IPL Data Analysis and Predicting Match outcome

Executive Summary

Cricket in India isn't just a game, it's a cultural phenomenon that brightens my childhood summers every March through May. With the IPL at its heart, I mined the 2008–2024 Kaggle dataset to deliver actionable insights for teams, auction strategists, sponsors, and cricket enthusiasts alike.

In my analysis, I first examined how teams have performed over the years, then dug into team and player statistics and venue information to identify any trends. After that, I applied machine learning techniques, using historical performance metrics and match conditions as inputs to predict the winning team of a match before it even started.

From this assignment, I found that Chennai Super Kings and Mumbai Indians are tied as the most successful franchises, each winning five IPL titles. Teams like the Gujarat Titans showed high consistency with win rates above 60%, while other teams lagged behind. On the player side, Virat Kohli emerged as the top run-scorer, but AB de Villiers boasted the highest strike rate. David Warner had the most 50+ innings, and Kohli led the century count. For bowlers, Yuzvendra Chahal took the most wickets, and Sunil Narine maintained the best economy rate. When it came to venues, Eden Gardens and Wankhede Stadium hosted the most matches.

Now, for modelling, I created a processed dataframe combining both categorical and numerical features. Categorical features included toss_winner and toss_decision (bat/bowl) to capture pre-game strategic choices. Numerical features comprised batting and bowling metrics (runs, run rates, economy rates), historical form indicators (overall and venue-specific win percentages), flags for whether top players were on each side, home-ground advantage, the average economy of each team's top three bowlers, and the average tallies of the top five run-scorers. Including this mix of variables helped improve the models' predictive power.

I trained and evaluated multiple machine learning models on these features and found that XGBoost delivered a test accuracy of 0.745, meaning it correctly predicted the outcome in about 74.5% of unseen matches. This level of predictive capability could be useful for teams making auction decisions, sponsors selecting players to back, and fans placing more informed bets.

I plan to enhance this analysis by integrating real-time factors such as weather conditions and real-time player availability updates. I also hope to build a simple, user-friendly dashboard for live match predictions, making these insights easily accessible to everyone interested.