IPL Match Outcome Prediction

Vihan Singh • **Dr. Prince Sharma**

Indian Institute of Information Technology (IIIT) Una — Department of Computer Science Engineering, Data Science

useR! 2026 - Poster

INTRODUCTION

The growing availability of structured sports data has created new opportunities for predictive modelling in cricket analytics. The IPL provides rich match-level and ball-by-ball records capturing team strength, venue behaviour and in-match momentum. This project builds a match outcome predictor by combining linked datasets through extensive feature engineering, a Random Forest classifier and Genetic Algorithm-based feature selection.

Three Linked Datasets

Matches
Venue, teams, toss info, winner & match metadata

Ball-by-Ball
Deliver level runs, wickets & player actions

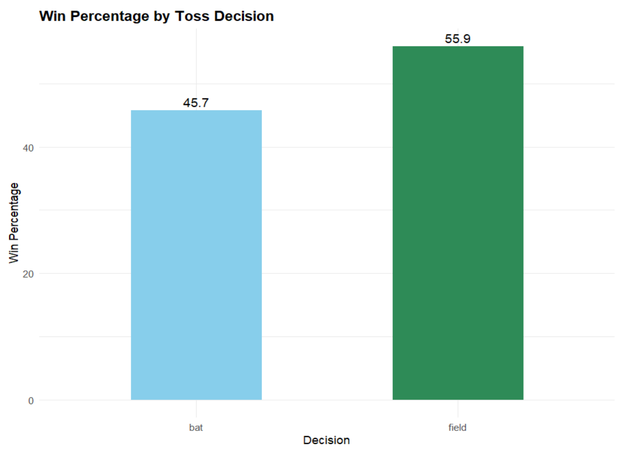
Teams
Team identifiers and names

Hypothesis Testing: Does the Toss Decide the Match?

Before modelling, we tested whether winning the coin toss predicts the match winner using a chi-square test of independence.

1187	$\chi^2 = 1.281$	$p = 0.257$
Matches Analysed	Test Statistic (df = 1)	Fail to Reject H_0 ($\alpha = 0.05$)

Verdict: toss winners won only 52.0% of matches, statistically no better than a coin flip.



Although the coin flip itself doesn't decide the game, the ***decision*** to bat or field shows a real gap:

- **Field first: 55.9% win rate**
- **Bat first: 45.7% win rate**

Interestingly, the predictive model reached a similar conclusion. The toss decision was an important feature, while the toss winner was not

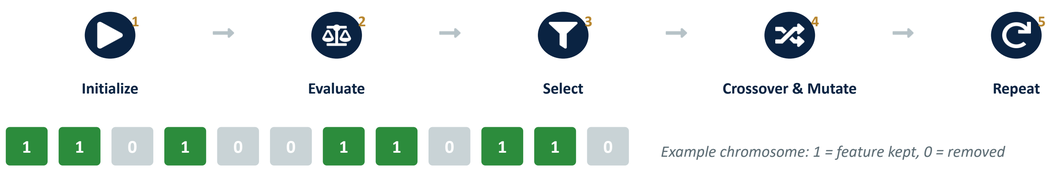
MATERIALS & METHODS

Random Forest (Baseline)

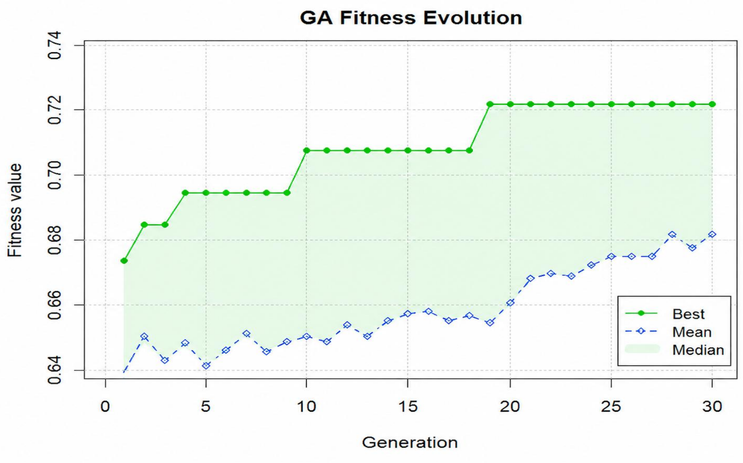
An ensemble of decision trees, each trained on a random subset of data and features. Every tree casts a vote on the outcome; the majority decision becomes the final prediction.

Genetic Algorithm: Feature Selection

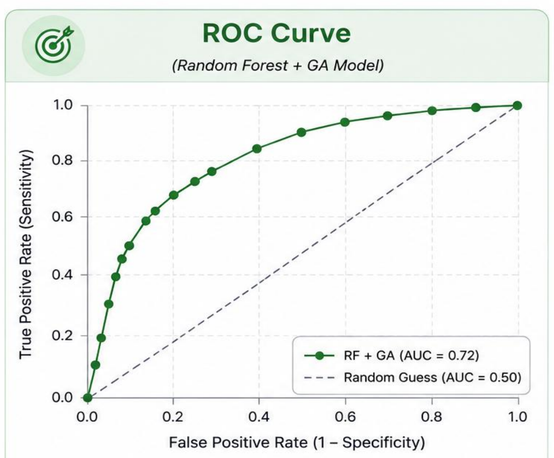
GA searches for the best subset of the 37 engineered features by evolving a population of candidate solutions toward higher predictive accuracy, generation after generation.

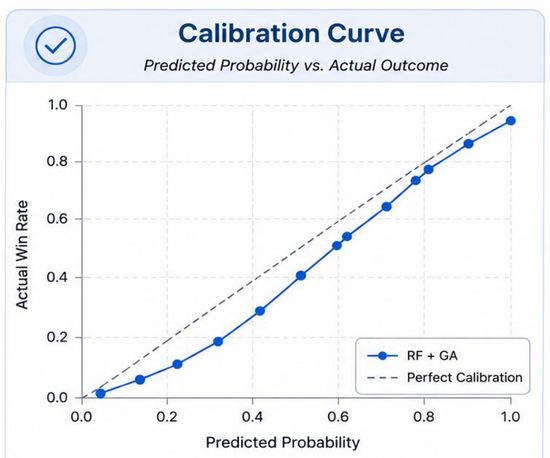


Example chromosome: 1 = feature kept, 0 = removed



Best fitness rises sharply in early generations, then converges by ≈ generation 20.
Extending the run from 30 → 60 generations produced **No further improvement was observed**, confirming the search had converged on the 13-feature subset.





RESULTS

37 → 13	+6.3 pts	+5.2 pts
Features selected by GA (64.9% reduction)	Accuracy improvement 64.3% → 70.6%	Precision improvement 60.0% → 65.5%

Model Performance Comparison

Model	Features	Accuracy	Precision	Recall	F1
Random Forest	37	64.3%	60.0%	78.3%	67.9%
RF + GA	13	70.6%	65.5%	82.6%	73.1%

Confusion Matrix: Team 2 Win Prediction

Random Forest (37 features)

	Actual Lose	Actual Win
Pred. Lose	90	60
Pred. Win	25	63

Random Forest + GA (13 features)

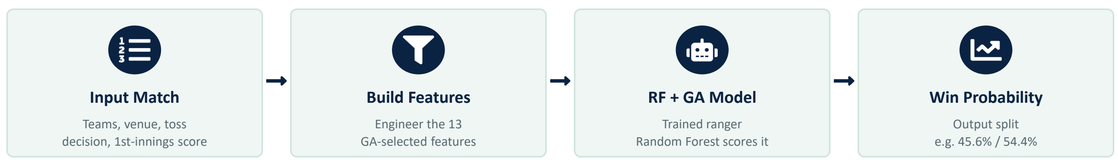
	Actual Lose	Actual Win
Pred. Lose	95	50
Pred. Win	20	73

Correct predictions rose from 153 to 168 matches (out of 238); false negatives fell by 10, meaning *fewer successful chases were missed.*

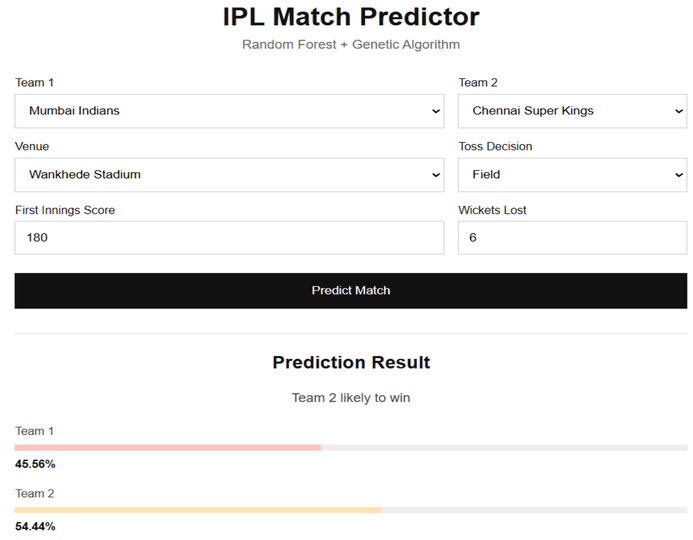
Top Predictive Features (by importance)

1. team2_avg_chase_score
2. score_vs_team1_opponent_avg
3. team2_avg_score_at_venue
4. team1_avg_score_at_venue
5. score_pressure
6. team1_avg_defend_score

How a Prediction Is Made



Live Demo: Deployed Predictor



Model served live via a `plumber` REST API wrapping the trained `ranger` Random Forest model.

Users enter teams, venue, toss decision and first-innings score to get a live win-probability split between sides.

CONCLUSIONS

- **Genetic Algorithm feature selection cut the feature space by 64.9% (37 → 13) while improving every evaluation metric**
- **Accuracy rose from 64.3% to 70.6%, with the largest gains in recall (78.3% → 82.6%), fewer missed chases**
- **Toss decision (bat/field), not toss outcome, proved to be the meaningful signal, validated via hypothesis testing before being engineered as a feature**
- **The final 13-feature model is simpler, faster to train, and deployed as a live R-based prediction API**

REFERENCES

- Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5–32.
- Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*.
- Cricsheet. *IPL Match and Ball-by-Ball Dataset*. https://cricsheet.org/

ACKNOWLEDGEMENTS

The author sincerely thanks Dr. Prince Sharma for his mentorship and guidance throughout this research. Appreciation is extended to the developers and maintainers of the R ecosystem, including dplyr, ranger, GA, ggplot2, and plumber, for providing the open-source tools used in this work. AI Agent assistance was used for implementation guidance, deployment support, API integration, and debugging