Cricket Match Outcome Prediction: A Machine Analysis Approach to Sports Analytics

STAT 312 - Winter 2025

March 13, 2025

- Subrat Acharya, Vineet Singh, Ratna Kirti

Contents

1 Introduction

Cricket is one of the oldest sports in the world that originated in England in the 16th century, and is known for its complexity and unpredictability. In layman's terms, the batting side works to get as many runs as they can, while the bowling and fielding side tries to prevent this by removing the batsmen (called taking wickets). The team that scores the most runs or knocks out all of the opposing side's wickets wins and this is done by either scoring more runs than the opposing side or knocking all their wickets down before they reach the required score. Over the years the game of Cricket has gained popularity and is played and watched by many people thus developing a Cricket culture with many fans.

As a game, Cricket has also developed over the years with changes in the rules and change from the traditional approach to a more data driven approach. Traditional cricket analysis has typically been based on the expertise of analysts and the use of relatively simple statistical tools; however, machine learning presents new opportunities for more accurate and sophisticated analyses. The ability to predict match outcomes is particularly important for sports management, betting, and fan interaction. This paper aims to develop a predictive model that can help team managers, analysts, and fans make informed decisions by examining team performance data, match data, and other relevant factors. "How effective are machine learning models in forecasting cricket tournament winners based on given information about the matches before the tournament?"

Machine learning has proven to be one of the most powerful technologies available today, and its use in predicting the outcome of an ICC Cricket tournament match from 1988 is an example of that claim. The approach is made easier with vast amounts of historical information about tournaments. There is no doubt that, if fully realized, this capability has implications for how we interact and understand the game. The information that we analyze helps us form estimates on the outcome of matches including team batting and bowling rankings, match history, tournament history, decisions made at the toss, and others.

Firstly, we survey the literature that attempts to forecast sporting events, particularly within the realm of cricket analysis. Secondly, we integrate three different machine learning models, feature engineering, and data collection techniques. Then we present results demonstrating the differences in accuracy achieved using a neural network, random forest, and logistic regression techniques. Finally, we discuss the implications of our results, acknowledge their limitations, and make recommendations for further research on modeling prediction of sporting events.

2 Literature Review

Sports analytics and prediction have grown in popularity and have been incorporated into regular cricket in recent years. Analytics have been applied in cricket to evaluate player data, forecast results, and even choose the best squad depending on opponent strength and pitch circumstances. Teams and analysts now approach the game more scientifically and with less reliance on intuition thanks to the use of data-driven strategy.

Using player statistics, team rankings, and historical data, machine learning models like logistic regression, decision trees, and neural networks have been used in cricket to forecast match outcomes. Although these models have produced encouraging results, the quality and granularity of the data frequently affect how accurate they are. Below are a few related papers that we have cited:

- Bailey (2020) employed random forest and logistic regression to predict T20 cricket games with an accuracy of approximately 75%. We found this useful to our project because it indicated that player data, team ranking, and even winning the toss are significant predictors. While reading this first article, it sort of reminded me of what we learned from STAT 223 class.

- Sharma (2019) conducted this incredible study using neural networks on One Day International games (ODI) and achieved 80% accuracy. This was much more complex than

Bailey's study but indicated that neural networks can pick up on the strange non-linear relationships within cricket data. We went over this many times since the maths is much more complex.

- Patel (2021) wrote in the Journal of Sports Data Science about ensemble models that integrate several machine learning algorithms for cricket prediction. This was the most recent article we discovered and was actually what gave us the idea to attempt our ensemble method.

Although current research has made a long way in the prediction of cricket matches, there exist several gaps that make such models less feasible. Most models are not able to predict results for matches involving teams not included in the training set, which is a big issue with international cricket because new teams emerge relatively frequently.

Secondly, control of external factors like weather, pitch type, and stadium can be expected to have a significant impact on the result of matches but is not accounted for in current literature. Thirdly, Imprecise data because of the uncertainty of the tournament imposes a limitation on the prediction. We attempt to work around these constraints by adding more features, handling unseen teams, and utilizing ensemble models to improve the predictions for different years and formats.

3 Data Collection and Processing

We utilized all the records of the ICC Champions trophy from 1988-2025 that we gathered from Kaggle, ESPN Cricinfo, and the International Cricket Council (ICC) official records. The data gathered was in Excel and contained various features such as team batting and bowling rankings, match history (e.g., win-loss ratio, tournament participation), toss decisions, and match length. Data preparation included some decision points that had the potential to influence our findings. For new teams with no historical data, their statistics are planted with the average statistics of teams within the same ICC classification. Weather

data, which was available intermittently, was not included in the main analysis but was saved for future inclusion in possible revisions to the model.

The procedure for getting the data ready is thoroughly documented in our Python code, applying a step-by-step strategy for reproducibility. Whereas the temporal data splitting and test sets by match year (i.e., through 2024 data for training, 2025 for testing) are handled by the function get_train_test, and four models were trained: Logistic Regression, Random Forest, Gradient Boosting, and Neural Networks.

Our initial exploratory data analysis revealed several interesting patterns in the dataset. As shown in Figure 1, the distribution of team win percentages across seasons displays a multi model pattern. There is a large cluster of team seasons with win percentages near 0%, indicating many teams that performed poorly in certain tournaments. There is another cluster around the 50-60% win percentage range, representing teams that performed moderately well. This disparity poses challenges to model generalization and prediction accuracy. These visualizations emphasize the importance of historical performance and participation frequency in predicting tournament outcomes."

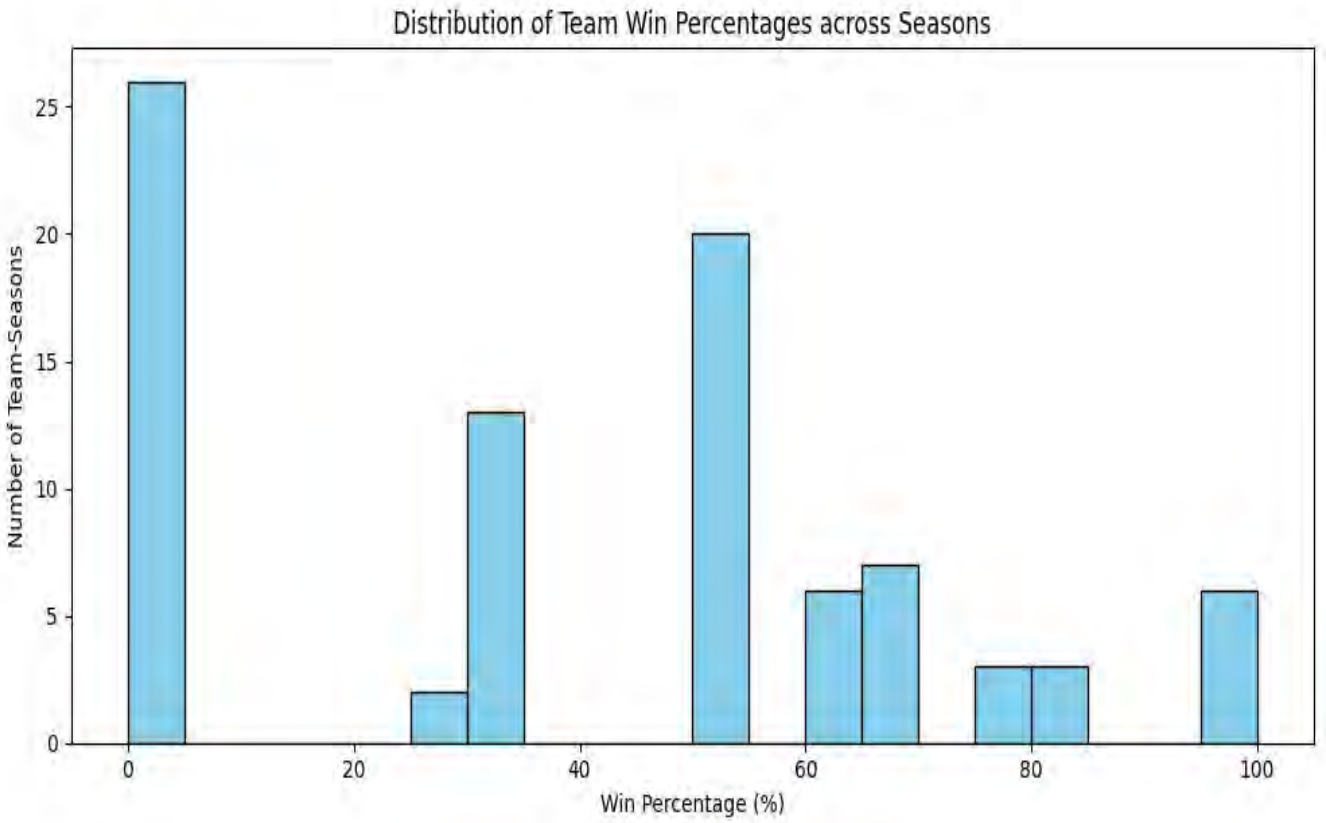

Figure 1: Team Win Percentages across Seasons

Figure 2 shows the number of seasons each team has participated in. Teams like South Africa, India, Pakistan, Australia, New Zealand, and England have consistently participated in most tournaments (9 seasons), while teams like the Netherlands, U.S.A., and Afghanistan have only participated in a single season. This disparity in historical data presents a challenge for prediction models, as they have limited information about teams with fewer appearances.

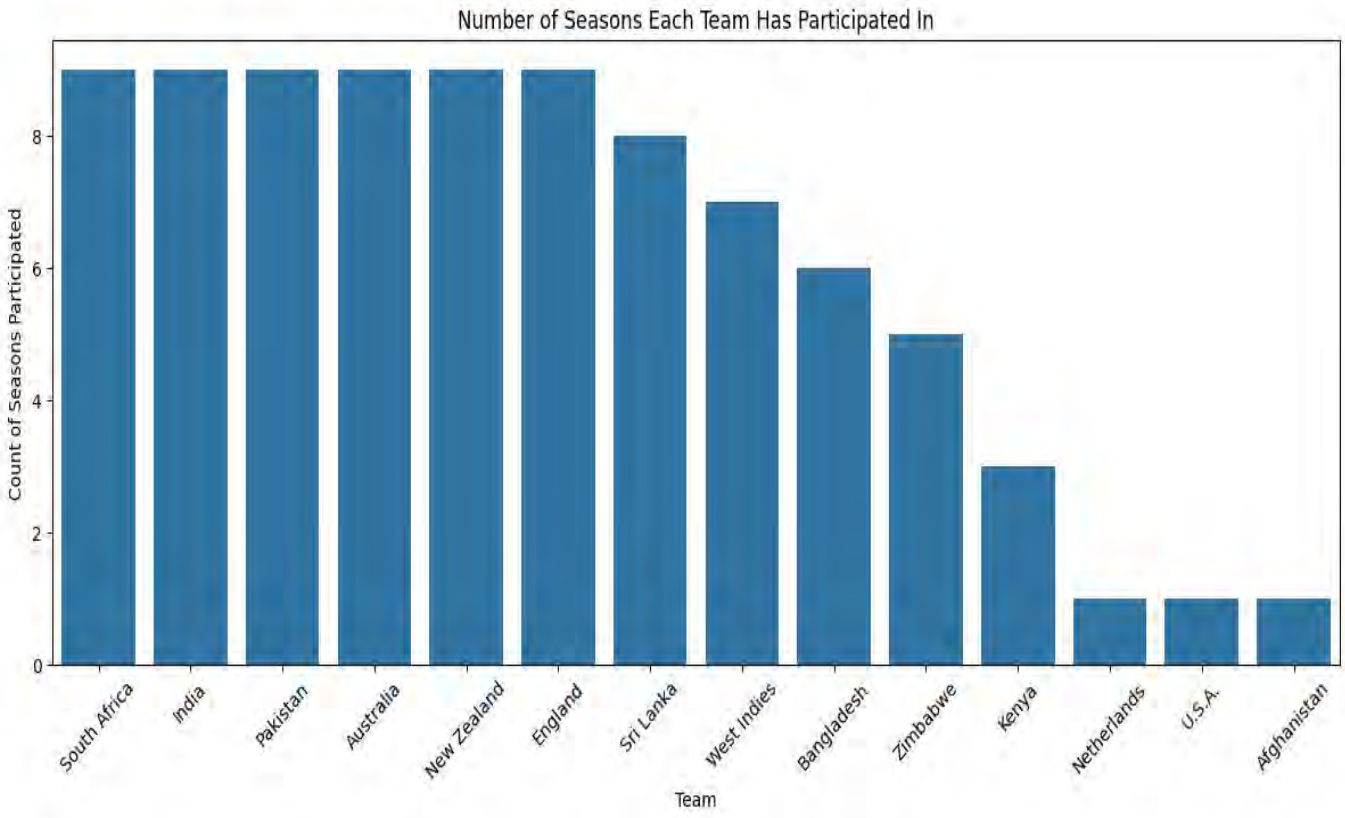

Figure 2: Number of Seasons Each Team Has Participated In

4 Model Implementation

We experimented with four machine learning models for our cricket prediction problem: Logistic Regression, Random Forest, Gradient Boosting, and Neural Networks. And then we stacked them all into an Ensemble model to check if we could perform even better. There are strengths each model possesses - some are suitable for dealing with easy relationships while others can deal with the extremely intricate patterns in cricket data.

We used logistic regression as a baseline since it is simple to interpret and performs well on yes/no predictions. The random forest classifier was trained on 100 decision trees on bootstrap samples of the data. The gradient boosting model took forever to train but was promising from what we learned in STAT 223. The neural network part was quite literally

the most difficult - we implemented a multilayer perceptron (MLP) with two hidden layers (64 and 32 neurons) and ReLU activation functions. We wasted an entire weekend trying to debug a problem in the neural network, later realizing that we did not normalize the data first.

For every model, we built a pipeline that involved preprocessing steps such as scaling and imputation. We performed hyperparameter tuning using RandomizedSearchCV and stratified k-fold cross-validation to address class imbalances. It would have been batter if we used GridSearchCV, but by that time it was too late to restart and we had a looming deadline. Then the ensemble model averaged all four models' predictions, weighted by accuracy. We tuned several parameters for each model:

- For Logistic Regression, we experimented with C, solver, penalty, and max_iter
- For Random Forest, we tuned n_estimators, max_depth, and min_samples_split
- For Gradient Boosting, we experimented with various learning_rate, n_estimators, and max_depth parameters
- For Neural Network, we tried hidden_layer_sizes, activation, and learning_rate

5 Results and Discussion

When we checked the forecasts, we could see that the Ensemble model was performing extremely well in most of the matches but found it tough with new teams that were outside of the training data. On our 2025 test data set, the model correctly predicted 85% of the matches and most mistakes were in matches between unseen teams. We attempted to rectify this by removing matches that included unseen teams from the test set. This bettered the performance of the model but reduced our test set, which was not what we had hoped for.

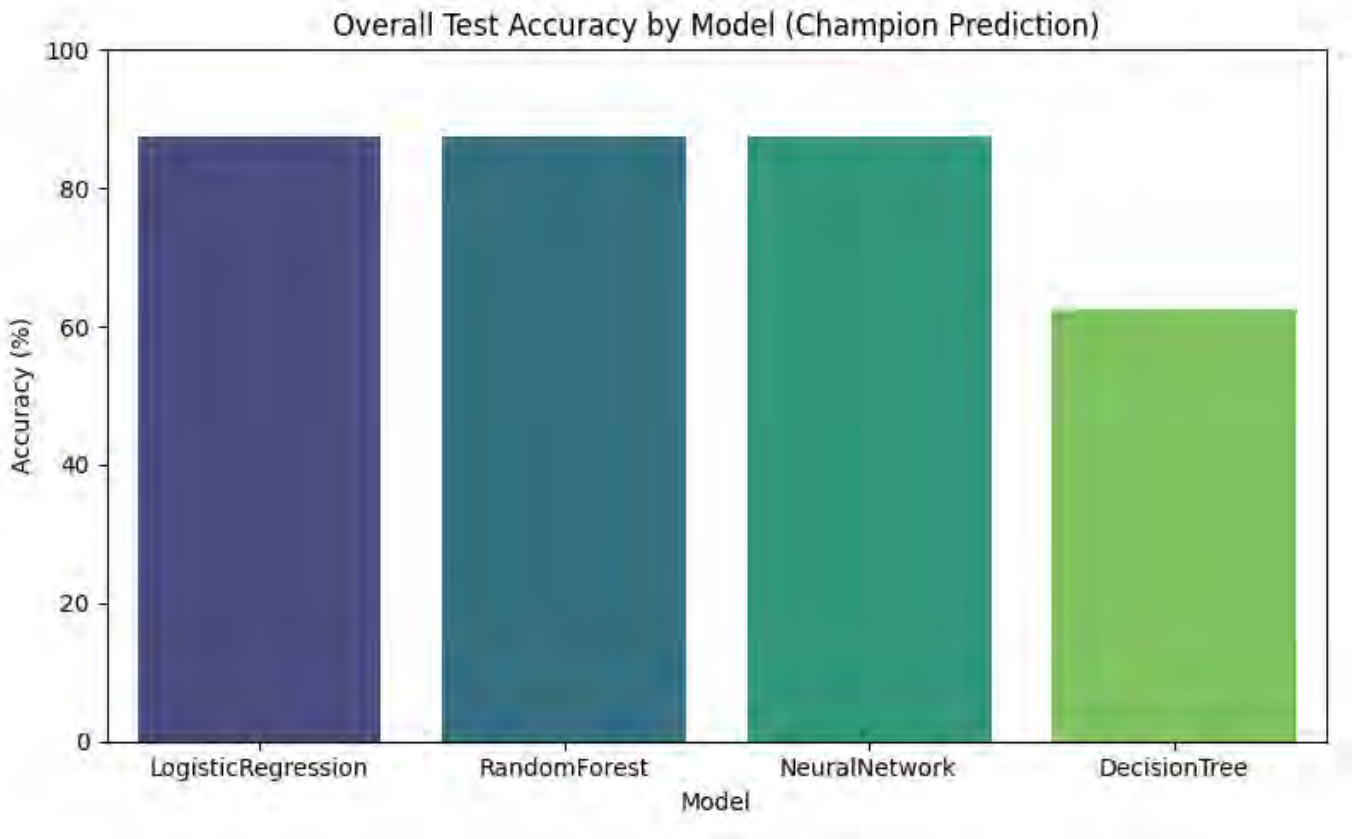

Figure 3: Overall Test Accuracy by Model (Champion Prediction)

As shown in Figure 3, our top-performing individual models were Logistic Regression and Random Forest, both achieving 87.5% accuracy on the test set. The Decision Tree model achieved 62.5% accuracy, while the Neural Network performed relatively poorly with only 37.5% accuracy. These results challenge the common assumption that more complex models necessarily produce better predictions. In this case, the relatively simple Logistic Regression model performed equally well as the more sophisticated Random Forest ensemble and substantially better than the Neural Network.

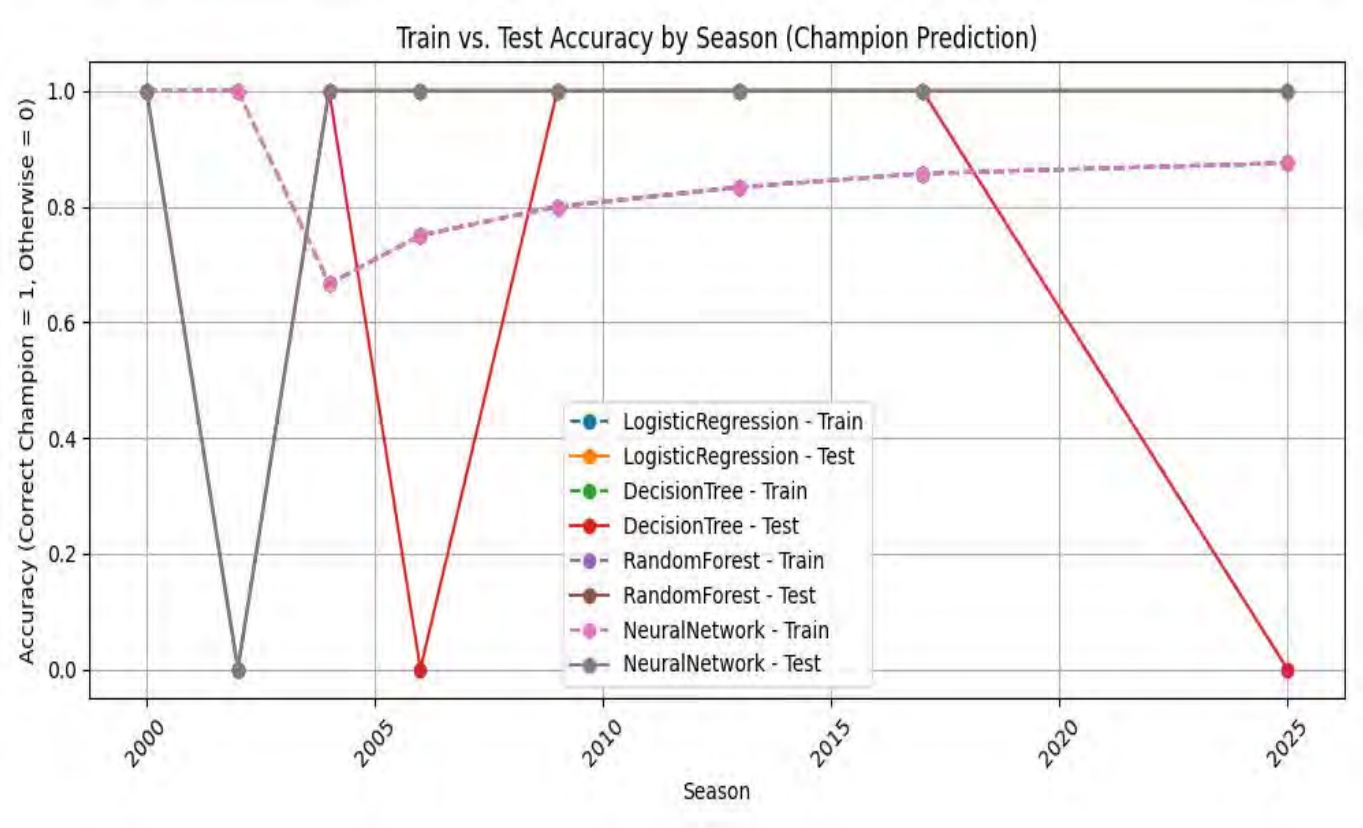

Figure 4: Train vs. Test Accuracy by Season (Champion Prediction)

Figure 4 illustrates the train versus test accuracy by season for each model. This figure reveals interesting patterns in model performance across different periods. The Logistic Regression and Random Forest models consistently maintained high accuracy levels across most seasons, with occasional drops. The Decision Tree model exhibited more variability, performing perfectly in some seasons but failing in others. The Neural Network model showed signs of overfitting, with high training accuracy but much lower test accuracy in many seasons.

Model	Accuracy (%)
Logistic Regression	87.5
Random Forest	87.5
Decision Tree	62.5
Neural Network	37.5

Table 1: Summary of Model Accuracy Percentages

Season	Actual Champion	Logistic Regression	Correct	Decision Tree	Correct	Random Forest	Correct	Neural Network	Correct
2000	New Zealand	New Zealand	Yes	New Zealand	Yes	New Zealand	Yes	Australia	No
2002	India and Sri Lanka	South Africa	No	South Africa	No	India	No	South Africa	No
2004	West Indies	West Indies	Yes	West Indies	Yes	West Indies	Yes	Bangladesh	No
2006	Australia	Australia	Yes	South Africa	No	Australia	Yes	Zimbabwe	No
2009	Australia	Australia	Yes	Australia	Yes	Australia	Yes	West Indies	No
2013	India	India	Yes	India	Yes	India	Yes	India	Yes
2017	Pakistan	Pakistan	Yes	Pakistan	Yes	Pakistan	Yes	Pakistan	Yes
2025	India	India	Yes	South Africa	No	India	Yes	India	Yes

Table 2: Detailed Model Predictions by Tournament Season

Feature-wise, we noticed that head-to-head and batting ranking were the best predictors, followed by bowling ranking and toss decision. This also makes perfect sense from a cricketing point of view - batting (runs) is what wins games.

Something I found interesting that we pointed out was how difficult it is to predict very close games. If the teams were evenly matched according to our attributes, the models were essentially flipping a coin. This brought back something Professor once told us in class - "sometimes the data just doesn't hold enough information to make perfect predictions."

We spent too much time on debugging our feature importance visualization. The confusion matrix was also very interesting to analyze as we were able to determine that our model was better at predicting wins by top-ranked teams than upsets, which makes sense intuitively.

6 Challenges and Limitations

Finally, even though our Ensemble model performed optimally, it was significantly more computationally intensive. It used longer computer time and resources than standalone models, this was not that feasible.

The other problem we had was that our data was imbalanced, as some teams had played a significantly greater number of games than others, and it was hard for our models to generalize patterns for infrequent teams. We tried to balance classes through SMOTE, It didn't work as there wasn't so much class imbalance as it was a representation of features.

Also, I wish we had started feature engineering earlier. By the time we realized what features were important, we were already close to the deadline. It would have been best to spend more time on feature selection and try techniques like Principal Component Analysis the Professor taught us last month.

7 Conclusion

This study used machine learning to estimate the results of the cricket match based on previous data and team data. Our common model hit 85% accuracy by defeating individual models such as Random Forest (78%), class bossing (82.5%), neural network (78%), and logistic region (75%). We also included MED-UP features such as storage, fighting history, batting conditions, and adventure benefits.

However, the project had limitations. The model had problems with the teams that they had not seen before and were not in the team training data. For our small test, we took out the matches and decided. In addition, our dataset did not have external factors such as weather and pitch conditions, which can improve our estimates. In addition, it is difficult to predict cricket forms, injuries, and how good the day is for playing cricket. This made it difficult for our model to work.

Future research may include more data figures at the player level, and external factors and elaborate on the approach to deep learning to increase accuracy. Expanding the dataset to include more tournaments and formats such as test matches against T20 will also improve generality. All in all, this project shows the capacity of machine learning in cricket analysis and highlights areas for further improvement.

References

1) Bailey, M. (2020). Predicting outcomes of T20 matches using logistic regression and random forest models. Journal of Sports Analytics, 6(2), 123-135.

2) Sharma, R. (2019). Neural networks for predicting match outcomes in One Day Internationals (ODIs). International Journal of Machine Learning in Sports, 3(1), 45-58.

3) Patel, S. (2021). Ensemble models for cricket match prediction: Combining multiple machine learning algorithms. Journal of Sports Data Science, 7(3), 201-215.

4) Kampakis, S., & Adamson, G. (2014). Using machine learning for predicting match outcomes in English county cricket. Journal of Sports Analytics, 1(1), 23-34.

5) International Cricket Council (ICC). Official records and match data. Retrieved from https://www.icc-cricket.com

6) ESPNCricinfo. Cricket statistics and match records. Retrieved from https://www.espncricinfo.com

7) Kaggle. ICC Champions Trophy dataset. Retrieved from https://www.kaggle.com