IPL Score Prediction Using Machine Learning

MINI-PROJECT

BACHELOR OF TECHNOLOGY

ELECTRONICS AND COMMUNICATION, ENGINEERING

DEPT. OF ECE

CENTRAL INSTITUTE OF TECHNOLOGY, KOKRAJHAR

BTAD, KOKRAJHAR-783370

October 2022

Submitted by:

JYOTIRMOY NATH (201902031020)

BITTU KUMAR MISHRA (201902033036)

Under the supervision of :-

Haradhan Chal SIR

Contents

0.1 ABSTRACT

In today's date data analysis is need for every data analytics to examine the sets of data to extract the useful information from it and to draw conclusion according to the information. Data analytics techniques and algorithms are more used by the commercial industries which enables them to take precise business decisions. It is also used by the analysts and the experts to authenticate or negate experimental layouts, assumptions and conclusions. In recent years the analytics is being used in the field of sports to predict and draw various insights. Due to the involvement of money, team spirit, city loyalty and a massive fan following, the Score of every matches is very important for all stake holders. In this paper, the past seven year's data of IPL containing the player's details, match venue details, teams, ball to ball details, is taken and analyzed to draw various conclusions which help in the improvement of a player's performance. Various other features like how the venue or toss decision has influenced the score of the match in last seven years are also predicted. Various machine learning and data extraction models are considered for prediction are Linear regression, Decision tree, K-means, Logistic Regression etc. The cross validation score and the accuracy are also calculated using various machine learning algorithms. Before prediction we have to explore and visualize the data because data exploration and visualization is an important stage of predictive modeling.

0.2 Introduction

Machine Learning is a branch of Artificial Intelligence that aims at solving real-life engineering problems. This technique requires no programming, whereas it depends on only data learning where the machine learns from pre-existing data and predicts the result accordingly. Machine Learning methods have benefit of using decision trees, heuristic learning, knowledge acquisition, and mathematical models. It thus provides controllability, observability, stability and effectiveness. The cricket game has various forms such as Test Matches, Twenty20 Internationals, Internationals one day, etc. IPL is also one of them, and has great popularity among them.There are eight teams which represent eight cities which are chosen from an auction. These teams compete against each other for the trophy.The match that is played before the day is also will make a change in the prediction. The stakeholders are much more benefited due to the huge popularity and the huge presence of people at the venue. The accuracy of a data depends on the size of the data we take for analysing and the records that are taken for predicting the Score. Cricket is a game played between two teams comprising of 11 players in each team. Considering unpredictability of this unpredictable game, there is a huge interest among the spectators to do some prediction either at the start of the game or during the game. Many spectators also play betting games to win money.

0.3 Plan of Implementation

The project can be broken down into 7 main steps which are as follows:

1. Understand the dataset.
2. Clean the data.
3. Analyse the candidate columns to be Features.
4. Process the features as required by the model/algorithm.
5. Train the model/algorithm on training data.
6. Test the model/algorithm on testing data.
7. Tune the model/algorithm for higher accuracy.

0.4 Methodology

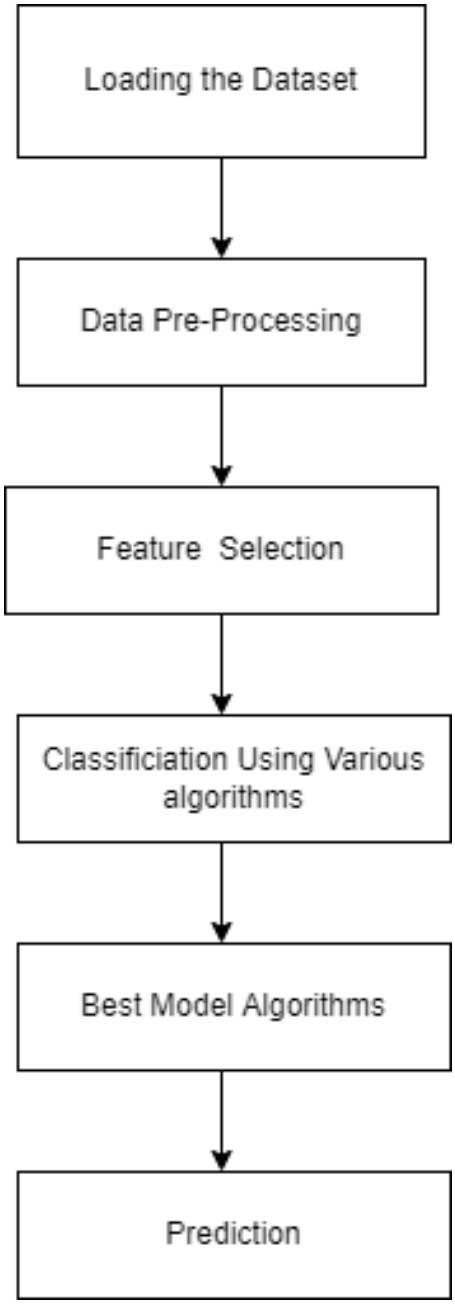

Figure 1: Process of IPL Score Prediction

0.5 Data Collection:-

Data collection is the process of gathering and measuring information from countless different sources. In order to use the data, we collect to develop practical machine learning solutions. Collecting data allows you to capture a record of past events so that we can use data analysis to find recurring patterns. From those patterns, you build predictive models using machine learning algorithms that look for trends and predict future changes. We take Kaggle's IPL Complete Dataset (2008-2017). It Contains one Csv files: Matches.csv which data size is 117,096 bytes The Matches dataset contains 18 attributes and 756 records. The dataset has the columns regarding match-number, IPL season year, the place where match has been held and the stadium name, the match winner details, participating teams, the margin of winning and the umpire details, player of the match etc. Here, some of the columns may contain null values and some of the attributes may not be required for Score prediction which is discussed in data preprocessing.

0.6 Data Preprocessing

0.6.1 Data cleaning

There are some null values in the dataset in the columns such as winner, city, venue etc. Due to the presence of these null values, the classification cannot be done accurately. So, we tried to replace the null values in different columns with dummy values.

0.7 Choosing Required Attributes

This step is the main part where we can eliminate some columns of the dataset that are not useful for the estimation of score prediction of the team. This is estimated using feature importance. The considered attributes have the following feature importance.

0.8 Data Visualization

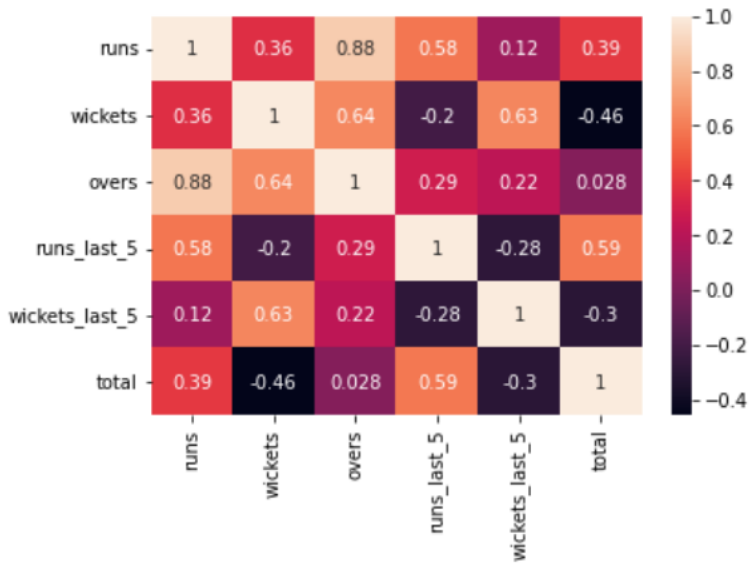

Figure 2: Data Visualization

- Matplotlib Library is used here for visualizing the graphs
- The data visualization is necessary to understand the solution in a better way. The below graphs were drawn based up on the previous seasons of the IPL matches.

0.9 Model Development and Evaluation

Here, we have developed a generic model and applied all classification methods. The data is split into training data and test data, we train the model using certain features and use it to predict the testing data, then we calculate the performance of the system. The various classification models used are: Logistic Regression, Gaussian Naïve Bayes Classifier, KNN (K Nearest Neighbor) algorithm, Support Vector Machines, Gradient Boost Algorithm, Decision Trees and Random Forest Classifier. Among these methods the Random Forest have given good results.

0.10 Model Used

For IPL Score Prediction We Compare various Machine learning algorithms.among these algorithms Random forest Shows 93 percent Accuracy.so we have choose Random Forest Algorithm.Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

0.11 Results

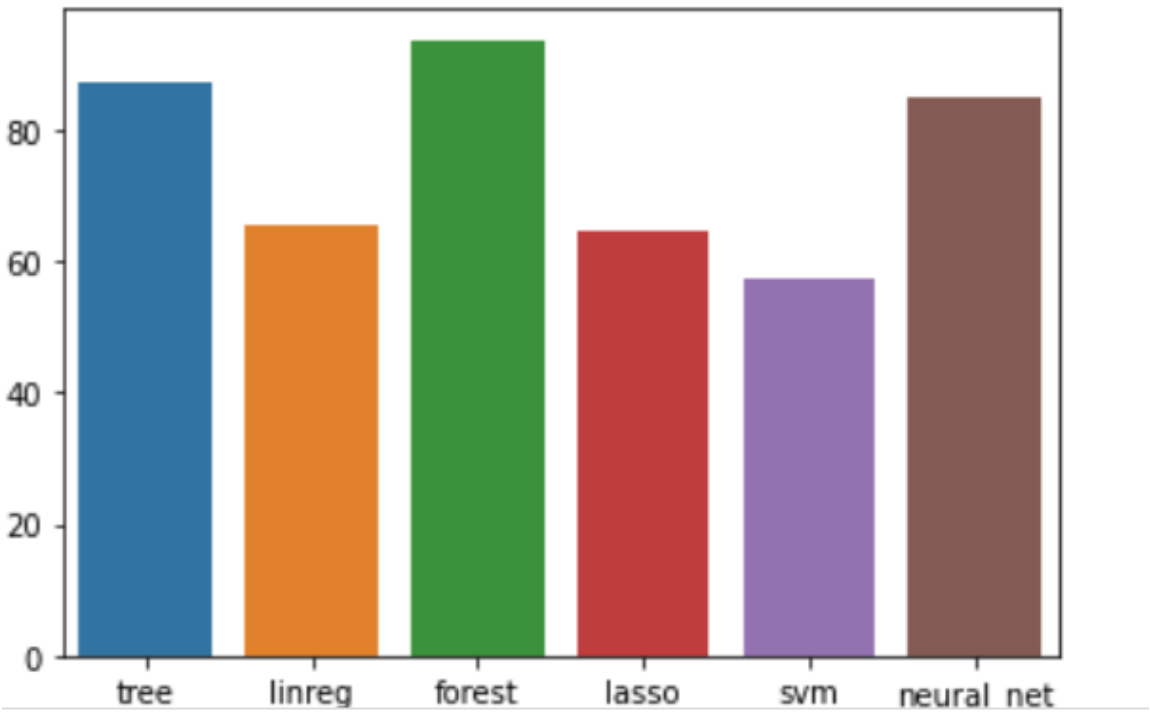

Figure 3: Accuracy of the Algorithms

- Batting Team : **Delhi Daredevils**
- Bowling Team : **Chennai Super Kings**
- Final Score : **147/9**

```
[ ]  batting_team='Delhi Daredevils'
     bowling_team='Chennai Super Kings'
     score = predict_score(batting_team, bowling_team, overs=10.2, runs=68, wickets=3, runs_last_5=29, wickets_last_5=1)
     print(f'Predicted Score : {score} || Actual Score : 147')

     Predicted Score : 150 || Actual Score : 147
     /usr/local/lib/python3.7/dist-packages/sklearn/base.py:451: UserWarning: X does not have valid feature names, but Ranc
       "X does not have valid feature names, but"
```

Figure 4: Test 1

- Batting Team : **Mumbai Indians**
- Bowling Team : **Kings XI Punjab**
- Final Score : **176/7**

```
[ ] batting_team='Mumbai Indians'
    bowling_team='Kings XI Punjab'
    score = predict_score(batting_team, bowling_team, overs=12.3, runs=113, wickets=2, runs_last_5=55, wickets_last_5=0)
    print(f'Predicted Score : {score} || Actual Score : 176')

    Predicted Score : 185 || Actual Score : 176
    /usr/local/lib/python3.7/dist-packages/sklearn/base.py:451: UserWarning: X does not have valid feature names, but Ran
      "X does not have valid feature names, but"
```

Figure 5: Test 2

```
    batting_team='Kings XI Punjab'
    bowling_team='Chennai Super Kings'
    score = predict_score(batting_team, bowling_team, overs=18.0, runs=129, wickets=6, runs_last_5=34, wickets_last_5=2)
    print(f'Predicted Score : {score}')

    Predicted Score : 147
    /usr/local/lib/python3.7/dist-packages/sklearn/base.py:451: UserWarning: X does not have valid feature names, but Randc
      "X does not have valid feature names, but"
```

Figure 6: Test 3

0.12 FUTURE SCOPE AND CONCLUSION

Prediction of the Score for a cricket match plays a significant role for the team's victory. The main goal of this paper is to analyse the IPL cricket data and predict the players' performance. Here, three classification algorithms are used and compared to find the best accurate algorithm. The implementation tools used are Googla Colab. Random Forest is observed to be the best accurate classifier with 93player performance. This knowledge will be used in future to predict the score of the matches for the next series IPL matches. . This project opens scope for future work in the field of cricket and predicting other important things like best team of players, best venue, best city, best fielding decision to win a match.