CrickXpert

Manan Chugh 2021335
Abhijeet Anand 2021509
Kartikey Dhaka 2021534
Abhijay Singh 2021226

Abstract

The project employs advanced machine learning techniques to accurately predict outcomes during cricket innings. This encompasses vital metrics such as runs scored, wickets taken, boundaries hit, and even the likelihood of specific events unfolding during a game. By meticulously analyzing historical data and identifying recurring patterns, the system offers invaluable insights. These insights are instrumental for teams and coaches in formulating tactical game strategies and making real-time decisions during matches. Beyond strategy, this application of machine learning also significantly enriches the understanding of cricket for enthusiasts and professionals alike. It underlines the ever-increasing role of data science in sports, transforming traditional cricket viewing into a more informed and immersive experience. Through this initiative, the project demonstrates the potential of merging technology with sports to enhance comprehension and enjoyment of the game.

- **Introduction**

Merging our love for sports and technology, the allure of predicting cricket scores using machine learning has captivated us. This endeavor enriches our passion for the game, sharpening our skills and introducing an added layer of thrill. Driven by a shared ambition, we embrace this journey, recognizing the transformative potential of data science in sports. This project harnesses machine learning to forecast cricket inning results, including key metrics like runs, wickets, boundaries, and event likelihoods. By revealing patterns, it aims to inform strategies and facilitate real-time analysis, highlighting data science's role in amplifying cricket comprehension and enjoyment. Our goal is to create adept machine learning models that predict crucial metrics in an IPL cricket match. By delving deep into the game's dynamics, we anticipate offering unparalleled insights. This initiative will not only contribute to the growing realm of sports analytics but also underscore the potential of AI in predicting cricket match outcomes. By merging technology with sports, we're setting the stage for a paradigm shift in how fans, team managements, and broadcasters engage with cricket. The culmination of our efforts will be a well-documented, functional system making real-time predictions during live matches, reshaping cricket analytics and augmenting fan engagement

- **Literature Survery**

1. https://www.analyticsvidhya.com/blog/2022/05/ipl-team-win-prediction-project-using-machine-learning/ The paper on 'IPL Team Win Prediction Using Machine Learning' discusses the prediction of IPL match winners based on match statistics. The project familiarizes readers with exploratory data analysis and feature engineering techniques essential for processing the data. The dataset used contains details like teams played, winner, venue, wickets, runs, toss decision, and whether DLS was applied, among other variables, which are analyzed to predict match outcomes 1.

2. https://github.com/PtPrashantTripathi/IPL-2020-Prediction The 'IPL 2020 Prediction' project employs predictive analysis to forecast outcomes of IPL matches, utilizing data from 2008 to 2020. The predictions are driven by data analysis. Their advanced algorithms meticulously analyze various factors such as team and player form, expected line-ups, batting orders, pitch,

weather conditions, and extensive historical match stats to provide precise predictions. This valuable insight enhances understanding and engagement for fans, teams, and cricket enthusiasts, elevating their comprehension of the game.

- **Dataset**

1. **Data Description**

The dataset consists of the data of **IPL matches** from 2007 to 2017. The dataset is of 76015x15 size. It consists of the following columns

1. **Mid:** Match identifier number.
2. **Date:** Date of the match.
3. **Venue:** Stadium where the match was played.
4. **Bat_Team:** Batting team's name.
5. **Bowl_Team:** Bowling team's name.
6. **Batsman:** Name of the batsman playing.
7. **Bowler:** Name of the bowler bowling.
8. **Runs:** Runs scored by the batsman.
9. **Wickets:** Number of wickets taken by the bowler.
10. **Overs:** Number of overs bowled.
11. **Runs_Last_5:** Runs scored in the last 5 overs.
12. **Wickets_Last_5:** Wickets taken in the last 5 overs.
13. **Striker:** Runs scored by the batsman currently facing.
14. **Non-Striker:** Runs scored by the batsman not facing.
15. **Total:** Total runs scored by the batting team in the match.

2. **Data Extraction**

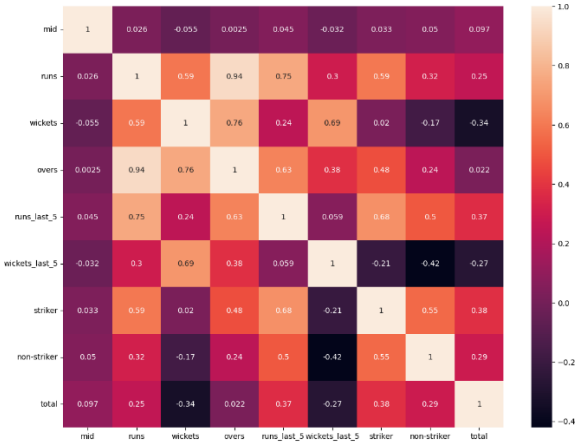
We scraped the internet for IPL matches data and we were able to collect the data for all the IPL matches from 2008 - 2017. After extracting this data we were left with 75000+ rows to work upon.

3. **Data Preprocessing**

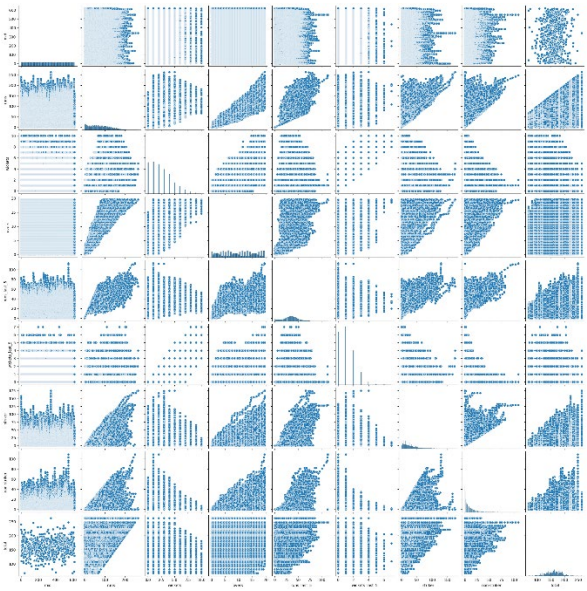
For optimal results we deleted all the rows which had any missing column in them. Then we thoroughly analyzed the data and came upon a conlusion that only 9 features were important for our model development("mid", "date", "bat_team", "bowl_team", "runs", "wickets", "overs", "runs_last_5", "wickets_last_5", "total", "runs_in_each_ball"). For more accurate results we computed the feature "runs_in_each_ball" and added it to our dataset. For better comparison of the dates we converted the "date" feature from String datatype to Date_Time datatype. We used One Hot Encoding to encode the categorical data ("bat_team", "bowl_team").

4. **Data Visualization**

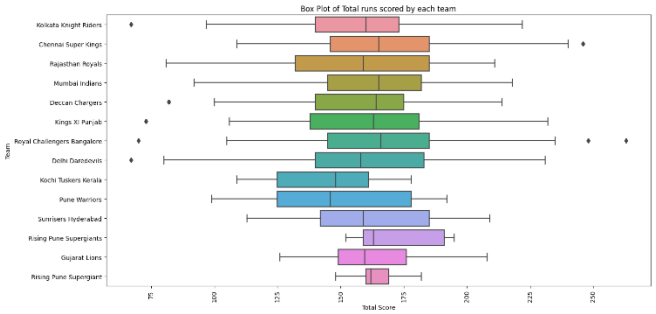
Correlation HeatMap:

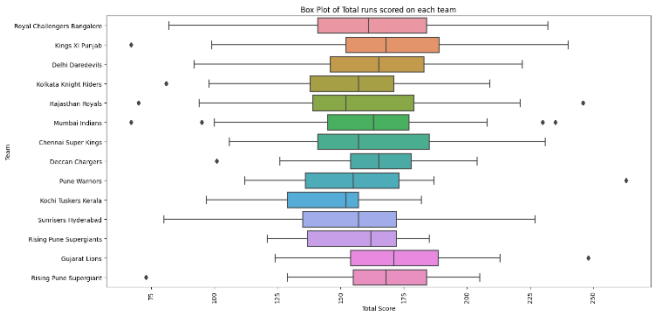


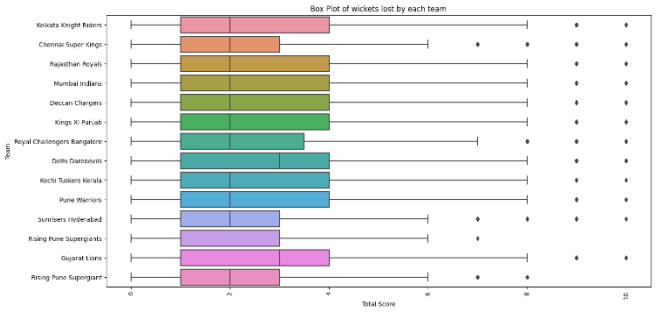
Pair Plots:

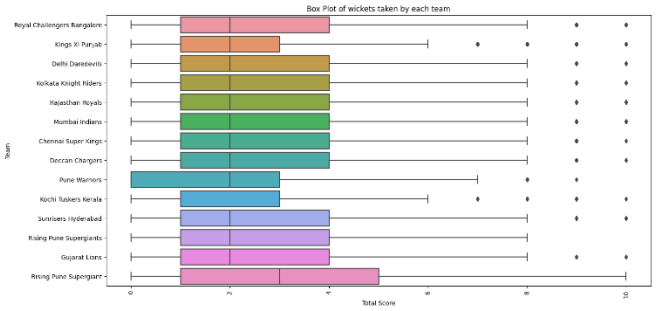


Box Plots:

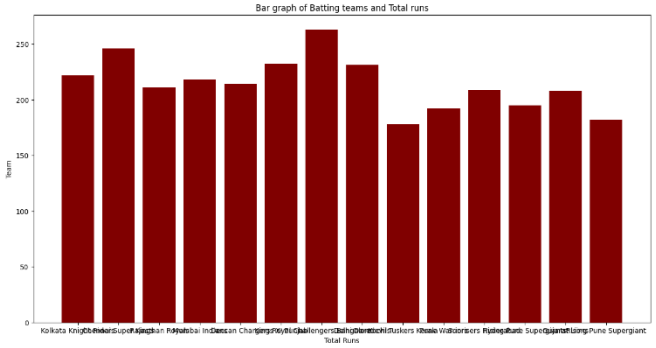


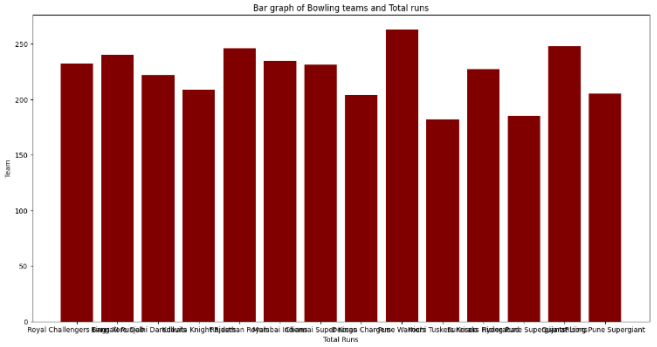




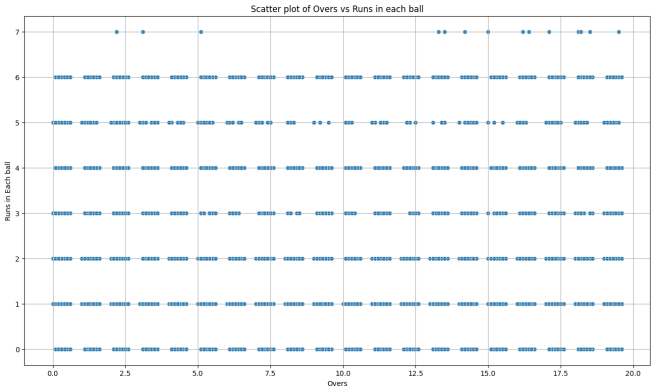


Bar Graphs:

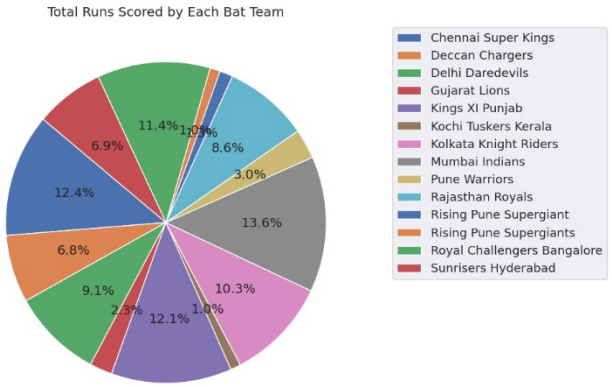




Scatter Plot:



Pie Chart:



Insights:

1. From the correlation heatmap we inferred the following:
   a) Runs and overs show very high positive correlation.
   b) Wickets and overs also show high positive correlation.
   c) Wickets and total runs are negatively correlated.
2. Runs and overs show linear dependency to some extent.

3. From the box plots we inferred:
   a) The median runs scored by each team is in the range of 150-175.
   b) Matches in which all 10 wickets fall are computed to be outliers.
   c) The median wickets taken in each match are 2-3.

4. There is only one team (Pune Warriors) which had matches with 0 wickets taken by them.

5. Mumbai Indians and Chennai Super Kings scored the highest number of runs per match on average.

6. Rising Pune Super Giants and Kochi Tuskers Kerala were the least-scoring teams per match on average.

7. The maximum number of runs that can be achieved on one ball is 7 assuming a wide ball and 6 runs.

- **Methodology**

The main objective of our model is to predict the detailed insights of a IPL Matches. Firstly, we will split the encoded dataset according to date, the datapoints before the year 2016 will be used in training the model and the datapoints after 2016 will be used in testing. We are dividing it along time to maintain a timeline continuity in the data. We will try many Regression techniques such as Linear Regression, Decision Tree Regression, Random Forest Regression and finally choose the most optimal algorithm. With our Regression techniques, we will use AdaBoosting to increase the performance of our model. Adaboost is a powerful technique that converts the weak learners(multiple models) to strong learners(final combined model). At the end, we will perform K-Cross Validation by dividing the test and train dataset according to different years.

- **Result and Analysis**

The evaluation metrics for the Linear Regression model came out as following: -

1. Mean Absolute Error (MAE): 12.11
2. Mean Squared Error (MSE): 251.01
3. Root Mean Squared Error (RMSE): 15.84

The evaluations metrics for the Decision Tree Regression came out as following: -

1. Mean Absolute Error (MAE): 17.08
2. Mean Squared Error (MSE): 531.05
3. Root Mean Squared Error (RMSE): 23.04

The evaluations metrics for the Random Forest Regression came out as following: -

1. Mean Absolute Error (MAE): 13.76
2. Mean Squared Error (MSE): 330.21
3. Root Mean Squared Error (RMSE): 18.17

After testing all the different models, the Linear Regression Model is giving the best accuracy, so we will continue with the Linear Regression Model.

- **Conclusion**

1. **Learnings from the Project**

In this project focused on Indian Premier League (IPL) data, we gained hands-on experience in constructing a machine learning pipeline. We realized the crucial role of data collection and utilized the IPL API to gather comprehensive match statistics, player performances, and team records. Utilizing various data visualization tools like matplotlib and seaborn, we analyzed trends and patterns in the IPL dataset. Additionally, we learned about mutual information classification for effective feature selection. Overall, this project equipped us with essential skills for handling IPL data and building predictive models for cricket analytics.

2. **Future Work**

We have successfully adhered to the outlined project timeline. We curated the necessary datasets, developed, and tested a model for predicting key metrics in a cricket match, achieving satisfactory results. Looking ahead, our plans for future enhancements include:

1. Hyper-parameter tuning along with different algorithms for predicting other cricket metrics.
2. Including other cricket-based datasets and using much more complex techniques for better accuracy.
3. Creating a pipeline that can aggregate the results of all the models to give the final optimal output.

- **Member Contribution**

All four members worked collaboratively on all the tasks in our timeline. The work was distributed equally during this deadline based on one's expertise in a certain topic and their availability at the time.

References

1. https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners/#:~:text=AdaBoost%20algorithm%2C%20short%20for%20Adaptive,assigned%20to%20incorrectly%20classified%20instances.
2. https://www.geeksforgeeks.org/decision-tree/
3. https://www.javatpoint.com/machine-learning-random-forest-algorithm