**Question-3**

**Problem Statement:** To predict the final score of the first innings of an IPL match given detailed information of what has happened in the innings till a randomly chosen point between the 8th and 12th over.

**Dataset Description:**

* The Dataset contains ball-by-ball information of the matches played between IPL Teams but the only relevant fields for us to make predictions would be more or less: [batting\_team, bowling\_team, over, ball, extra\_runs, total\_runs].
* Since the total\_runs column only contains the information of the total run scored on that particular ball, we required two more field of the cumulative sum of data fields of the total\_runs column and final score in the training dataset in order to predict the final score after that inning.
* We removed all the rows after the 12th over from the training dataset since they were irrelevant to us.
* While data preprocessing, we performed label encoding, one-hot encoding, and column transformation.

**Algorithms on which it was trained and tested:**

| Algorithms used: |
| --- |
| Decision Tree Regressor |
| Linear Regression |
| Random Forest Regression |
| Lasso Regression |
| Support Vector Machine Regression |
| Neural Network Regression |

I chose Random Forest Regressor for the final model since the training score was 93.26%, Test Score was 58.68% and Mean absolute error was the minimum in it of all the algorithms used of 14.71.

* Created a prediction\_score array for all the teams of IPL