logistic model could achieve just as accurate forecasts compared to some more complex alternatives.

Concerning football a research (Hamadani, B., 2005) offered approach that in the study 3 seasons were tested and it was discovered that each season a different set of features had more significance which means either that football is an unstable game or that optimal features lie somewhere between those found by the study.

In (Elfrink, T., 2018) the author has investigated if it can be predicted which team will win individual MLB games? They used historical data of games and used different machine learning algorithms such as random forests and XGBoost to do get the result from their research. When using the XGBoost algorithm it showed the best results with an accuracy of 0:5552. The author claimed that this result can be improved when using more data, more computing power and better feature engineering.

The article (Karlis, D., and Ntzoufras, J., 2003) indicates a bivariate Poisson model to forecast number of goals scored by each team in a match. (Baio, G., and Blangiardo, M., 2010) proposed a Bayesian hierarchical model to predict the outcomes of the Italian Serie A league in the 2007/08 season. On the other hand, the reference (Hill, I. D., 1974) proves a

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 3, ISSN No. 2455-2143, Pages 83-91
Published Online July 2020 in IJEAST (http://www.ijeast.com)

significant correlation between the predictions made by football experts and the final league tables of the 1971–1972 season. The article (Dixon, M. J., and Coles, S. G., 1997) shows that the prediction of the well performed teams are not so hard but the prediction of the individual game is more challenging.

The authors in (Bunker, R. P., and Thabtah, F., 2017) creates a model which records a team's attacking and defensive abilities. On the matches of the 2013/2014 and 2014/2015 English Premier League seasons, their model outperformed the model by Dixon and Coles (1997) based on number of predicted goals (Dixon, M. J., and Coles, S. G., 1997).

To predict outcomes of the 2002 FIFA World Cup, the authors in (O'Donoghue, P. G., et al, 2004) used a different methods consisting probabilistic neural networks, linear and logistic regression, bookmakers' odds, computer simulations, and expert forecasts. There also have been done investigations on rating systems to predict the result of football games. Which the rating system in (Elo, A. E., 1978) is the best known method that was proposed by Arphad Elo on prediction the chess games and later deployed to football (Hvattum, L. M., and Arntzen, H., 2010).

Multilayer perceptron (MLP) with back-propagation learning rule is deployed in (Huang, K. Y., and Chen, K. J., 2011) to predict the winning rates of two teams according to their official statistical data of 2006 World Cup Football Game at the earlier stages. The training samples are used from three class: win, draw, and loss. At the new stage, new training samples are selected from the previous stages and were added to the training samples, then they construct the neural network. In this model they finally achieved the accuracy of 75% by excluding the games which finishes as a draw.

III. METHODOLOGY

3.1. Challenges and limitations

The data set contains the English premier league matches results from season 2013/2014 till 2018/2019 [http://www.football-data.co.uk/englandm.php] as a public data. Dataset in CSV format containing 64 columns (features) and 380 rows (matches) for each season. In here we use the public data so the accuracy will be not as much as the odds and betting houses. And the data for all teams which are relegating from the English premier league to championship would not be on hand therefore its considered to be predicted from the available data and the data which is not on hand there would be two zeroes it means the teams have made a draw.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
1	Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee	HS	AS	HST	AST	HF	AF	HC	AC	HY	AY	HR	AR	B365H	B365D	B365A	BWH	BWD	BWA	IWH	IWD	IWA	LBH
2	E0	13/08/16	Burnley	Swansea	0	1	A	0	0	D	J Moss	10	17	3	9	10	14	7	4	3	2	0	0	2.4	3.3	3.25	2.45	3.1	2.95	2.5	3.3	2.7	2.5
3	E0	13/08/16	Crystal Palace	West Brom	0	1	A	0	0	D	C Pawson	14	13	4	3	12	15	3	6	2	2	0	0	2	3.3	4.5	2	3.2	3.9	2.1	3.3	3.3	2
4	E0	13/08/16	Everton	Tottenham	1	1	D	1	0	H	M Atkinson	12	13	6	4	10	14	5	6	0	0	0	0	3.2	3.4	2.4	2.95	3.2	2.4	2.7	3.3	2.5	3.1
5	E0	13/08/16	Hull	Leicester	2	1	H	1	0	H	M Dean	14	18	5	5	8	17	5	3	2	2	0	0	4.5	3.6	1.91	4.33	3.4	1.9	3.3	3.3	2.1	4.5
6	E0	13/08/16	Man City	Sunderland	2	1	H	1	0	H	R Madley	16	7	4	3	11	14	9	6	1	2	0	0	1.25	6.5	15	1.22	6	11.5	1.3	5.5	10	1.3
7	E0	13/08/16	Middlesbrough	Stoke	1	1	D	1	0	H	K Friend	12	12	2	1	18	14	9	6	3	5	0	0	2.38	3.2	3.4	2.25	3.1	3.25	2.3	3.3	2.9	2.3
8	E0	13/08/16	Southampton	Watford	1	1	D	0	1	A	R East	24	5	6	1	8	12	6	2	1	2	0	1	1.8	3.75	5	1.8	3.4	4.5	1.8	3.5	4.2	1.8
9	E0	14/08/16	Arsenal	Liverpool	3	4	A	1	1	D	M Oliver	9	16	5	7	13	17	5	4	3	3	0	0	2.4	3.5	3.1	2.35	3.3	2.9	2.3	3.3	2.9	2.4
10	E0	14/08/16	Bournemouth	Man United	1	3	A	0	1	A	A Marriner	9	11	3	7	7	10	4	2	0	1	0	0	4.75	3.6	1.85	4.6	3.5	1.75	4.5	3.5	1.8	4.8

	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM
1	LBD	LBA	PSH	PSD	PSA	WHH	WHD	WHA	VCH	VCD	VCA	Bb1X2	BbMxH	BbAvH	BbMxD	BbAvD	BbMxA	BbAvA	BbOU	BbMx	BbAv	BbMx	BbAv	BbAH	BbAHh	BbMxAHH	BbAv	BbM	BbAvAHA	PSCH	PSCD	PSCA
2	3.3	3.1	2.47	3.32	3.19	2.5	3.2	2.9	2.5	3.2	3.3	55	2.55	2.43	3.35	3.21	3.3	3.1	40	2.4	2.3	1.7	1.6	32	-0.25	2.13	2.1	1.9	1.81	2.79	3.16	2.89
3	3.3	4.3	2.06	3.29	4.32	2.05	3.1	4	2	3.3	4.4	56	2.1	2.01	3.4	3.23	4.5	4.16	38	2.7	2.5	1.6	1.5	33	-0.5	2.07	2	1.9	1.85	2.25	3.15	3.86
4	3.4	2.4	3.25	3.43	2.37	3.1	3.1	2.4	3.3	3.4	2.4	55	3.3	3.12	3.45	3.32	2.5	2.36	41	2.1	2.1	1.9	1.8	32	0.25	1.91	1.9	2.1	2	3.64	3.54	2.16
5	3.5	1.9	4.43	3.55	1.95	4.2	3.25	1.95	4.4	3.5	2	55	4.5	4.17	3.6	3.43	2.33	1.95	40	2.3	2.2	1.7	1.7	31	0.25	2.35	2.3	2	1.67	4.68	3.5	1.92
6	6.5	13	1.27	6.48	13.2	1.25	5.5	13	1.3	6.5	15	56	1.3	1.25	6.8	6.11	15	12.55	39	1.6	1.5	2.7	2.5	34	-1.5	1.81	1.7	2.2	2.14	1.25	6.5	14.5
7	3.2	3.4	2.33	3.24	3.53	2.4	3.1	3.1	2.4	3.2	3.4	56	2.4	2.31	3.3	3.16	3.65	3.38	38	2.6	2.5	1.6	1.5	32	-0.25	1.99	1.9	2	1.92	2.2	3.38	3.7
8	3.6	5	1.88	3.68	4.64	1.83	3.4	4.5	1.8	3.6	5	56	1.88	1.82	3.8	3.56	5	4.62	42	2.1	2.1	1.8	1.8	33	-0.75	2.16	2.1	1.9	1.8	1.8	3.83	4.91
9	3.4	3.1	2.41	3.53	3.1	2.5	3.1	3	2.4	3.5	3.1	55	2.5	2.36	3.55	3.42	3.2	3.04	42	2	1.8	2.1	2	31	-0.5	2.41	2.3	1.8	1.64	2.8	3.44	2.68
10	3.6	1.8	4.7	3.62	1.88	4.5	3.4	1.85	4.8	3.6	1.9	55	5	4.5	3.75	3.51	1.95	1.86	42	2.1	2.1	1.9	1.8	33	0.75	1.8	1.8	2.2	2.11	5.4	3.65	1.78

Chart 1 the original dataset in CSV format which contains lots of unnecessary features

The dataset from the public libraries from 2013/2014 till 2018/2019 season of English premier league which is taken and the first five seasons are used as training data and just one current year is, due to time consuming, remained as testing data.

3.2. Feature engineering and selection

After lots of research and doing PCA in the data set the features which were in multiple dimensions *(chart 1*) changed into two dimension of Home/Away which are treated as global features. And another most important feature is the (Team Ratings) that comes from [https://www.fifaindex.com/] which shows all statistics of individual teams (*Attack*, *Midfield*,

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 3, ISSN No. 2455-2143, Pages 83-91
Published Online July 2020 in IJEAST (http://www.ijeast.com)

Defense, *Overall*) and over all team ratings generated by the football video game series FIFA *(chart 2).*

The Goal difference between two teams which is the feature shows the result of the played game come out from the total scored goals minus total of conceded goals for the kth match between teams as shown in equation 2.

$$GD_k = \sum_{j-1}^{k-1} GS_j - \sum_{j-1}^{k-1} GC_j \ldots (2)$$

There has been selected some metrics to assess the performance of the teams by counting the Corners, Overall rating of the teams will be concluded from the differentiation of home team rating (R_i^H) over the away team rating (R_i^A) as in equation 1.

$$R_i = R_i^H - R_i^A \ldots (1)$$

Shots on Target and Goals in the match. The Average value of past k matches of the
teams will be calculated as $\lambda_j^i \in$ {Corners, Shots on Target, Goals}, λ_p^i for a team's jth match as indicated in equation 3.

$$\lambda_j^i = \frac{\left(\sum_{p=j-k}^{j-1} \lambda_p^i\right)}{k} \ldots (3)$$

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	HST	AST	HC	AC
2	E0	13/08/16	Burnley	Swansea	0	1	A	0	0	D	3	9	7	4
3	E0	13/08/16	Crystal Palace	West Brom	0	1	A	0	0	D	4	3	3	6
4	E0	13/08/16	Everton	Tottenham	1	1	D	1	0	H	6	4	5	6
5	E0	13/08/16	Hull	Leicester	2	1	H	1	0	H	5	5	5	3
6	E0	13/08/16	Man City	Sunderland	2	1	H	1	0	H	4	3	9	6
7	E0	13/08/16	Middlesbrough	Stoke	1	1	D	1	0	H	2	1	9	6
8	E0	13/08/16	Southampton	Watford	1	1	D	0	1	A	6	1	6	2
9	E0	14/08/16	Arsenal	Liverpool	3	4	A	1	1	D	5	7	5	4
10	E0	14/08/16	Bournemouth	Man United	1	3	A	0	1	A	3	7	4	2

Chart 2 the most important features

After feature engineering the most relevant and useful features are selected which are; FTHG and HG = Full Time Home Team Goals, FTAG and AG = Full Time Away Team Goals, HTHG = Half Time Home Team Goals, HTAG = Half Time Away Team Goals, HST = Home Team Shots on Target, AST = Away Team Shots on Target, HC = Home Team Corners, AC = Away Team Corners, HTR = Half Time Result (H=Home Win, D=Draw, A=Away Win) and FTR and Res = Full Time Result (H=Home Win, D=Draw, A=Away Win).

3.3. Modeling

Support vector machine: as the literature review showed that the most successful algorithm for this prediction is to use from SVM, so in this research it is intended to test the prediction model on SVM. In the result section you may see that SVM has higher accuracy than other technique.

Random Forest: as it's a comparative predicting model, the model is also tested by random forest classifier. This classifier works well using less time and selecting features by itself but the accuracy is not as much as SVM.

3.4. Tools used to create the prediction model

Python version 3.4 is used for coding the model by using the Pycahrm and GitHub Desktop as environments. The libraries (Selenium and Urllib) are used for data extraction. The Sklearn is used for machine learning algorithms. The Numpy, Pandas and Matplotib are used for data analysis and other parts.

IV. RESULTS AND DISCUSSIONS

As it mentioned in the methodology part the model has been tested on SVM and random forest machine learning algorithms. The second objective of the paper was to model the ranking of the final table of teams in

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 3, ISSN No. 2455-2143, Pages 83-91
Published Online July 2020 in IJEAST (http://www.ijeast.com)

the league which is shown in the following figures by each algorithm.

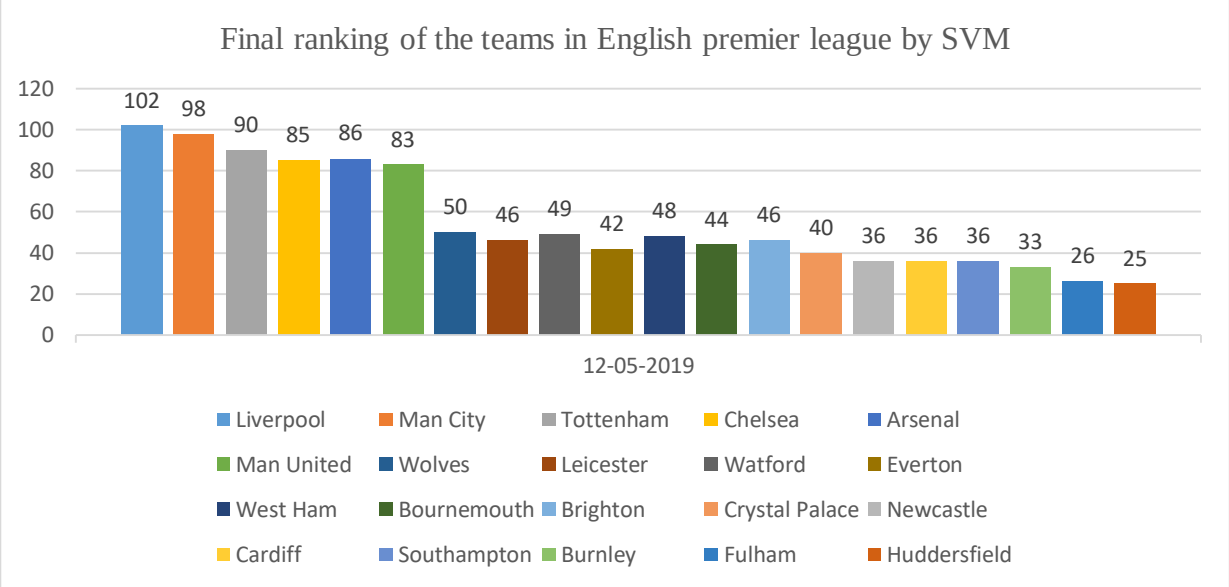

Figure 1 final ranking of the teams in EPL using the model in SVM

The above figure *(figure 1)* shows that the table of the EPL is quiet similar to the original table which is resulted in the season of 2018/2019 in English premier league, the first 6 teams are so similar in ranking numbers but just differ on the first and second team which should be swapped.

The figure *(figure 2)* shows that the table of English premier league is more similar to the original table of the season 2018/2019, the first six teams are the same on the ranking just the final scores of the ranking is a little different.

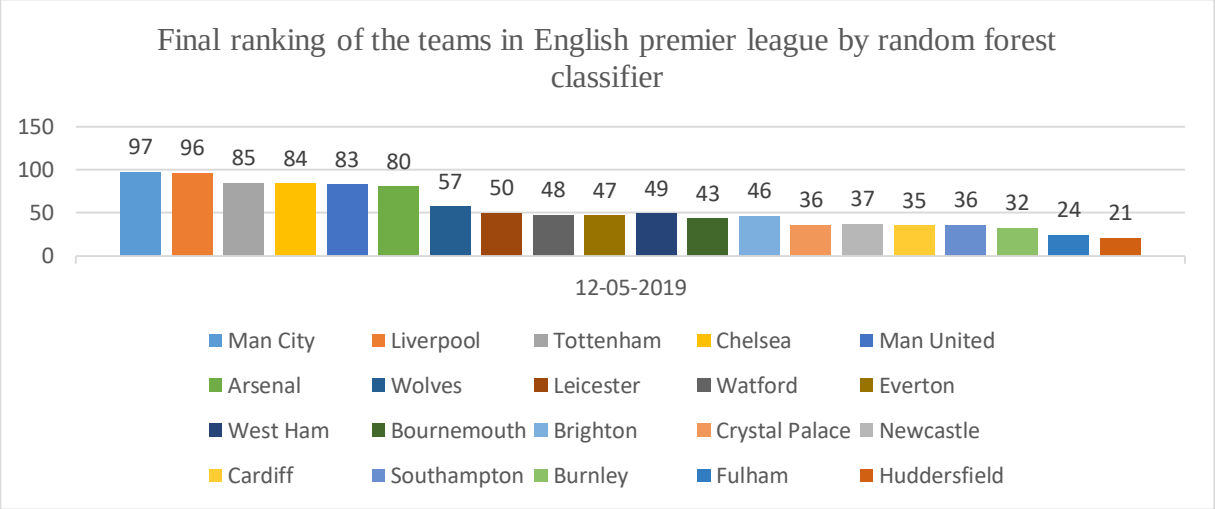

Figure 2 final ranking of the teams in EPL using the model in random forest classifier.

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 3, ISSN No. 2455-2143, Pages 83-91
Published Online July 2020 in IJEAST (http://www.ijeast.com)

#	Team	MP	W	D	L	F	A	D	P
1	Manchester City	38	32	2	4	95	23	+72	98
2	Liverpool	38	30	7	1	89	22	+67	97
3	▲ Chelsea	38	21	9	8	63	39	+24	72
4	▼ Tottenham Hotspur	38	23	2	13	67	39	+28	71
5	Arsenal	38	21	7	10	73	51	+22	70
6	Manchester United	38	19	9	10	65	54	+11	66
7	Wolverhampton ...	38	16	9	13	47	46	+1	57
8	▲ Everton	38	15	9	14	54	46	+8	54
9	▼ Leicester City	38	15	7	16	51	48	+3	52
10	▲ West Ham United	38	15	7	16	52	55	-3	52
11	▼ Watford	38	14	8	16	52	59	-7	50
12	Crystal Palace	38	14	7	17	51	53	-2	49
13	Newcastle United	38	12	9	17	42	48	-6	45
14	AFC Bournemouth	38	13	6	19	56	70	-14	45
15	Burnley	38	11	7	20	45	68	-23	40
16	Southampton	38	9	12	17	45	65	-20	39
17	Brighton & Hov...	38	9	9	20	35	60	-25	36
18	Cardiff City	38	10	4	24	34	69	-35	34
19	Fulham	38	7	5	26	34	81	-47	26
20	Huddersfield Town	38	3	7	28	22	76	-54	16

Figure 3 table of English premier league for the season 2018/2019

The figure *(figure 3)* shows the original table of the premier league for the season 2018/2019 which is taken from online public sports site www.soccerway.com.

This shows that the model in both techniques has their effectiveness so that random forest has more similar result than SVM by considering the ranking to the original table of the EPL.

The model confidence which is shown in figures *(figure 4)* and *(figure 5)* in the SVM (0.534) 53.4% is higher than random forest classifier (0.498) 49.8%.

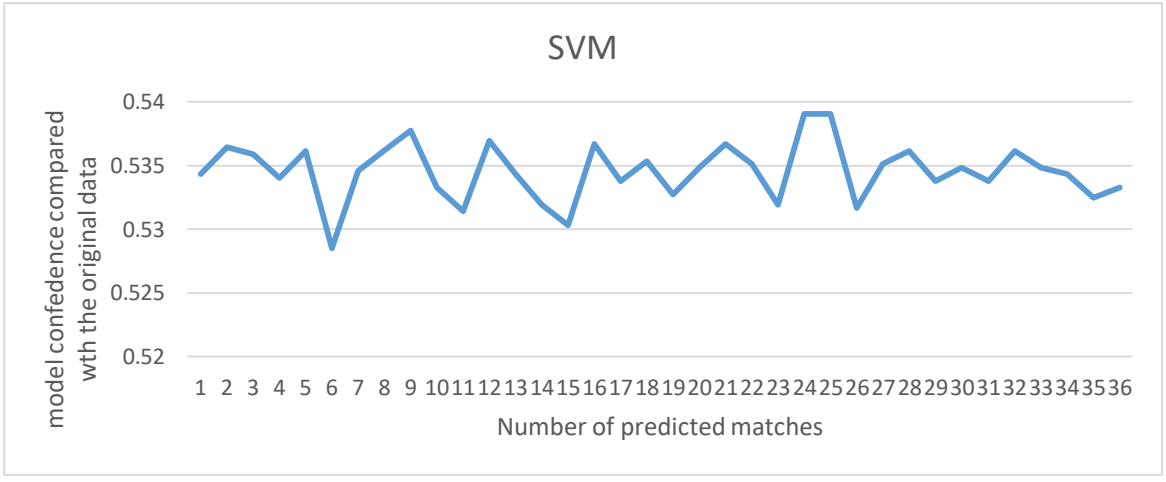

Figure 4 the confidence of the model in SVM

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 3, ISSN No. 2455-2143, Pages 83-91
Published Online July 2020 in IJEAST (http://www.ijeast.com)

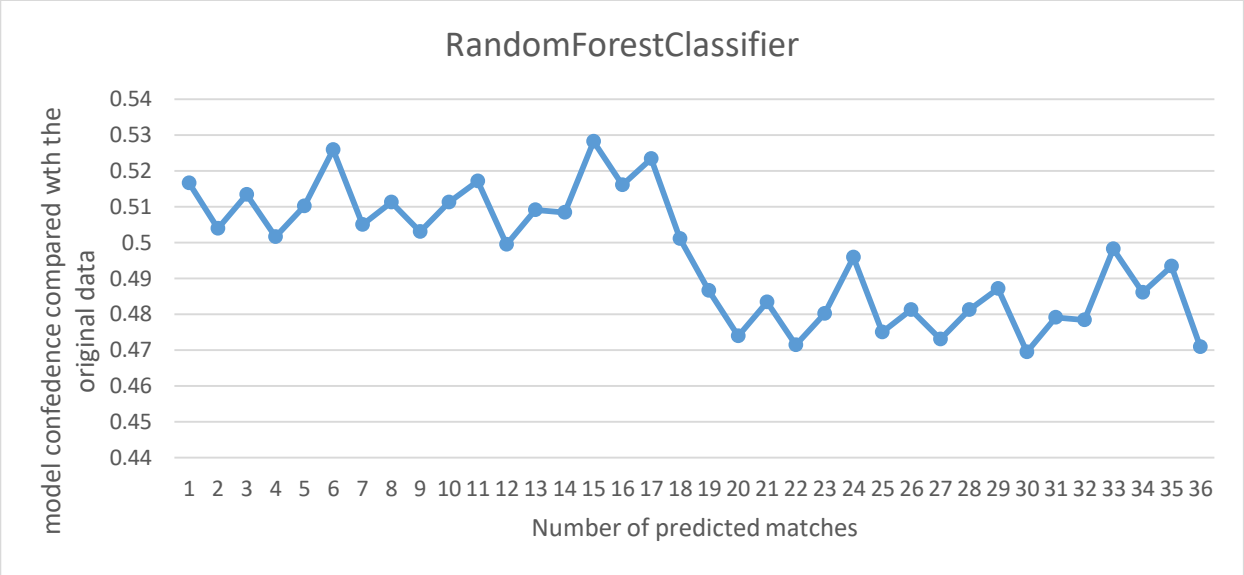

Figure 5 confidence of the model in random forest classifier

V. CONCLUSION

The result of this research which has the accuracy of 49.8% for random forest and 54.3% for SVM is comparable to what is found in the literature review, but in the literature review there hasn't done any work that shows the overall matches results for a special league season based on the other hand in this project the features which are related to the players are not included so the confidence and the accuracy is lower. Machine learning can be used to predict in different filed, including sports, especially football. An example of English premier league is indicated with this research that the possibility of finding a classifier that can predict the incoming matches by accuracy of up to 54.3% can be gained. And also there are lots of works to do and research more on the features in the future.

VI. FUTURE WORK

It would be better to add more data sets like Spanish league datasets and do a comparative work to get a more accurate and confident model. By adding some more features such as team streak and team form to provide better performance and get higher accuracy and Testing on more ML algorithms such as Naïve Bayes and gradient boosting models would be a better idea to create a more accurately predicting model.

Acknowledgments

The authors would like to thank Mr. Ali Mohammad dean of computer science faculty of Bakhtar University for his great supervision and effective guidance.

VII. REFERENCES

1. Angelini, G., and De Angelis, L. (2017). PARX model for football match predictions. Journal of Forecasting, 36(7), 795-807.
2. Baio, G., and Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. Journal of Applied Statistics, 37(2), 253-264.
3. Blundell, J. (2009). Numerical algorithms for predicting sports results (Doctoral dissertation, University of Leeds, School of Computer Studies).
4. Bunker, R. P., and Thabtah, F. (2017). A machine learning framework for sport result prediction. Applied Computing and Informatics.
5. Constantinaou, A. C., Fenton, N. E., and Neil, M. (2012). A Bayesian network model for forecasting Association Football match outcomes. Knowledge-Based Systems, 36, 322.
6. Dixon, M. J., and Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. Journal of the Royal Statistical Society: Series C (Applied Statistics), 46(2), 265-280.
7. Dixon, M. J., and Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. Journal of the Royal Statistical Society: Series C (Applied Statistics), 46(2), 265-280.
8. Elfrink, T. (2018). Predicting the outcomes of MLB games with a machine learning approach. Vrije universiteit Amsterdam.

9. Elo, A. E. (1978). The rating of chessplayers, past and present. Arco Pub.
10. Esumeh, E. O. (2015). Using machine learning to predict winners of football league for bookies. Int. J. Artif. Intell, 5, 22.
11. Goddard, J. (2005). Regression models for forecasting goals and match results in association football. International Journal of forecasting, 21(2), 331-340.
12. Goddard, J. (2005). Regression models for forecasting goals and match results in association football. International Journal of forecasting, 21(2), 331-340.
13. Hamadani, B. (2005). Predicting the outcome of NFL games using machine learning. URL http://cs229.stanford.edu/proj2006/BabakHamadani-PredictingNFLGames.pdf
14. Hill, I. D. (1974). Association football and statistical inference. Applied statistics, 203-208.
15. Huang, K. Y., and Chen, K. J. (2011). Multilayer perceptron for prediction of 2006 world cup football game. Advances in Artificial Neural Systems, 2011, 11.
16. Hvattum, L. M., and Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. International Journal of forecasting, 26(3), 460-470.
17. Joseph, A., Fenton, N. E., and Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. Knowledge-Based Systems, 19(7), 544-553.
18. Kampakis, S., and Thomas, W. (2015). Using machine learning to predict the outcome of english county twenty over cricket matches. arXiv preprint arXiv:1511.05837.
19. Karlis, D., and Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. Journal of the Royal Statistical Society: Series D (The Statistician), 52(3), 381-393.
20. Karlis, D., and Ntzoufras, J. (2003, April). Bayesian and non-Bayesian analysis of soccer data using bivariate Poisson regression models. In 16th Panhelenic Conference in Statistics. Kavala (Vol. 20).
21. Liu, F., Shi, Y., and Najjar, L. (2017). Application of Design of Experiment Method for Sports Results Prediction. Procedia computer science, 122, 720-726.
22. Machine learning, https://www.datarobot.com/wiki/machine-learning/ retrieved on April 24, 2020.
23. O'Donoghue, P. G., Dubitzky, W., Lopes, P., Berrar, D., Lagan, K., Hassan, D., and Darby, P. (2004). An evaluation of quantitative and qualitative methods of predicting the 2002 FIFA World Cup. Journal of Sports Sciences, 22(6), 513-514.
24. Reep, C., and Benjamin, B. (1968). Skill and chance in association football. Journal of the Royal Statistical Society. Series A (General), 131(4), 581-585.
25. Rue, H., and Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in a league. Journal of the Royal Statistical Society: Series D (The Statistician), 49(3), 399-418.
26. Spann, M., and Skiera, B. (2009). Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. Journal of Forecasting, 28(1), 55-72.
27. Steffen. S, (2107). Beating the Bookies: Predicting the Outcome of Soccer Games. Stanford University, USA
28. SteffenSmolka-BeatingTheBookies-poster – Beating the..., https://www.coursehero.com/file/23471228/SteffenSmolka-BeatingTheBookies-poster/ retrieved on June 26, 2020
29. Stenerud, S. G. (2015). A study on soccer prediction using goals and shots on target (Master's thesis, NTNU).
30. Stenerud, S. G. (2015). A study on soccer prediction using goals and shots on target (Master's thesis, NTNU).
31. Timmaraju, A. S., Palnitkar, A., and Khanna, V. (2013). Game ON! Predicting English Premier League Match Outcomes.
32. Yezus, A. (2014). Predicting outcome of soccer matches using machine learning. Saint-Petersburg University.
33. Zaveri, N., Tiwari, S., Shinde, P., Shah, U., and Teli, L. K. (2018). Prediction of Football Match Score and Decision Making Process. International Journal on Recent and Innovation Trends in Computing and Communication, 6(2), 162-165.