Predicting the Best Team Players of Pakistan Super League using Machine Learning Algorithms

Jawaria Ashraf
MS Scholar
Institute of Information and Communication Technology, Mehran UET, Jamshoro, Pakistan

Sania Bhatti, PhD
Associate Professor
Department of Software Engineering, Mehran UET, Jamshoro, Pakistan

Shahnawaz Talpur, PhD
Associate Professor
Institute of Information and Communication Technology, Mehran UET, Jamshoro, Pakistan

ABSTRACT
Owing to short and fast paced play, T20 is the adored format of cricket sport. In T20 cricket, Pakistan super league (PSL) is one of the most famous professional leagues founded to strengthen Pakistan cricket by scrutinizing the young talent. However, the selection of the best players for PSL teams is a very critical phase which certainly affects the final results of the play. To avoid biasness caused by the human nature in selection process, this study aims to select and rank the team of top fifteen players based on their batting and bowling performance in previous five seasons of PSL using Machine learning approach. For this purpose, Support vector machine (SVM), Random forest, Naive Bayes, Linear regression and K-nearest neighbor (classification) techniques have been employed for the development of predictive model from individual batting and bowling features sets. Based on comparison of applied techniques, the evaluated results have been plotted in term of accuracy, precision, recall and "f1score". For the selection of both batsman (in term of runs scored) and bowlers (in term of wickets taken), Random Forest performed well by yielding an accuracy of 100%. Findings of this research also ascertain that batting performance leads over bowling performance.

Keywords
Pakistan Super League, Batting, Bowling, Machine Learning, Prediction, Classification, Ranking

1. INTRODUCTION
Cricket is renowned, simple and entertainment sport around the globe especially in Asian countries. In the southeast England, cricket sport was initiated in the 16th century and became an international team play of the world. Cricket is played between two teams each consists of 11 players. Three main formats of cricket including one day international, test and T20s are played internationally [1]. T20 is the most popular format among fans due to its shorter and fast paced play. Therefore, this research focuses on T20s format due to its popularity at international as well as domestic level. In T20 cricket, both squads are mainly concerned with a maximum of 20 overs in each inning. There are a number of T20 domestic leagues played around the globe. Pakistan super league is the one of the most famous T20 cricket domestic leagues which is admired due to its competitiveness in term of batting and bowling. PSL was established in 2015 with 5 cricket teams including Peshawar Zalmi, Quetta Gladiators, Islamabad United, Lahore Qalandars, Karachi Kings and Multan Sultan was introduced in season 3 as 6th team. PSL usually starts in February, in which every team play 2 matches against each team. Top 4 best performing teams are qualified for semifinal. Out of 4 teams, only two teams qualify for final match and terminating the tournament by winning PSL Cup. Each PSL team selects highly qualified international players from 11 different countries of the world. Selection committee selects the players including wicket keeper, batsmen and bowlers. Each type of player has its specialized skills like batsman is enough competent to bat at various locations, angles and different types of bowlers including spin bowler, fast bowler and medium fast bowler. People of various era and surroundings are extreme buffs of cricket owing to well performance of cricket team within recent bygone. Though, several pregame misperceptions ascend regarding team combo as which batsman and bowler should be nominated or dropped and which batting player have to play in which position for forthcoming match because it's a sport of uncertainty to predict the final results. Several ordinary aspects, matches' instructions, players' talent, their coordination and practice arrangements are very significant to analyze the performance metrics of final outcome of the match which is very helpful for team managers, instructors and speculators too [2]. Machine learning approaches are used by emerging classification methods based on significant features which directly impact on the final outcome of the match as weather, position of the players, location, home team, toss decision to estimate match outcome [3, 4]. The model is established to calculate the proficiency of the final result of the match using method contains training the data based on historical matches [5]. Usually, efficiency of machine learning approach is estimated in performance metric as prediction accuracy [6]. From the previous record of past 5 PSL seasons, it has been observed that selection of the players highly affects the final results of the match and it is also difficult to take a decision whether batting capabilities lead over bowling capabilities or not. Hence, to improve the team winning chance of tournament, the team players must be composed and categorized. Moreover, winning prediction also rely on features including toss winning, home ground, pitch , batting and bowling classification in term of run rate, batting average, bowling average , wickets taken and so on. So, one of the foremost techniques applied in selection of the top players research is machine learning which plays an important role in evaluation parameters of the final results of the match. Subsequently final outcome of the matches is evaluated using machine learning classification methods with several input factors based on previous matches played.

In this research, various machine learning approaches are employed to determine the best models which can evaluate the final result of the match with high performance metrics. The research interrogation we try to response is:

- For cricket matches allied to PSL-T20, which machine learning technique procure best probabilistic model and which classification

approaches are accurate in term of evaluation parameters including "accuracy", "precision", "recall" and "f1-score"?

The riposte of research query is that various machine learning techniques are empirically observed from previous studies containing decision tree, random forest, SVM and naive Bayes [7, 8] for best model which can evaluate the final result of the match with high accuracy. In this research we used previous 5 years of PSL data which is collected from PSL-T20 tournaments. Further description on data and experimental outcomes are deliberated in section 3 and section 4 correspondingly. The aim of this research is to identify the features influencing the performance which have direct impact on the final outcome of the match in cricket field. This research fascinated on producing an intelligible and simplified model which is effective to analyze the match result based on batting and bowling performances using machine learning prediction techniques for upcoming match series. Main reason behind the prediction of the match result is to enhance team capabilities and to uplift the team winning chances of tournament. Significance of team winning is beneficial for television indentures, enthusiast stock market, membership and maintenance, financial and media sponsorships, enterprises and stadium arrangements. This research can help the cricket management and researchers interested in cricket data prediction.

To have clear understanding, this paper is distributed in five different sections. After introduction, literature review is presented in section 2. Section 3 describes the methodology of the paper. Section 4 discusses classification results. Finally, section 5 of the paper concluded the findings with future research directions.

2. LITERATURE REVIEW

In this section, previous research to be acquainted with comprehensive study in cricket field is completed. Investigations fanatical to this field are deliberated in detail here. Many researchers have done prediction of cricket match result with several essential features like toss decision, home ground, day /night, performance of players, bowling and batting using machine learning prediction techniques classification via regression. The research mechanism is interconnected to several cricket match problems which are enlightened below.

Jayalath, K.P. [1] described classification, regression tree and logistic regression algorithms to predict ODI cricket match impact. In this research, the authors have examined the significance of home field advantage for several clubs including Pakistan, South Africa, India and New Zealand with respect to challenger's field location. In the main output result amongst all the team, by using CART and Logistic regression the South Africa has maximum winning possibility about 72% in home team advantage. Kumashkapadia et al. [3] examined machine learning algorithms to predict the IPL-T20 match. Authors have used filter-based model to identify significant features of the dataset and approved four Machine learning techniques comprising K-Nearest Neighbor, Naïve Bayes, Model Trees and Random Forest to predict the match outcomes from these distinctive features. Experimental result showed that Naïve Bayes algorithm generated better results for home team advantage feature to improve predictive outcomes. Munir, F., et al. [4] reciprocated in-play and pre-play data to estimate optimal output. They evaluated T20 format of international matches and IPL cricket match data as training dataset. In deep study, they divided the data on the base of various features like batting first, location, and home team vs. opponent team etc. Decision tree algorithm gave the best results with the accuracy of 78% and 75% for first inning and second inning respectively. Kampakis et al. [5] applied gradient decision trees, naive Bayes and logistic regression etc. approaches on particular dataset to predict the match outcome. In this model naïve Bayes generate better result and gave high accuracy about 64% while gradient decision trees gave lowest accuracy in the comparative study. Ahmed, et al. [6] utilized machine learning models with features selection and splitting of data for pre-match prediction. Pakistan's cricket performance prediction was done by random forest method with 82% accuracy. Pre-competition scrutiny which they used was the major reason to avoid any intolerant view and unsafe reaction. AkhilNimmagadda et al. [9] introduced T20 cricket match prediction model while the match is in progress [4]. They used statistical techniques to get the best results for predictive system. Regression model and random forest classifier applied to predict outcome of both first and second innings which based on runs scored per over. They obtained winner of the match by using random forest technique. NeerajPathak and HardikWadhwa [10] worked on predicting the optimum result for ODI match based on many features including toss decision, innings, home ground, performance of team players etc. They have used new classification approaches like naïve bayes, SVM, and random forest [3, 8] to generate all possibilities of winning or losing of an ODI cricket match (2001-2015). Experimental results showed that the SVM approach performed better with 62% accuracy while other models gave 60% accurate result.MadanGopalJhanwar and VikramPudi[11] applied supervised learning approach to predict the outcome of ODI cricket. Firstly KNN, SVM, Random forests, Logistic regression and Decision trees techniques were applied [12] on 22 player's performance to calculate the final results of match. KNN performed better results than other classifier and also gave high accuracy about 71%.Somaskandhan, P. [13] determined the important set of features which are highly affected on cricket match end results. They introduced twenty three various attributes to describe the facts of inning level and calculated them. In this research, various machine learning techniques are used but SVM is best approach to obtain optimal result with high accuracy.Thenmozhi, D et al. [14] elucidated several classification techniques to predict that the home team can win the match or not when the match is in progress. In the end result, the optimal output is obtained through Random Forest model because this technique provided highest accuracy for all of the team.

From the previous research, it is evident that most of studies have been conducted on machine learning methods for the prediction of final results of the match using different features and it is also identified that all researchers have used different numeral of attributes and machine learning methods in their paper. They are also directing different formats of cricket in their research. Some scholars have deliberated only features in their work while some scholars have evaluated which machine learning model will be accurate to predict the cricket match results. From above literature the low performance accuracy is achieved in study [10] which is 60% and highest performance accuracy is achieved in study [6] which is 82%. In certain studies, insufficient parameters are used which may decrease the performance accuracy of evaluated output. Therefore, we have selected important features and topmost approach too which can improve the performance metrics as prediction accuracy with best results. However, a limited research has been directed on the effects of players' selection on the final

outcomes of the match. Therefore, this study focuses on the prediction of the final results based on the selection of the players using different machine learning approaches before match. In this work, we used random forest, SVM, naive Bayes, KNN and linear regression methods because in literature these classifiers are ideal for such dataset.

3. METHODOLOGY & DATA DESCRIPTION

3.1 Methodology

In this research, several machine learning approaches are evaluated to tackle the problem of predicting final result of PSL match series. Intellectual models are depicted to analyze the result of the match based on the influence of batting and bowling performances respectively to improve the winning chance of tournament. For this purpose, two predictive models representing the effect of batting capabilities and bowling capabilities of PSL players are formulated to select the top batsmen and bowlers correspondingly. To estimate the batting and bowling performance of top players of PSL using machine learning methods, the Scikit library in Anaconda as package and Jupyter notebook as IDE is used. Figure 1 signifies the methodology steps of this research to evaluate the final outcomes.

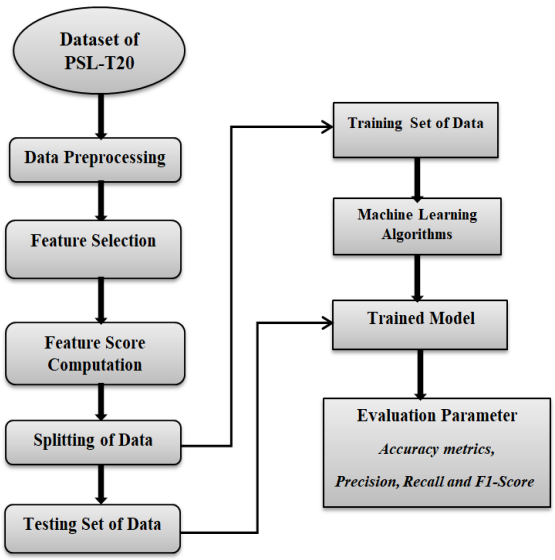

Fig 1: Proposed methodology

First, the input features of batting and bowling dataset are preprocessed by excluding insufficient parameters and choosing the key features which directly impact on training set's performance using feature selection method. Then extract the individual value of each feature using correlation matrix and compute the feature score. After feature computation, the dataset is split into various sizes of dualistic percentages (training-data and testing-data) to obtain accurate combination which provides high accuracy. Selection of top players is based on maximum runs made by batsmen and maximum wickets taken by bowlers respectively by means of different parameters. Then machine learning approaches are applied to develop classification models. These predictive models are finally compared with precision, accuracy, recall and "f1-score" to analyze the final results.

3.2 Data Collection

Previous five years of Pakistan super league (PSL-T20) data is apprehended to execute valuation scrutiny. So, this research uses the secondary statistics which is easily accessible on Cricinfo website. From [15] we extracted the required data for series of PSL season I-V (2016-2020). Data consists of several features as 16 factors and 526 instances for batsmen dataset and bowlers' dataset respectively. Then, we analyzed the data, selected the significant features and created a CSV file. As per PSL procedures only 5 separate teams take part in every single season of tournament. Though, one more team has been made in season 3 (2018) which is Multan Sultan. Owing to this, PSL dataset has 6 separate teams. Majority of PSL tournaments are played by Peshawar Zalmi, Quetta Gladiators, Islamabad United, Lahore Qalandars and Karachi Kings whereas Multan Sultan played only three PSL seasons. Figure 2 shows the dataset of batsmen and bowler's performance before preprocessing of the data.

Batsman	Team	Matche	Innings	Not outs	Runs	Highest score	Batting averag	Balls faced	Strike rate	100s	50s	0s	4s	6s	Season
0 Aamer Yamin	Peshawar Zalmi	2	-	-	-	-	-	-	-	-	-	-	-	-	2016
1 Abdur Rehman	Peshawar Zalmi	1	-	-	-	-	-	-	-	-	-	-	-	-	2016
2 Adnan Rasool	Lahore Qalandars	3	-	-	-	-	-	-	-	-	-	-	-	-	2016
3 Ahmed Shehzad	Quetta Gladiators	10	10	0	290	71	29	202	143.56	0	2	0	36	8	2016
4 Aizaz Cheema	Quetta Gladiators	4	1	1	0	0	-	0	-	0	0	0	0	0	2016
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
521 D Wiese	Lahore Qalandars	11	8	5	131	48	43.66	77	170.12	0	0	0	10	8	2020
522 Yasir Shah	Peshawar Zalmi	4	1	1	1	1	-	1	100	0	0	0	0	0	2020
523 Zafar Gohar	Islamabad United	3	1	1	13	13	-	10	130	0	0	0	1	0	2020
524 Zahid Mahmood	Quetta Gladiators	1	1	1	19	19	-	19	100	0	0	0	2	0	2020
525 Zeeshan Ashraf	Multan Sultans	11	11	0	202	52	18.36	155	130.32	0	2	0	25	6	2020

Bowlers	Team	Match	Inns	Overs	Mdns	Total Runs	Wickets	Ave	Econ	SR	4w	5w	Ct	St	Season
0 Aamer Yamin	Peshawar Zalmi	2	2	7	0	54	1	54	7.71	42	0	0	0	0	2016
1 Abdur Rehman	Peshawar Zalmi	1	1	3	0	28	0	-	9.33	-	0	0	1	0	2016
2 Adnan Rasool	Lahore Qalandars	3	2	6	0	67	0	-	11.16	-	0	0	0	0	2016
3 Ahmed Shehzad	Quetta Gladiators	10	-	-	-	-	-	-	-	-	-	-	4	0	2016
4 Aizaz Cheema	Quetta Gladiators	4	4	14	0	104	6	17.33	7.42	14	0	0	3	0	2016
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
521 D Wiese	Lahore Qalandars	11	10	30	0	252	12	21	8.4	15	0	0	6	0	2020
522 Yasir Shah	Peshawar Zalmi	4	4	12	0	112	3	37.33	9.33	24	0	0	2	0	2020
523 Zafar Gohar	Islamabad United	3	3	10	0	62	5	12.4	6.2	12	0	0	1	0	2020
524 Zahid Mahmood	Quetta Gladiators	1	1	2.5	0	26	0	-	9.17	-	0	0	0	0	2020
525 Zeeshan Ashraf	Multan Sultans	11	-	-	-	-	-	-	-	-	-	-	7	1	2020

Fig 2: Batsmen and Bowler's dataset

3.3 Feature Selection

Data is preprocessed using Pandas library in Jupyter notebook to select the effective parameters from batsman and bowlers dataset to predict the final outcomes of the match which minimizes overfitting, improves performance accuracy and fast to train the model. To get rid of missing records in dataset, the independent variables are preprocessed as input dataset by excluding inadequate proceedings. PSL statistics with no prediction of match outcome was eradicated from machine learning classification models. Finally 13 features are chosen for batting performance and 9 features for bowling performance that acknowledged as accurate set of attributes using wrapper and filter based feature selection method. Several attributes influence the cricket match [16]. Moreover significant features which affect the cricket match based on batting and bowling proficiencies are used to rank the top players which are listed below in Table 1 and Table 2 respectively.

Table 1. Batting attributes

Batting features	Description
Batsmen	Name of the Batsmen who played in PSL-T20 tournament.
Runs	Total runs recorded by batsman in PSL series. Maximum record shows the best performance of batsman.
Matches	Total matches played by batsman in PSL series.
Highest score	Highest score made by batsman in PSL series.
Batsman' strike rate	Total runs score recorded for every hundred balls which faced by batsman voguish PSL series. **Batting strike rate = (Batsman runs*100)/balls faced by batsman**
Batting average	Total runs scored by batsman per total intervals in which he baptized as out in PSL series. **Batting Average = (Batsman runs/ number of innings)**
Innings	Total innings in which Batsman essentially played in PSL series.
Balls faced	Sum of balls faced by batsman in PSL series containing no balls.
Not outs	Batsman who not outs in PSL series.
Centuries (100s)	Total innings in which batsman recorded hundred score in PSL series.
Fifties (50s)	Total innings in which batsman recorded fifty score in PSL series.
Sixes (6s)	Batsman made six runs in PSL series.
Fours (4s)	Boundaries in which batsman made four runs in PSL series.

Table 2. Bowling attributes

Bowling Features	Description
Bowlers	Name of the Bowlers who bowled in PSL series.
Wickets	Total wickets taken by bowlers in PSL series.
Match	Total matches bowled by bowlers in PSL series.
Bowling average (Ave)	Average of total runs scored per wickets in PSL series. **(ave = runs / wkts)**
Bowlers economy rate (Econ)	Average of total runs recorded per over bowled in PSL series.
Bowlers strike rate (SR)	Average of total balls which is bowled per wicket taken in PSL series . **(SR = balls/wkts)**
Catch outs (Ct)	Total catches taken by bowlers in PSL series.
Overs	Total overs have been bowled by bowlers in PSL series.

3.4 Relative Feature Score Computation

Estimate the individual value of each feature using correlation. With the help of correlation matrix this work identified how the factors are interrelated to target variable which is run score made by batsmen and wickets taken by bowlers from batting and bowling attributes respectively. The value of target variable may be maximize or minimize due to positive and negative correlation.

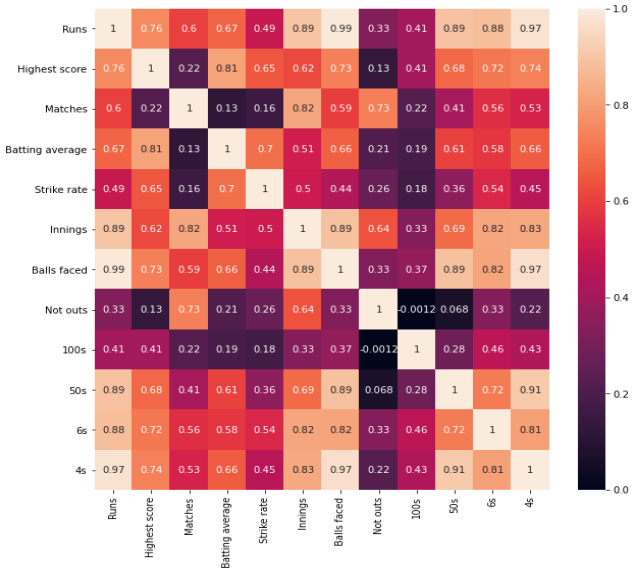

Fig 3: Correlation matrix of batting attributes

Figure 3 shows the correlation of batting features in which Runs score is target variable correlated to other factors as Balls faced is extremely correlated with Runs followed by 50s, innings, 6s, highest score, batting average and matches whereas not outs, strike rate and 100s are slightly correlated with Runs. Finally feature score of batting attributes are calculated using following formula:

Feature score for batting attributes (FS_1) = (Runs*0.97+Highestscore*0.74+Matches*0.53+Battingaverage*0.66+Strikerate*0.45+Innings*0.83+Ballsfaced*0.97 +Not outs*0.22+100s*0.43+50s*0.91+ 6s*0.81+4s*1)

Where **0.97**, 0.74, 0.53, 0.66, 0.45, 0.83, 0.97, 0.22, 0.43, 0.91, 0.81 and 1 are experimental variables used to rank the individual factors of top fifteen batsmen pronounced above in computation of feature score for batting attributes.

Feature score for bowling attributes (FS_2) = (Wickets*0.96+Match*0.89+Ave*(-0.038) +Econ*(-0.29) + SR*0.058+Ct*0.66+Overs*1)

Where 0.96, 0.89, -0.038, -0.29, 0.058, 0.66 and 1 are analytically observed variables to rank the top fifteen bowlers of PSL show in Figure 4.

Fig 4: Correlation matrix for bowling attributes

4. RESULTS ANALYSIS & DISCUSSION

4.1 Batting Analysis

In first predictive model, top batsmen are estimated considering batting competencies of PSL players. In this model, thirteen batting attributes are used as batsmen name, runs made by batsmen, batsmen innings, batting average, sixes, fours, centuries, total matches, highest score of batsmen, balls faced by batsmen, not outs, strike rate and fifties made by batsmen which have been exclusively normalized using correlation matrix. For implementation, model is trained and Scikit library is used for ML algorithms. Predictive model individually examines the proficiencies of batsmen and then fused all batting capabilities together. For batsmen analysis, group the batsmen data by batsmen names and merge all their batting aptitudes organized. After that sort the calculated values in decreasing order and ranking top to bottom best fifteen batsmen of PSL tournament based on feature score shows in Figure 5. Babar Azam and Kamran Akmal have best feature score of 3102.67 and 3033.88 respectively.

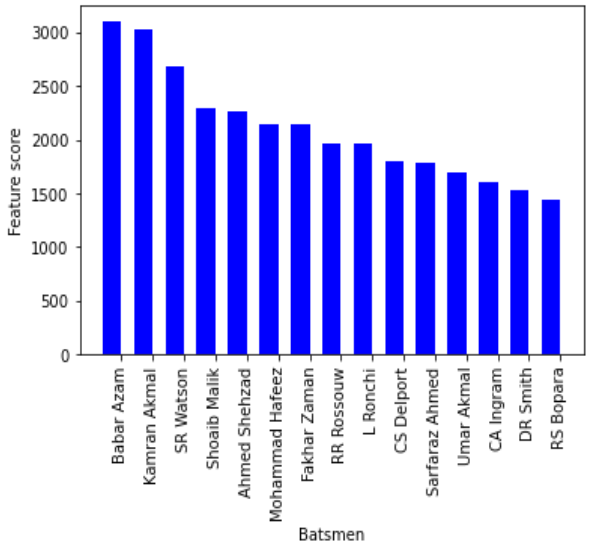

Fig 5: Top Batsmen of PSL based on Feature score

Figure 6 shows the ranking of top fifteen batsmen of PSL whereas Figure 7 indicates a whole list of top 119 PSL batsmen ranked who played at least eight matches in PSL season (2016-2020).

Batting-Index= (Batting-Average)*(Strike-rate)/100

Batsmen performance also depends on highest value of batting index shows in Figure 8.

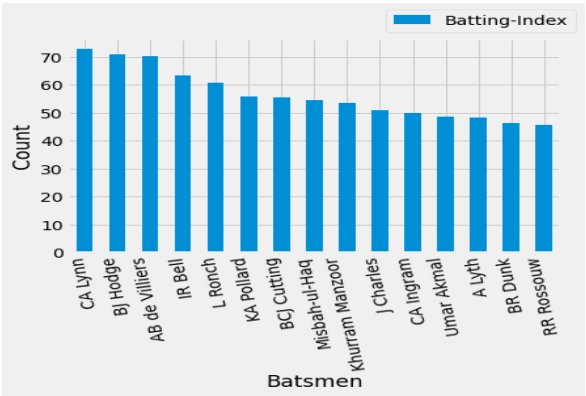

Fig 8: Batting index of PSL players

	Batsman	Runs	Highest score	Matches	Batting average	Strike rate	Innings	Balls faced	Not outs	100s	50s	6s	4s	Feature score	Ranking
0	Babar Azam	1516.0	78.0	47.0	33.920000	106.374000	45.0	1285.0	5.0	0.0	14.0	23.0	163.0	3102.675500	1.0
1	Kamran Akmal	1537.0	107.0	56.0	28.566000	137.040000	55.0	1111.0	2.0	3.0	9.0	77.0	158.0	3033.881560	2.0
2	SR Watson	1361.0	91.0	46.0	31.916000	138.294000	46.0	982.0	4.0	0.0	9.0	81.0	123.0	2683.586860	3.0
3	Shoaib Malik	1127.0	68.0	48.0	32.254000	122.090000	44.0	919.0	8.0	0.0	7.0	41.0	74.0	2288.468140	4.0
4	Ahmed Shehzad	1077.0	99.0	45.0	27.128000	116.652000	43.0	897.0	2.0	0.0	9.0	32.0	109.0	2261.527880	5.0
5	Mohammad Hafeez	1002.0	98.0	48.0	24.818000	119.752000	45.0	854.0	6.0	0.0	6.0	37.0	97.0	2139.648280	6.0
6	Fakhar Zaman	1064.0	94.0	40.0	26.350000	138.065000	40.0	775.0	0.0	0.0	7.0	42.0	111.0	2138.700250	7.0
7	RR Rossouw	962.0	100.0	40.0	34.792500	131.405000	37.0	734.0	9.0	1.0	3.0	34.0	85.0	1970.805300	8.0
8	L Ronchi	1020.0	94.0	31.0	36.833333	164.690000	31.0	614.0	3.0	0.0	10.0	50.0	116.0	1961.380500	9.0
9	CS Delport	860.0	117.0	34.0	24.654000	134.700000	33.0	666.0	2.0	1.0	5.0	33.0	79.0	1800.246640	10.0
10	Sarfaraz Ahmed	868.0	56.0	52.0	28.328000	124.944000	43.0	694.0	12.0	0.0	3.0	16.0	74.0	1787.081280	11.0
11	Umar Akmal	833.0	93.0	32.0	37.567500	129.657500	30.0	604.0	5.0	0.0	7.0	42.0	66.0	1695.200425	12.0
12	CA Ingram	793.0	127.0	32.0	32.843333	151.803333	31.0	519.0	7.0	1.0	3.0	40.0	67.0	1603.398100	13.0
13	DR Smith	701.0	73.0	28.0	31.937500	107.080000	25.0	611.0	4.0	0.0	5.0	31.0	67.0	1529.054750	14.0
14	RS Bopara	667.0	71.0	37.0	23.626000	100.038000	34.0	574.0	8.0	0.0	3.0	17.0	51.0	1434.010260	15.0

Fig 6: Top fifteen batsmen in PSL

	Runs	Highest score	Matches	Batting average	Strike rate \
count	119.000000	119.000000	119.000000	119.000000	119.000000
mean	322.941176	50.000000	23.201681	19.037301	108.836283
std	333.654248	30.629553	13.461546	11.124896	34.129366
min	1.000000	1.000000	8.000000	0.000000	0.000000
25%	76.000000	23.500000	12.000000	10.614584	95.235000
50%	221.000000	50.000000	19.000000	18.265000	115.817500
75%	448.500000	73.000000	32.000000	27.861666	130.127500
max	1537.000000	127.000000	56.000000	50.192500	179.740000
	Innings	**Balls faced**	**Not outs**	**100s**	**50s **
count	119.000000	119.000000	119.000000	119.000000	119.000000
mean	17.016807	252.655462	3.394958	0.067227	1.638655
std	12.115953	254.731711	3.380448	0.337767	2.516828
min	2.000000	2.000000	0.000000	0.000000	0.000000
25%	8.000000	58.500000	1.000000	0.000000	0.000000
50%	13.000000	177.000000	2.000000	0.000000	1.000000
75%	23.000000	356.500000	5.000000	0.000000	2.000000
max	55.000000	1285.000000	14.000000	3.000000	14.000000
	6s	**4s**	**Ranking**		
count	119.000000	119.000000	119.000000		
mean	13.226891	28.184874	60.000000		
std	15.382930	32.458309	34.496377		
min	0.000000	0.000000	1.000000		
25%	2.000000	5.000000	30.500000		
50%	8.000000	17.000000	60.000000		
75%	17.000000	40.000000	89.500000		
max	81.000000	163.000000	119.000000		

Fig 7: Descriptive analysis of top batsmen in PSL

In this analysis various sizes of training and testing sets are performed to idealize the particular percentage of given data. Five machine learning approaches are applied as SVM, Naive Bayes, Random Forest, KNN and Linear Regression to estimate batting performance of top batsmen in PSL series with best combination of training and testing model accuracy. The final results of dualistic percentage are given in Table 3 which shows the accuracies estimated by different approaches to predict the Runs score.

Table 3. Runs prediction with different sizes of training and testing set

Classifier	Accuracy (%)				
	85% train & 15% test	80% train & 20% test	70% train & 30% test	65% train & 35% test	60% train & 40% test
Random Forest	100	100	100	99.89	99.52
SVM	94.00	93.24	90.03	89.68	87.57
Naive Bayes	73.02	69.24	66.40	64.24	64.04
Linear Regression	80.56	86.32	80.56	82.90	81.98
KNN	76.23	84.45	86.00	77.8	89.45

It is apparent from Table 3 that random forest gives best prediction results for batting feature in all training and testing sets. Accuracies of Random forest, SVM and Naive Bayes increase as size of training set is increased except linear regression and KNN. In this case prediction accuracies increase when the size of training dataset is decreased. In prediction of Runs score the random forest gives an accuracy of 100% when 85%, 80% and 70% of data is used as training set whileat 65% and 60% of training data this model providesaccuracies of 99.89% and 99.52% respectively. For predicting the runs, SVM provides highest accuracy of 94% when training data is 85% used whereas it gives low accuracy of 87.57% when training model used as 60%. Naive Bayes estimate runs score with lowest accuracy of 64.04% when training data is used 60% and predicts the Runs with highest accuracy of 73.02% at 85% of training set. Linear regression evaluated the runs score with minimum accuracy of 80.56% when training set is used as 70% and 85% respectively and at 80% of training data this model estimated the maximum accuracy of 86.32% to predict the Runs score. KNN predicts the runs score with high accuracy of 89.45 when 60% training data is used and predicts the runs with low accuracy of 76.23% at 85% of training set. The top five batting players of PSL are acknowledged here as Babar Azam, Kamran Akmal, SR Watson, Shoaib Malik and Ahmed Shehzad who made maximum runs of 1516, 1537, 1361, 1127 and 1077 respectively. Figure 9 indicates the plot of batting performance of top batsmen prediction in term of accuracy metrics when training set is 85%.

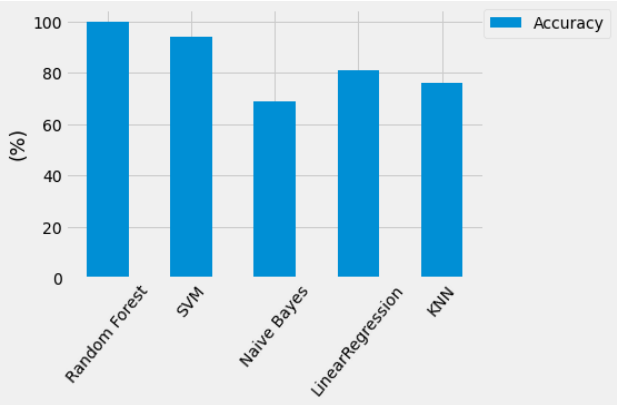

Fig 9: Accuracy metrics evaluated using machine learning algorithms for runs prediction of top batsmen in PSL using 85% of data as training

Estimating precision, recall and f1-score results for Runs (total number of runs made by top batsmen) are shown in Figure 10 compatible with accuracy when 85% of data is used as training model. Random Forest approach outstripped as compared to other techniques with the precision, recall and f1-score of 100% for Runs prediction owing to null character false positives (FPs) and false negatives (FNs). This model is capable to properly classify all samples rendering to confusion matrix while SVM, Naive Bayes, KNN and Linear regression incorrectly classified 7, 25, 17 and 32 instances respectively for predicting the Runs. These wrongly predicted samples enlarged the FPs and FNs and eventually gave lower results for precision, recall and f1-score. These performance metrics are acquired by SVM is 89%, 94% and 91% as precision, recall and f1 score respectively and Naive Bayes gave 84%, 86% and 90% respectively. Linear regression provides 83% precision, 87 % recall and 90% f1-score while KNN gives 86% precision, 90% recall and 92% f1-score for runs prediction.

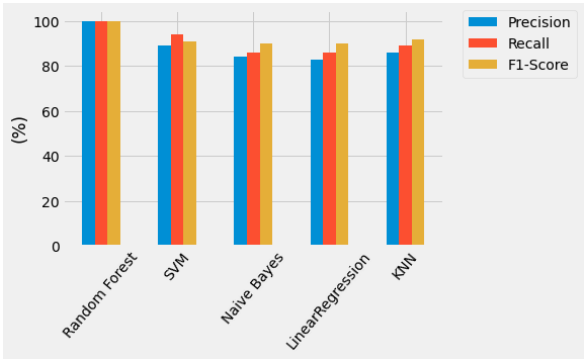

Fig 10: Precision, recall and f1_score evaluated using machine learning algorithms for top batsmen in PSL using 85% of data as training set

4.2 Bowling Analysis

In second predictive model, eight bowling features have been used as response to scrutinize the top fifteen bowlers in PSL using feature score formula defined in section 3. The factors used during analysis are: bowlers, wickets taken by bowler's name, matches played by bowlers, bowling average, bowler's innings, bowler's economy rate, bowling strike rate, total catches, and total overs bowled by bowlers in PSL series. Top fifteen batsmen based on feature score shows in Figure 11. WahabRiaz and Mohammad Irfan have top feature score of 333.66 and 322.3 respectively.

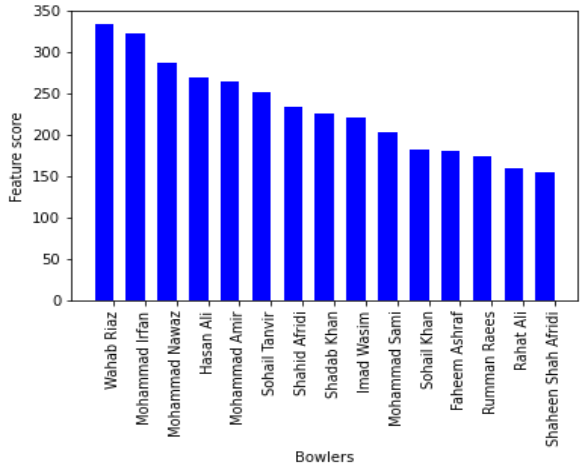

Fig 11: Top bowlers based on feature score

	Bowlers	Wickets	Match	Ave	Econ	SR	Ct	Feature score	Ranking
0	Wahab Riaz	76	55	19.078000	7.020000	16.100000	18	333.663036	1.0
1	Mohammad Irfan	55	64	32.040000	7.885000	24.912500	10	322.300755	2.0
2	Mohammad Nawaz	49	52	26.796000	7.184000	22.640000	30	286.731512	3.0
3	Hasan Ali	59	45	26.166000	7.464000	20.960000	15	269.346812	4.0
4	Mohammad Amir	49	48	25.712000	7.078000	21.780000	3	263.973564	5.0
5	Sohail Tanvir	48	46	34.792000	7.618000	27.440000	12	251.700204	6.0
6	Shahid Afridi	42	46	35.524000	6.822000	31.300000	13	234.727108	7.0
7	Shadab Khan	37	41	28.032500	7.192500	23.400000	15	225.516140	8.0
8	Mohammad Sami	42	36	24.470000	6.847500	20.450000	11	204.090465	9.0
9	Sohail Khan	37	34	40.564000	8.624000	26.640000	3	182.562728	10.0
10	Faheem Ashraf	46	31	21.323333	8.560000	14.533333	10	180.400247	11.0
11	Rumman Raees	34	32	26.150000	7.206000	21.440000	10	175.080080	12.0
12	Rahat Ali	34	28	24.203333	7.776667	18.433333	3	159.234173	13.0
13	Shaheen Shah Afridi	34	27	23.853333	7.970000	17.833333	3	154.366607	14.0
14	Usman Shinwari	35	31	24.116667	8.456667	17.433333	5	151.232267	15.0

Fig 12: Top fifteen bowlers in PSL

	Wickets	Match	Ave	Econ	SR \
count	98.000000	98.000000	98.000000	98.000000	98.000000
mean	16.265306	17.765306	29.964723	8.101371	22.269328
std	14.978026	14.297863	10.724092	1.103296	8.035061
min	1.000000	2.000000	10.600000	5.373333	9.600000
25%	5.000000	6.250000	22.637500	7.362500	16.693750
50%	10.500000	14.000000	27.595000	7.935000	20.980000
75%	23.000000	27.000000	35.023000	8.623000	25.525000
max	76.000000	62.000000	71.000000	11.500000	54.000000
	Ct	Overs	Ranking		
count	98.000000	98.000000	98.000000		
mean	4.908163	52.046939	49.500000		
std	5.197316	48.031242	28.434134		
min	0.000000	8.000000	1.000000		
25%	1.000000	16.000000	25.250000		
50%	3.000000	33.200000	49.500000		
75%	7.000000	70.500000	73.750000		
max	30.000000	208.000000	98.000000		

Fig 13: Descriptive analysis of top bowlers in PSL

Figure 12 indicates top fifteen bowlers based on their bowling capabilities in PSL tournament while Figure 13 shows complete ranking list of top 98 PSL bowlers who bowled in at least eight overs in PSL season (2016-2020). Both Figures also indicate the best bowling performance based on thoroughgoing taking outs by bowlers from batsmen, minimum number of average, strike rate and economy rate as well.

Bowling-Index = (Bowling-Average)*(Bowling-Strike rate)/100

A best bowling performance depends on lowest value of bowling index shows in Figure 14.

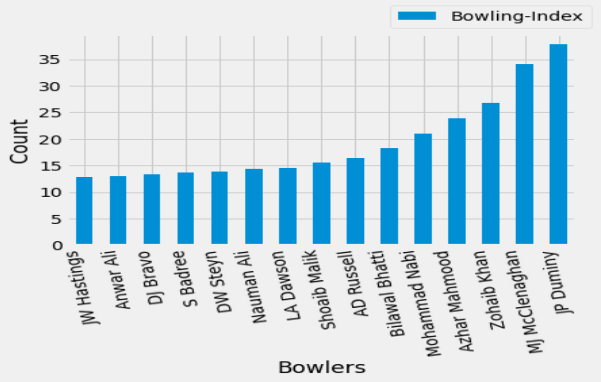

Fig 14: Bowling index of PSL players

In this examination several train's and test's sizes are used to predict the wickets taken by top bowlers. Table 4 shows the different sizes of train and test set to choose the best model accuracy applying five machine learning techniques to predict the wickets.

Table 4. Wickets Prediction with different sizes of training and testing sets

Classifier	Accuracy (%)				
	85% train & 15% test	80% train & 20% test	70% train & 30%test	65% train & 35% test	60% train & 40% test
Random Forest	100	100	100	98.54	99.85
SVM	89.00	73.44	70.29	68.46	66.23
Naive Bayes	47.74	49.15	47.30	53.55	58.50
Linear Regression	67.09	48.98	47.00	47.97	39.68
KNN	41.56	45.78	45.31	48.99	44.34

From Table 4 training and testing models are analyzed that random forest is best approach to predict the wickets from bowling attributes with high accuracy of 100% when 85%, 80% and 70% of data applied as train the model. Random forest gives lowest accuracy of 98.54% at 65% of training model. Accuracies of random forest, SVM and linear regression maximize as size of training model is maximized except in case of KNN and Naive Bayes. Accuracies of Naive Bayes and KNN increase as size of training model is decreased. SVM and linear regression provide the high accuracies of 89% and 67.09% respectively when 85% used as training set for the prediction of wickets and these both model provide low accuracies of 66.23% and 39.68% respectively when 60% training model is used. Naive Bayes gives lowest accuracy of 47.3% at 70% train set. Naive Bayes provides high accuracy of 58.50% at 60% of training set while KNN gives low accuracy of 41.56% when 85% of data is used as train model. KNN predicts the wickets with high accuracy of 48.99% when training model is used as 65%. Figure 15 indicates the plot of bowling features of top bowler's estimation in term of accuracy metrics when 85% of data used as train model.

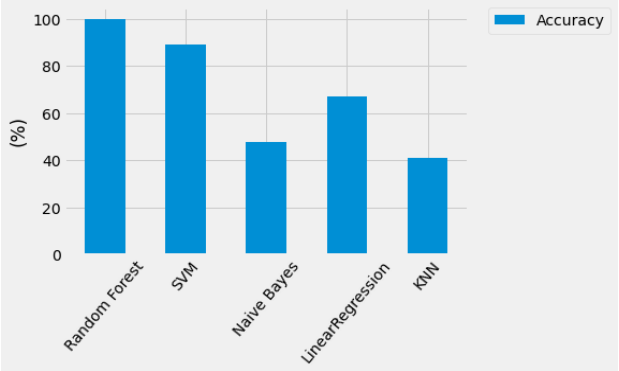

Fig 15:Accuracy metrics estimated using machine learning algorithms for wickets prediction of top bowlers in PSL

Precision, recall and f1 score results are evaluated for scrutinizing the "Wickets" prediction against the top batsmen feature when 85% of data is used as train model which shown in Figure 16 attuned with accuracy. In this, Random Forest approach outshined like first predictive model by means of all other techniques with the precision, recall and f1 score metrics of 100% as a result of null character false positives (FPs) and false negatives (FNs). This model proficient to correctly classify all samples interpreting to confusion matrix while SVM, Naive Bayes, Linear regression and KNN misclassified 32, 42, 9 and 17 samples respectively. These wrongly predicted instances achieve higher FPs and FNs values which ultimately decreasing the precision, recall and f1 score performance metrics. SVM approach gave precision, recall and f1 score results about 56%, 82% and 67% respectively and Naive Bayes provides the results of 36%, 73% and 48% respectively due to higher FPs and FNs samples for prediction of wickets at 85% train set. Linear regression predicts the wickets with 67% of precision, 96% of recall and 80% of f1-score whereas KNN provides precision, recall and f1-score of 44%, 94% and 64% respectively. It means, naive Bayes approach was incapable to increase the predictive results via all other techniques exposed an undesirable outcomes regarding to precision, recall and f1 score. Top five bowling players of PSL are predicted as WahabRiaz, Hassan Ali, Mohammad Irfan, Mohammad Amir and Mohammad Nawaz who has taken 76, 59, 55, 49 and 48 wickets respectively from batsmen.

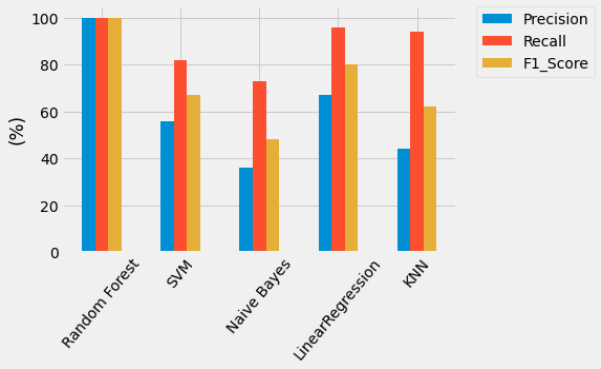

Fig 16: Precision, recall and f1_score evaluated using machine learning algorithms for top bowlers in PSL

5. CONCLUSION & FUTURE WORK

This study takes an imperious step to select the best players for PSL matches assuming a huge part in team's triumph and used previous five years of PSL data to predict the match outcome. Machine learning techniques are applied to estimate the final predictions by ranking the top players of PSL by dint of batting and bowling capabilities regarding accuracy, precision, recall and f1-score. In this research, set of 119 batsmen and 98 bowlers has been examined who played and bowled in at least eight matches and eight overs of PSL respectively. Four distinct machine learning classification techniques for scrutinizing the final outcomes have been executed and compared. First predictive model gave marginally high performance than second predictive model with evaluation parameters. Random forest performed better for both models as compared to other algorithms with an accuracy of 100% for analyzing runs made by batsmen and wickets taken by bowlers respectively. SVM gave equally high results concerning accuracy for predicting runs by batsmen and wickets taken by bowlers which are 94% and 89% respectively where as Naive Bayes gave low accuracy of 69% for runs prediction and 48% for wickets prediction. Linear regression predicted the runs with accuracy of 80.56% and predicted the wickets with accuracy of 67.09% while KNN predicted the runs and wickets with accuracy of 76.23 and 41.56% respectively. Random Forest classifier outperformed with the precision, recall and f1-score of 100% by means of zero result value of FPs and FNs for prediction of both runs made by batsmen and wickets taken by bowlers. It is concluded that random forest and SVM gave approximately equal high result for batting feature in accuracy metrics whereas Naive Bayes, Linear Regression and KNN gave low results which are not much satisfactory.Ranking of top fifteen batsmen and bowlers are depicted in Figure 6 and Figure 12 respectively. An exact prediction of top players' selection in match series will help the team managers and researchers involved in Cricket field to select the best players for team winning chances. These comparable models can also be worked for other formats of cricket like Test matches and ODI tournaments. In future, deep learning classification approach will be used in order to apprehend more valuable factors that can conceivably increase the prediction accuracy.

6. REFERENCES

[1] Jayalath, K. P. 2018. A machine learning approach to analyze ODI cricket predictors. Journal of Sports Analytics, 4(1), 73-84.

[2] Tekade, P., Markad, K., Amage, A., &Natekar, B. 2020. Cricket match outcome prediction using machine learning. International journal, 5(7).

[3] Kapadia, K., Abdel-Jaber, H., Thabtah, F., &Hadi, W. 2020. Sport analytics for cricket game results using machine learning: An experimental study. Applied Computing and Informatics.

[4] Munir, F., Hasan, M., & Ahmed, S. 2015. Predicting a T20 cricket match result while the match is in progress (Doctoral dissertation, Brac University).

[5] Kampakis, S., & Thomas, W. 2015. Using machine learning to predict the outcome of english county twenty over cricket matches. arXiv preprint arXiv:1511.05837.

[6] Ahmed, W., Amjad, M., Junejo, K., Mahmood, T., & Khan, A. 2020. Is the performance of a cricket team really unpredictable? a case study on pakistan team using machine learning. Indian Journal of Science and Technology, 13(34), 3586-3599.

[7] Shetty, M., Rane, S., Pandita, C., &Salvi, S. 2020. Machine learning-based Selection of Optimal sports Team based on the Players Performance. In 2020 5th International Conference on Communication and Electronics Systems (ICCES) (pp. 1267-1272). IEEE.

[8] Singh, S., Aggarwal, Y., &Kundu, K. 2020. Quantitative Analysis of Forthcoming ICC Men's T20 World Cup 2020 Winner Prediction using Machine Learning. International Journal of Computer Applications, 975, 8887.

[9] Nimmagadda, A., Kalyan, N. V., Venkatesh, M.,Teja, N. N. S., &Raju, C. G. 2018. Cricket score and winning prediction using data mining. International Journal for Advance Research and Development, 3(3), 299-302.

[10] Pathak, N., &Wadhwa, H. 2016. Applications of modern classification techniques to predict the outcome of ODI cricket. Procedia Computer Science, 87, 55-60.

[11] Jhanwar, M. G., &Pudi, V. 2016. Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach. In MLSA@ PKDD/ECML.

[12] Kumar, J., Kumar, R., & Kumar, P. 2018. Outcome prediction of ODI cricket matches using decision trees and MLP networks. In 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC) (pp. 343-347). IEEE.

[13] Somaskandhan, P., Wijesinghe, G., Wijegunawardana, L. B., Bandaranayake, A., &Deegalla, S. 2017, December. Identifying the optimal set of attributes that impose high impact on the end results of a cricket match using machine learning. In 2017 IEEE International Conference on Industrial and Information Systems (ICIIS) (pp. 1-6). IEEE.

[14] Thenmozhi, D., Mirunalini, P., Jaisakthi, S. M., Vasudevan, S., Kannan, V. V., &Sadiq, S. 2019. MoneyBall-Data Mining on Cricket Dataset. In 2019 International Conference on Computational Intelligence in Data Science (ICCIDS) (pp. 1-5). IEEE.

[15] [Online]. Available: https://www.espncricinfo.com.

[16] İshrat, R. İ. A. Z., Mushtaq, N., Jillani, M. M., &Navaz, U. 2019. Performance analysis of Pakistan super league players using principle component analysis approach. Scientific Journal of Mehmet AkifErsoy University, 2(4), 127-135.