Cricket Match Analytics + Prediction using ML

Abstract

- Goal: Predict cricket match outcomes during the **2nd innings**.

- Features: target, runs left, wickets, player-specific metrics.

- Algorithms: Random Forest, SVM, Logistic Regression, Naïve Bayes.

- Innovation: **Player Consistency formula** (mix of stats + dynamic ratings).

- Contribution: Review + predictive model guidance for future cricket analytics.

Introduction

- Cricket's popularity + data-rich nature → great for ML.

- ML can analyze player stats, pitch/weather, team history.

- **Deep learning** can analyze ball trajectory, biometrics, playing styles.

- Fantasy cricket apps (Dream11) = real-world testbed for predictive models.

- Potential future: injury prediction, real-time predictions, wearable integration.

Literature Review (Highlights)

- **Tree-based models (Random Forest, Gradient Boosted Trees)** generally outperform statistical/probabilistic models.

- Features used: toss, venue, form, pitch, weather, social media sentiment.

- Studies range from predicting match winners to player performance (batting runs, wickets).

- Random Forest often reported **>85–95% accuracy** across studies.

- Research gaps:

    1. Focus on single league (mostly IPL/ODI).

    2. Missing external factors (weather, injuries, toss impact).

    3. No comparison with betting odds or expert predictions.

    4. Limited interpretability of chosen features.

Methodology

1. **Data Sources**

    - Player data → scraped via Selenium from **iplt20.com**.

    - Match data → JSON from **cricsheet.org**.

2. **Data Processing**

    - Compiled multi-year player stats into single profiles.

    - Filtered irrelevant match attributes.

3. **Feature Engineering**

    - Developed **Player Rating Formula** (batting SR, bowling SR, high scores, wickets, etc.).

    - Used **deep learning** to optimize feature weights.

    - Added **Player Consistency metric**.

4. **Feature Selection**

    - Correlation-based filtering + Chi-squared test.

    - Identified most influential attributes.

5. **Models Used**

- ○ Random Forest 🌟

- ○ SVM

- ○ Logistic Regression

- ○ Naïve Bayes

Conclusion

- Random Forest proved to be **most reliable model**.

- Strong predictive power for real-world cricket match analytics.

- Potential use in strategy, team selection, fan engagement.

Future Work

- Refine **Player Consistency** metric.

- Add more player-centric metrics.

- Explore **deep learning + ensemble methods**.

- Collaboration with cricket experts for interpretability.

- Real-time dynamic prediction (wearables, live data).