FINAL PROJECT REPORT – USABLE AI

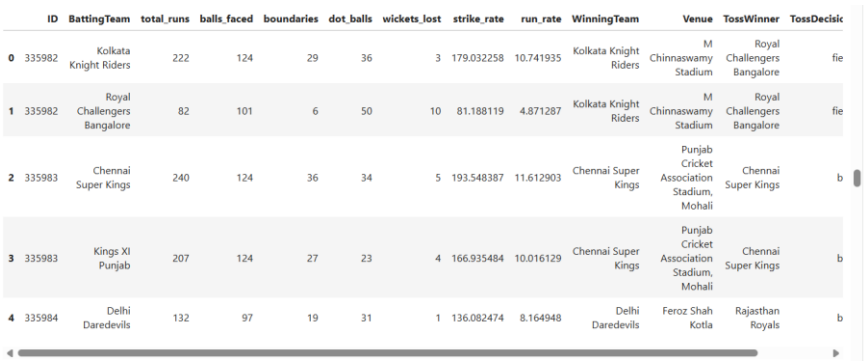
SAKSHI NAIR (saknair)

- **INTRODUCTION:** With elite domestic and international players, the Indian Premier League (IPL) is one of the world's most competitive and data-rich cricket competitions. The league offers a unique opportunity to use machine learning techniques for predictive analysis because of its size and the depth of historical data that is available. This project aims to forecast two key outcomes of an IPL match: (1) the total number of runs a team is expected to score, and (2) the likely match outcome (win or loss) for a team based on the factors. By analyzing ball-by-ball match events, team compositions, player statistics, and match conditions such as venue and toss decisions, I developed models that provide insights into what influences match performance. These predictions can support strategic decision-making for teams and provide fans and analysts with deeper, data-driven perspectives on the game. The project includes detailed data cleaning, exploratory data analysis (EDA), feature engineering, and the implementation of regression and classification models to address the key questions.

- **DATASET DESCRIPTION:** This project utilizes three key datasets sourced from publicly available IPL archives. These datasets cover both granular ball-level events and match-level summaries, allowing for a comprehensive analysis of team and player performance.

  1. Ball-by-Ball Dataset (IPL_Ball_by_Ball_2008_2022.csv): This dataset contains detailed information about each ball played, including player statistics, runs scored, wickets taken, and extra runs. This dataset provides granular insights into individual player performance.
  2. Match Results Dataset (IPL_Matches_Result_2008_2022.csv): Includes data on match outcomes, venue, toss decision, and match margin. This dataset helps us understand the broader context of each match, such as team composition and match location.
  3. Additional Data (data.csv): Contains other match-related details, including predicted runs, batting team, bowling team, and venue specifics.
  These datasets are merged to form a comprehensive dataset, which is then pre-processed for model training.

  I then combined important variables from all 3 datasets into one dataframe and named it as merged_data.



	ID	BattingTeam	total_runs	balls_faced	boundaries	dot_balls	wickets_lost	strike_rate	run_rate	WinningTeam	Venue	TossWinner	TossDecisic
0	335982	Kolkata Knight Riders	222	124	29	36	3	179.032258	10.741935	Kolkata Knight Riders	M Chinnaswamy Stadium	Royal Challengers Bangalore	fie
1	335982	Royal Challengers Bangalore	82	101	6	50	10	81.188119	4.871287	Kolkata Knight Riders	M Chinnaswamy Stadium	Royal Challengers Bangalore	fie
2	335983	Chennai Super Kings	240	124	36	34	5	193.548387	11.612903	Chennai Super Kings	Punjab Cricket Association Stadium, Mohali	Chennai Super Kings	b
3	335983	Kings XI Punjab	207	124	27	23	4	166.935484	10.016129	Chennai Super Kings	Punjab Cricket Association Stadium, Mohali	Chennai Super Kings	b
4	335984	Delhi Daredevils	132	97	19	31	1	136.082474	8.164948	Delhi Daredevils	Feroz Shah Kotla	Rajasthan Royals	b

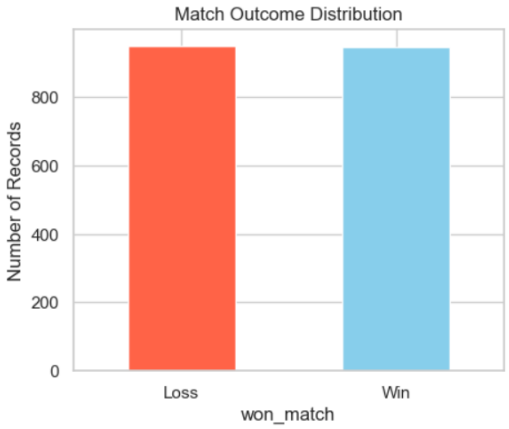
- **DATA CLEANING:**

Firstly, I checked for the missing values in each dataset. Significant missing values were present in the Ball-by-Ball dataset in columns like extra_type, player_out, kind, and fielders_involved. This is to be expected in rows where there were no additional runs or wickets. To keep things consistent without adding bias, the placeholder "None" was used to complete these. The mode, or most frequently occurring value, was used to fill in the missing values in the City and Winning Team columns of the Match Results dataset. This was appropriate because the features are categorical and had few missing entries. For a number of rows, the method column—which usually indicates the usage of specific match-deciding techniques like Duckworth-Lewis—was absent. To show that no unique decision-making process was applied, these were filled in with the label "normal." The extra_data (3rd dataset) consist of no missing values. The extra_data (data.csv) was already one-hot encoded, but inconsistencies in team and venue names (e.g., "Kings XI Punjab" vs. "Punjab Kings") were resolved by merging similar columns to ensure consistency. Outliers in total_runs were removed using the IQR method to prevent skewing regression models, especially when extremely low or high scores (e.g., 49 or 250) were anomalies not representative of typical team performance. These steps improved data quality and ensured consistency across features used for modeling.

- **DATA PROCESSING:**

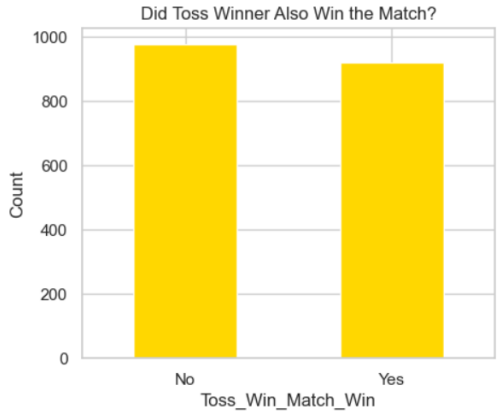
For the exploratory data analysis, I used Python with pandas, seaborn, and matplotlib to visualize and interpret team-level match performance.

1. I began by analyzing the distribution of match outcomes, confirming the dataset was balanced between wins and losses.



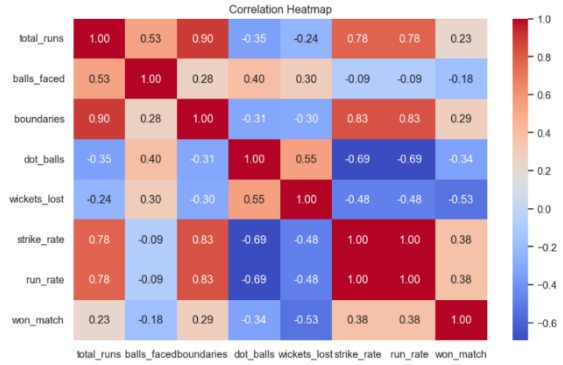
The bar plot shows a nearly equal distribution of wins and losses, suggesting that match outcomes are well-represented across the dataset. The similar bar heights indicate no major skew or bias in how matches ended. This balance enables meaningful comparisons between winning and losing teams in further analysis. A check for imbalance confirmed that outcomes were generally balanced overall, though minor fluctuations across seasons or teams were retained to preserve real-world context.

2. I explored the impact of toss results, finding only a slight advantage for toss winners.



The bar chart shows that in slightly more cases, the team that won the toss did not go on to win the match. While the difference is not large, it suggests that winning the toss does not guarantee a win, and other in-game factors play a more significant role in determining the outcome.

3. I generated a correlation heatmap.



The correlation heatmap reveals strong positive relationships between total_runs, boundaries, strike_rate, and run_rate, indicating these features tend to increase together. Importantly, won_match is moderately positively correlated with strike_rate, run_rate, and boundaries, suggesting that aggressive scoring is linked to match wins. On the other hand, dot_balls and wickets_lost show negative correlations with won_match, meaning that fewer dot balls and wickets tend to favor winning outcomes.

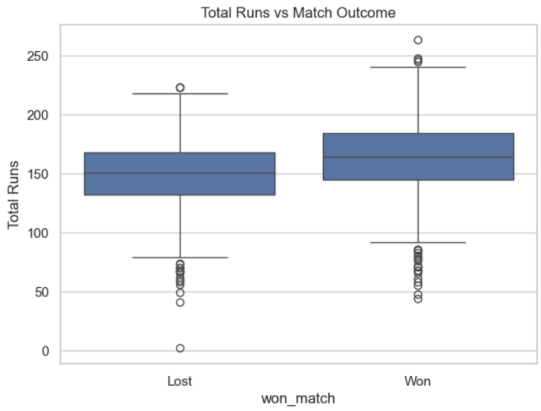
4. I carried out the summary statistics for win vs loss.

Mean Values for Teams that Won vs Lost

won_match	total_runs	boundaries	dot_balls	wickets_lost	strike_rate	run_rate
Loss	148.762605	17.380252	44.815126	7.191176	122.375996	7.342560
Win	163.292812	20.739958	38.677590	4.550740	140.221327	8.413280

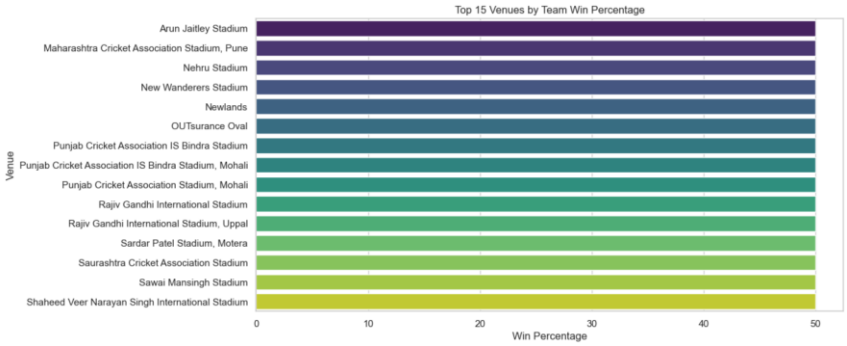
The summary table clearly shows that teams that won matches had better overall performance metrics. On average, winning teams scored more total runs (163 vs. 149), hit more boundaries, had fewer dot balls, lost fewer wickets, and maintained significantly higher strike rates and run rates. These indicators suggest that consistent scoring pressure and efficient batting were key contributors to match victories.

5. I plotted Boxplot of total_runs by match result.



The boxplot shows that teams which won matches generally scored more total runs than those that lost. The median total runs for winning teams is higher, and their interquartile range (IQR) is shifted upward. There are also more high outliers in the winning group, indicating that very high scores more often resulted in victories. This supports the conclusion that scoring more runs is a strong predictor of match success.

6. I calculated win% by Venue.



The bar plot shows the top 15 venues by win percentage, revealing that all listed venues have fairly close win rates, hovering around 45–50%. This suggests that no single venue overwhelmingly favors the batting team in terms of match outcomes, but some grounds like *Shaheed Veer Narayan Singh International Stadium* and *Sawai Mansingh Stadium* show slightly higher win rates, indicating a possible mild influence of venue on team performance.

7. I visualized the Toss Decision.

The bar chart shows that teams winning the toss tend to choose **fielding first** far more often than batting. This suggests a strategic preference in IPL matches, possibly due to pitch conditions, dew factor, or chasing advantages. The imbalance also highlights a common perception among teams that bowling first may offer a better chance of controlling the game.

Feature Engineering - To predict match outcomes, I aggregated team-level performance statistics per match using features like total_runs, boundaries, dot balls, and strike rate. Each row in the modeling dataset represents a team's performance in a single match, allowing the classifier to learn win/loss patterns from structured team-based feature

- **KEY QUESTIONS:**

1. Can we predict the number of runs a team will score based on historical match data and team composition?

A. I utilized a Random Forest Regressor to predict team ratings because of its capacity to manage non-linear relationships and feature interactions. Without requiring a lot of feature engineering, it offers dependable performance and performs well with structured data. To estimate the effects of variables like strike rate, boundaries, and match circumstances on total runs, Random Forest was the perfect option due to the complexity and unpredictability of cricket data.

```
Random Forest Regressor: RMSE ≈ 8.77 runs; R² ≈ 0.92
```

These findings imply that I may use match characteristics and previous data to forecast a team's score with a high degree of accuracy. In the context of IPL cricket, an RMSE of approximately 9 runs indicates that the Random Forest model's predictions are, on average, within **8 to 9 runs** of the actual total. This is a very accurate estimate. A very strong fit is confirmed by the model's R2 value of 0.92, which shows that it accounts for **92%** of the variance in team ratings. Strike rate, boundaries, dot balls, and run rate are important factors that influence these

forecasts. Other factors include match context, such as whether the team won the toss or batted first.

A team that plays first on a favorable pitch and has an aggressive batting lineup, for example, is expected to score a lot more runs. As predicted in cricket, there is still some degree of unpredictability, but this model offers a trustworthy means of assessing team performance and guiding strategy.

2. What factors influence the success of a team in the IPL?

A. For predicting team success in the IPL, I used a Logistic Regression model. This model is well-suited for binary classification problems like win or loss and provides a straightforward interpretation of how each feature affects the outcome. Logistic Regression is particularly useful when we want to understand the direction and strength of relationships between match performance metrics (like total runs, strike rate, dot balls, and wickets lost) and match results. Its simplicity, efficiency, and interpretability made it a strong choice for identifying key indicators of winning a match in a statistically sound way.

```
Logistic Regression: Accuracy ≈ 0.75
Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.72      0.75       199
           1       0.72      0.78      0.75       181

    accuracy                           0.75       380
   macro avg       0.75      0.75      0.75       380
weighted avg       0.75      0.75      0.75       380
```

Several factors were found to significantly influence a team's success in the IPL. Higher total runs and strike rates are strong indicators of winning, reflecting the importance of aggressive and efficient scoring. Lower dot balls and fewer wickets lost also contribute positively, highlighting the value of maintaining momentum and preserving wickets. The model, with **~75%** accuracy, confirms that consistent batting performance and minimizing errors (like losing quick wickets or playing dot balls) are key drivers of victory. It shows balanced precision, recall, and f1-scores of 0.75 for both classes, indicating consistent and reliable classification performance. **These insights align with core IPL strategies and validate the predictive power of team-level match statistics.**

3. Can player statistics (e.g., batting average, strike rate) be used to predict team performance?

A. Using the XGBoost Classifier, I assessed whether aggregated player-level variables like run and strike rates alone might predict team success. The impact of individual batting performance on match outcomes is isolated in this question, in contrast to earlier ones that involved a wider match context. XGBoost was chosen because it can efficiently handle small feature sets and model intricate, non-linear relationships. Without depending on team names, toss results, or venues, this enabled me to evaluate the extent to which player statistics alone can predict win.

```
XGBoost Classifier (Player Stats Only): Accuracy ≈ 0.64
Classification Report:
              precision    recall  f1-score   support

           0       0.69      0.56      0.62       199
           1       0.60      0.73      0.66       181

    accuracy                           0.64       380
   macro avg       0.65      0.64      0.64       380
weighted avg       0.65      0.64      0.64       380
```

The XGBoost Classifier trained on player performance metrics — specifically strike rate and run rate — achieved an accuracy of approximately **64%.** The model showed balanced precision and recall across both winning and losing classes, with an F1-score of 0.66 for wins and 0.62 for losses. **These results suggest that player-level stats carry meaningful predictive power, particularly in identifying strong batting performances that often correlate with victories.** However, the moderate accuracy also indicates that other contextual features (like venue, toss outcome, and bowling stats) are needed to further improve prediction reliability.

4. Does the venue (stadium) have an impact on the match outcome?

A. For this question, I used Logistic Regression to test whether the match venue alone could help predict the outcome of an IPL match. Logistic Regression was chosen for its simplicity and interpretability, allowing me to directly assess whether certain venues consistently favor winning outcomes. The goal was to isolate venue influence without involving team performance or match conditions.

```
Logistic Regression (Venue Only): Accuracy ≈ 0.38
Classification Report:
              precision    recall  f1-score   support

           0       0.40      0.37      0.38       199
           1       0.37      0.40      0.38       181

    accuracy                           0.38       380
   macro avg       0.39      0.39      0.38       380
weighted avg       0.39      0.38      0.38       380
```

The above model achieved an accuracy of only **38%**, confirming that **venue alone is not a strong predictor** of match outcomes. Both win and loss classes had low precision and recall, meaning the model struggled to learn meaningful patterns based solely on where the match was played. This result supports the idea that **while venue might have minor effects**, it does **not significantly influence results on its own**.

- **<u>Discussions and Conclusion:</u>**

This project explored the use of machine learning models to predict IPL match outcomes and team performance based on historical data. The analysis highlighted that team batting performance, particularly variables like strike rate, boundaries, and total runs, plays a significant role in determining match outcomes. These features were consistently influential across both regression and classification tasks. The Random Forest Regressor provided highly accurate predictions for total runs scored by a team, achieving an R^2 of 0.92 and an RMSE of ~8.8, indicating strong predictive power. Similarly, Logistic Regression and XGBoost classifiers showed that team success can be reasonably predicted using batting-related stats, with accuracies ranging from 64–75%.

However, not all features proved useful. For instance, when venue was tested as a sole predictor using both Logistic, performance dropped significantly (accuracy ≈ 38–39%). This indicates that venue alone does not significantly influence the outcome and must be combined with other contextual features to add predictive value.

- **Limitations and Challenges:**
  1. Limited feature diversity: Some potentially important variables, like real-time weather conditions or individual player matchups, were not available in the dataset.
  2. Outliers and noise: Despite cleaning, cricket data can be unpredictable — due to collapses, exceptional innings, or rare events — which models may not fully capture.
  3. Venue inconsistency: Different naming formats for the same venue required merging, but this might still have introduced noise.

- **Handling Negative Results:**
  The venue-only models are an ideal example of a negative outcome. Despite the initial assumptions that certain stadiums would favor particular results, our investigation showed that the venue has no effect on its own. Acknowledging such results is crucial because they aid in the improvement of theories and direct further research.

- **Future Work:**
  1. Add individual player form and bowling performance metrics.
  2. Include features like home vs. away performance or match phase (powerplay/death overs).
  3. Investigate ensemble models that combine classification and regression to create more complex prediction layers.