Prediction of IPL Match Score and Winner Using Machine Learning Algorithms

Mrs. G V Gayathri
Assistant Professor, Department of CSE, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam-531162, India.

Deepika Gonnabattula
Student, Department of CSE, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam,531162, India.
deepika.17.cse@anits.edu.in

Srinivasa Reddy Gajjala
Student, Department of CSE, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam,531162, India.
srinivasareddy.17.cse@anits.edu.in

Manideep Kanakam
Student, Department of CSE, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam,531162, India.
manideep.17.cse@anits.edu.in

Sai Swaroop Medisetty
Student, Department of CSE, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam,531162, India.
mswaroop.17.cse@anits.edu.in

Abstract

Cricket is a very familiar and exciting sport that people of all age groups are insane to see and play. For many it's a billion-dollar market as they speculate financially, hoping to be able to earn profit in the form of gambling and various other ways. In this project, a model using machine learning algorithms is proposed to predict the score of each match and winning team based on past datasets available from 2008 to 2019 IPL matches in Kaggle. This proposed methodology includes the following steps like Pre-processing of collected datasets, Feature selection from raw data, Conversion of categorical data into numerical data, Partitioning of samples into training and test samples, Training, and classification. Few machine learning algorithms like Support Vector Machine, Random Forest, Naive Bayes were already used in previous papers. In this project, algorithms like Lasso Regression, Ridge Regression, and Random Forest regression models are proposed for a score prediction, and SVM(Linear, RBF), Logistic Regression classifier is for the match-winning prediction. The accuracy of the above machine learning algorithms is used to predict the winner of an IPL match along with its Precision, Recall and F-Measure measured and the model with better accuracy is considered.

Keywords: Cricket, IPL, Machine Learning, Regression, Classification.

1. Introduction

Cricket is the most widespread and much-loved game of everyone. it's delighted in by the overall population of all age mass because it is an exceptionally fascinating and suspicious game. Cricket is also referred to as the Game of Uncertainty and there is no precise forecast that a selected team would win in any given conditions. Finally, a team wins which multiplies the energy of every team member. There turn into a major jam of cricket darlings within the stadium and television rooms to see the cricket at whatever point i.e., a world level, national level, or any test match. The magnetism of cricket has also included businessmen that became a source of income for them as they gamble over their favorite teams. The popularity of cricket increased when ICC (International Cricket Council) started the concept of fast cricket within the sort of twenty-20(T-20) matches. In 2007, the first twenty-20 world cup was held within South Africa that was won by India which increased the popularity of this game in India. BCCI (The Board Of Control For Cricket In India) cashed the chance and created a league referred to as the Indian Premier League (IPL) in 2008 and got it approved by ICC. IPL is one of the finest twenty-20 cricket competitions in the present cricketing world that is based on the EPL (English Premier League) league and NBA (National Basketball Association) Basketball League [5]. During its first edition, IPL gained huge popularity which opened avenues for many stakeholders. In every IPL season 8 teams play with one another within the first stage, after the first stage 4 teams attend the eliminator round (next stage) and after the eliminator round 2 teams attend the final match and at last, there will be one winner. Each team is owned by a franchise that is owned by a group of people. These franchises hire players, evaluate them on the idea of their national, international, T-20 experience and performance, and hire them at the time of auctions [1]. Results of each match within the IPL depend on the varied conditions like venue, player performance, toss, performance in power-play, etc. Results of a match can only be predicted to some extent if previous player performance, venue, and other match-related data are available. In this paper, the authors predict the results of IPL match using three machine learning algorithms namely SVM(Linear, RBF), Logistic Regression classifier on the idea of previous data available [3][6].

The rest of the paper is organized as follows: Previous related work associated with the prediction of matches has been discussed in Section-II. The proposed prediction model is presented in Section-III, Section IV deals with results and subsequent discussions. Conclusion and future work are given in Section V

2. Literature Survey

According to the findings of the literature review, there is a need for a machine learning model that can predict the outcome of an IPL match before it starts.

The Twenty20 format of cricket, more than any other, sees a lot of changes in the game's momentum.

A game can be radically changed by an over.

As a result, forecasting the outcome of a Twenty20 game is a difficult undertaking.

Developing a prediction model for a league that is entirely based on auction is also a challenge.

IPL matches cannot be anticipated just based on statistics derived from historical data.

Players are bound to change teams as a result of player auctions, which is why the continued performance of each player must be taken into account while constructing a prediction model.

➢ Sasank used the batsman and bowler ratings, as well as the team's relative strength, to dynamically forecast the outcome of an IPL match's second innings. Chellapilla is a type of cactus. Deep Prakash, C. Patwardhan, and C. Vasantha propose a variety of methods for predicting the IPL season 9 champion.

➢ Dr. KB, Priyanka S, Vysali K Priyalyer used data from previous IPL editions and data mining algorithms to forecast the outcome of the 2020 edition of the IPL.

>Ananda Bandulasiri investigated the advantage a team has when playing on their home turf, discovering a link between the winner of the coin toss and the match's outcome, as well as analyzing and demonstrating the usefulness of the Duckworth Lewis Method.

>Shilpi Agrawal, Suraj Pal Singh, and Jayesh Kumar Sharma created a model that used three separate machine learning algorithms to predict the winner of a match and obtained excellent accuracy with all three techniques.

>The article also took into account the batting strike rate and the bowling run rate, which were both taken into account independently even during the Power Play overs.

>Amal Chaminda Kaluarachchi and Aparna S. Varde discovered that classification is the best method for predicting a match's winner.

>The authors investigated the factors that influence the outcome of a match and developed a program that could predict a team's odds of winning a match based on a set of characteristics.

3. Methods

The system architecture model is a five-step machine learning approach that can be explained. They are as follows:

- Obtain the dataset
- Information Feature extraction and pre-processing
- Data format conversion
- Model Training
-

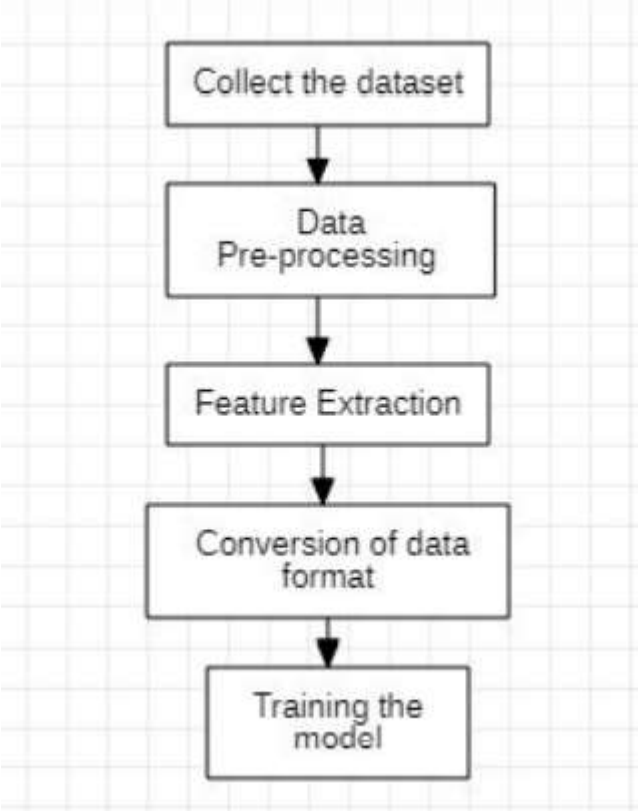

Figure 1: System Architecture

3.1 Obtain the dataset:

.The dataset for analysis and prediction was obtained from www.kaggle.com, which included data from past IPL editions from 2008 to 2019.

There were two datasets used in this study. On each ball of the match, the first gives us ball-by-ball information from every match ever played in the IPL, including the batsman, bowler, runs, wicket, and more. The second dataset contains a summary of each match, including the teams involved, the winner, the toss winner, and other information for every match played in the IPL.

3.2 Information Feature extraction and pre-processing:

The pre-processing stage cleans the dataset by deleting data that isn't necessary for obtaining results. During the pre-processing stage, data that has not been declared or tagged is eliminated.
To extract the essential analysis, as well as for the prediction module, the data must be pre-processed and cleaned. The dataset was created using records from the last 11 years, or from season 2008 to 2019.
Methods such as eliminating outliers, normalizing, and standardization are used to pre-process the data.

3.3 *Data format conversion:*

Because a few of the dataset's attributes are categorical, classification is rather difficult.
It may potentially have an impact on the model, resulting in incorrect predictions.
Except for the target attribute (Winner), all categorical data in the dataset has been transformed to numeric format and standardized on a scale basis in this step.
Ordinal encoding and one-hot encoding are the two most used approaches.

- *Model Training:*

The datasets were divided into two sections for training and testing before the model was trained. On one of the datasets, three regression models are used to predict the score: Lasso Regression, Ridge Regression, and Random Forest Regression. On the one hand, three predictive modeling classifiers, Support Vector Machine (SVM), Logistic Regressions are used for classification.

3.4.1 Random Forest:

Random forest is an ensemble-based supervised learning methodology. Ensemble learning is a type of learning in which many decision trees are constructed and then combined to produce more accurate prediction models. The resulting of trees termed "Random Forest" is a mix of multiple decision trees. The random forest algorithm is not biassed because it is based on a majority vote and delivers the final prediction based on that voting. Random Forest and Decision Tree both employ the identical Equation(1) and Equation(2) formulas.

The Random Forest method is based on the following principle:

Phase 1: In this step, random rows from the training data set will be selected by assigning an arbitrary value.

Phase 2: The decision tree is built based on the selection of random rows in this step. The output is then created from each decision tree.

Phase 3: Using the frequency method, voting will be done on the created output.

Phase 4: Based on the number of votes collected from the decision trees, the ultimate result is anticipated from step 3.

3.4.2 Ridge Regression:

When the number of predictor variables in a set exceeds the number of observations, or when a data set exhibits multicollinearity, ridge regression is an approach to develop a parsimonious model (correlations between predictor variables).
Ridge regression employs a ridge estimator, which is a sort of shrinkage estimator.
Theoretically, shrinkage estimators generate new estimators that are closer to the "actual" population parameters.
The ridge regression cost function is as follows:

$$\min\left(||Y - X(\theta)||_2^2 + \lambda||\theta||_2^2\right)$$

3.4.3 Lasso Regression:

Lasso regression is a sort of linear regression that makes use of shrinkage. Data values are shrunk towards a central point, such as the mean, in shrinkage.
The least absolute shrinkage and Selection Operator is an acronym that stands for Least AbsoluteShrinkage and Selection Operator.
Quadratic programming challenges, such as Lasso solutions, are best tackled with software (like Matlab).
The intensity of the L1 penalty is controlled by a tuning parameter. is the amount of shrinkage in terms of:

D = least-squares + lambda * summation (absolute values of the magnitude of the coefficients)

3.4.4 Support Vector Machine:

Support vectors are data points that are closer to the hyperplane and have an impact on the hyperplane's location and orientation. We maximize the classifier's margin by using these support vectors.
The goal of the SVM method is to maximize the distance between the data points and the hyperplane. Hinge loss is a loss function that aids in margin maximization.

3.4.5 *Logistic Regression:*
Under the Supervised Learning approach, Logistic Regression is one of the most used Machine Learning algorithms. It's a method for predicting a categorical dependent variable from a set of independent variables.
The steps below will be used to implement the Logistic Regression:
Fitting Logistic Regression to the Training Set Predicting the Test Result (Creation of Confusion Matrix) Visualizing the Test Set Result

4. Performance Analysis and Results

This part illustrates the obtained outputs of both Regression and Classification algorithms. The dataset used for win prediction has 757 records of IPL matches. Similarly, the runs prediction dataset has a record of 617 matches. For both predictions, the dataset is divided with 90% as training data and 10% as testing data.

Firstly, the algorithms used for score prediction are Random Forest, Ridge Regression, Lasso Regression, and the results are achieved based on the Cross Validation method. Secondly, the algorithms used for predicting match winners are SVM Linear, SVM RBF, Logistic Regression, and the results are evaluated based on accuracy, precision, recall, F-measure.

Table-1: Comparing Accuracies of various Regression Algorithms

Method	Random Forest	Ridge Regression	Lasso Regression
Accuracy	.75	.69	.67

Table-1 compares how different regression algorithms (Ridge Regression, Lasso Regression, Random Forest) got various Accuracies.

Figure-2: The bar graph gives information about how three different regressors performed on a dataset and got accuracies accordingly.

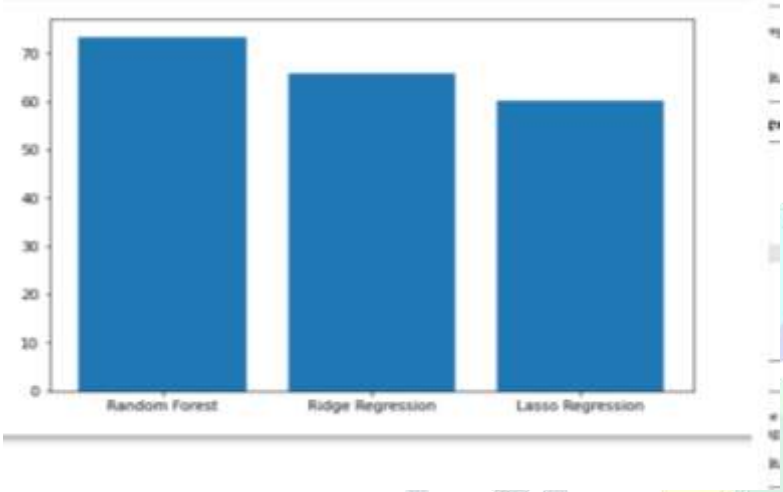

Figure-2

Table-2: Calculation of Precision, Recall, F- measure for SVM(Linear)

Accuracy	Precision	Recall	F1 Score
0.65	0.65	0.95	0.77

Table-2 shows various performance measures(Accuracy, Precision, Recall, F1Score) of the SVM Linear Classifier where the data set is divided with 90% training data and 10% testing data.

Figure-3: The below diagram illustrates how many number of true positives, true negatives, false positives, and false negatives are observed after applying SVM Linear classifier.

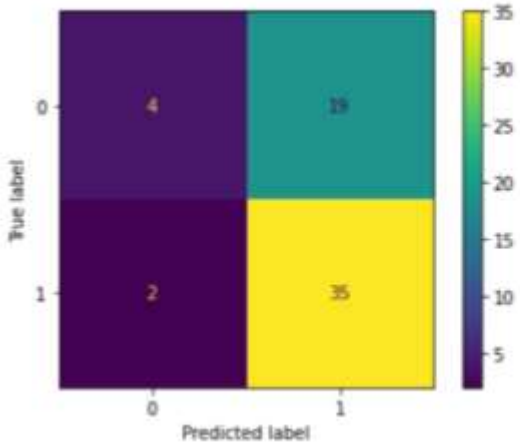

Figure-3

Table-3: Calculation of Precision, Recall, F- measure for SVM(RBF)

Accuracy	Precision	Recall	F1 Score
0.83	0.83	0.92	0.83

Table-3 shows various performance measures(Accuracy, Precision, Recall, F1Score) of the SVM RBF Classifier where the data set is divided with 90% training data and 10% testing data.

Figure-4: The below diagram illustrates how many number of true positives, true negatives, false positives, and false negatives are observed after applying the SVM RBF classifier.

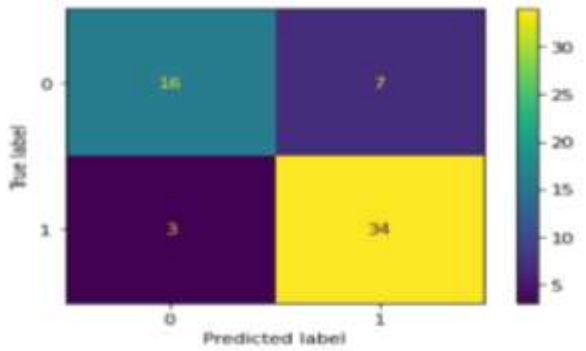

Figure-4

Table-4: Calculation of Precision, Recall, F- measure for Logistic Regression

Accuracy	Precision	Recall	F1 Score
0.63	0.64	0.76	0.68

Table-4 shows various performance measures(Accuracy, Precision, Recall, F1 Score) of Logistic Regression where the data set is divided with 90% training data and 10% testing data.

Figure-5: The below diagram illustrates how many number of true positives, true negatives, false positives, and false negatives are observed after applying Logistic Regression.

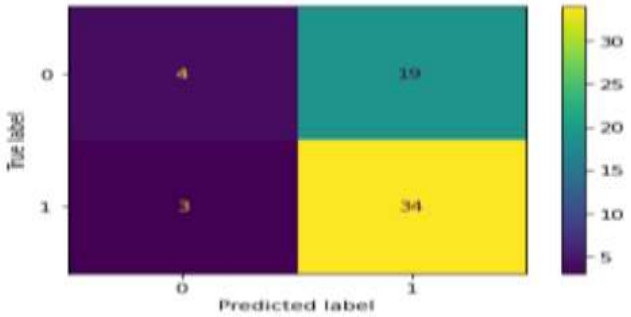

Figure-5

Table-5: Comparing Accuracies of various Classification Algorithms

Method	Accuracy	Precision	Recall	F1 Score
SVM Linear	0.65	0.65	0.95	0.77
SVM RBF	0.83	0.83	0.92	0.83
LOGISTIC REGRESSION	0..63	0.64	0.76	0.68

Table-5 compares how different algorithms (SVM Linear, SVM RBF, Logistic Regression) got various Performances (Accuracy, Precision, Recall, F1 Score).

5. Conclusion

In this work, the data sets used have been collected from real IPL cricket matches and impractical features have been removed in preprocessing with few other data cleaning steps. Additionally, suitable data is converted to a numeric form. First and foremost, the cleaned data which is used for win prediction is trained and classified with three classifiers SVM Linear, SVM RBF, Logistic Regression. Subsequently, the cleaned dataset which is used for runs prediction is trained with three regressors Random Forest, Ridge Regression, Lasso Regression, and python tool is used in both the predictions. Good results have been achieved using the SVM RBF classifier for a win and Random Forest Regressor for runs with an overall accuracy of 83% and 75% respectively.

As our approach well predicts the IPL in
the current scenario that is based on the historical records, it can be further extended after youngsters join the team, their history records are made available. Moreover, new season data can be added, and adding some new features like head-to-head win which are beneficial in increasing accuracy.

6. References

1.Dynamic Winner Prediction in Twenty20 Cricket: Based on Relative Team Strengths, Sasank Viswanadha, Kaustubh Sivalenka, Madan Gopal Jhawar, and Vikram Pudi. 19

2.Predicting the Winner in One-Day International Cricket, by Ananda Bandulasiri

3.Chellapilla is a type of chellapilla. Deep Prakash, C. Patwardhan, and C. Vasantha, Deep Mayo Predictor for IPL-9 based on Data Analytics.

4.Prediction of the Indian Premier League-IPL 2020 using Data Mining Algorithms, Priyanka S, Vysali K, Dr. K B PriyaIyer.

5.Predicting Results of Indian Premier League T-20Matches Using Machine Learning, Shilpi Agrawal, Suraj Pal Singh, and Jayesh Kumar Sharma.

6. Amal Chaminda Kaluarachchi and Aparna S. Varde: CricAI: A Classification-Based Tool for Predicting ODI Cricket Outcomes.

7.Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach, Madan Gopal Jhanwar, and Vikram Pudi.

8.ECML-98, Springer Berlin Heidelberg, 1998, 137-142. 8. Joachims T, Nedellec C, and Rouveirol C., "Text categorization with Support Vector Machines: Learning with Many Relevant Features Machine Learning," ECML-98, Springer Berlin Heidelberg, 1998, 137-142.

9. Kansal, P., Kumar, P., Arya, H., and Methaila, A., "Player value in the Indian Premier League auction using data

mining technique," International Conference on Contemporary Computing and Informatics (IC3I), 2014, 197-203.