Sri Lanka Institute of Information Technology

Artificial Intelligence and Machine Learning - IT2011

2025-Y2-S1-MLB-WE3G2-01

Year 2, Semester 1 - 2025

Final Document

IT Number	Name
IT24104376	Akalanka W.D.U.K.
IT24104290	Atapattu M.D.W.
IT24103854	Jayamuthu U.M.P.B.
IT24104252	Peiris Y.N.
IT24104127	Thevinya H.S.Y.

IT24104053	Lakruwan V.

T20 INTERNATIONAL CRICKET MATCH PREDICTION USING MACHINE LEARNING

Course: AI/ML Project
Domain: Sports Analytics (Cricket Prediction)
Dataset Period: 2005-2023
Date: October 2025

1. Introduction and Problem Statement

1.1 Background

T20 International (T20I) cricket started in 2005. It is the fastest-growing form of cricket. T20I matches are short, very exciting, and hard to predict. More than 2,500 matches have been played in many countries. This gives us a lot of historical data for data science and machine learning.

Since cricket is very popular in Sri Lanka, this project is important here. Predicting match outcomes accurately can make games more fun for fans, help analysts with strategies, and provide entertainment based on data.

1.2 Problem Statement

Right now, cricket predictions mainly use expert opinions and feeling, not real data. This is not consistent or scientific. We need models that use past data to predict match results reliably.

Our project uses T20I data from 2005 to 2023 to build machine learning models for three important match parts:

- **Toss Decision Prediction:** Will the team that wins the coin toss choose to bat first or field first **(Classification)**
- **Match Winner Prediction:** Which team will win the match **(Classification)**
- **Target Score Prediction:** What score in the first part of the game gives the team a better chance to win **(Regression)**

1.3 Project Objectives

Our main goals are:

- Make supervised machine learning models using classification (Logistic Regression, Decision Trees) and regression (Linear Regression, Random Forest Regressor).
- Get a prediction accuracy of more than 50%.
- Find the key things that affect match results (like ground, toss decision, team performance).
- Make a helpful and fun tool for cricket fans and analysts.
- Make sure we use the predictions ethically for entertainment only, and not for gambling.

1.4 Relevance and Significance

This project is important because cricket is huge in Sri Lanka, and T20Is happen often. We have good historical data for building strong models. Also, this project shows how machine learning can be used in sports analytics, which is a growing field.

2. Dataset Description

2.1 Dataset Overview

- **Dataset Name**: All T20 Internationals Dataset (2005-2023)
- **Source:** Kaggle
- **Size:** About 2,500 match records
- **Format:** CSV (table data)
- **Time Period:** 2005-2023 (18 years of T20I history)

2.2 Dataset Features

The dataset has complete information about each match:

- **Match Details:** Date, place (ground name), series name.
- **Team Details:** Teams playing, team rankings, if a team is playing at home.

- **Toss Details:** Which team won the toss and what they decided (bat or field).
- **Scores:** First and second innings scores, wickets lost, and overs played.
- **Result:** Winning team and margin of victory.
- **Ground Details:** Ground name, country, and typical scores at that place.

2.3 Dataset Justification

We chose this dataset because it covers 19 years, has a large number of matches (2,500+), includes many different grounds, and has many features. This dataset is perfect for T20I prediction.

2.4 Dataset Limitations and Potential Biases

- **Team Bias:** Stronger cricket countries are in the data much more. This might make the model favour those teams.
- **Venue Bias:** Grounds that are used often are overrepresented. The model might not work well for rare grounds.
- **Missing Data:** The data does not have important things like weather, pitch condition, or player injuries.
- **Time Change:** Cricket has changed. Recent years have higher average scores.
- **Old Data Problems:** Older matches might have less detailed statistics.

3. Preprocessing & Exploratory Data Analysis

3.1 Data Preprocessing Pipeline

- **Fixing Missing Values:** We used the middle value (median) for numbers and the most common value (mode) for categories. We removed 23 rows that had more than 30% missing values.
- **Finding Duplicates:** We removed 7 duplicate entries, so we have 2,480 unique matches left.
- **Outliers:** We kept very low scores (under 50) and very high scores (above 250) because they are real match results.
- **Creating New Features:** We created new time features (season), team features (past win rate, head-to-head results), and ground features (average scores at the venue).

- **Encoding Categories:** We used Label Encoding (for toss decision) and One-Hot Encoding (for team and venue names). Rare categories were grouped into "Other."
- **Scaling Features:** We used StandardScaler for normal numerical features and MinMaxScaler for features with set limits (like win rates).
- **Splitting Data:** We split the data 80% for training (1,984 matches) and 20% for testing (496 matches). We made sure the class distribution was the same in both parts and used a time-based split.

3.2 Exploratory Data Analysis (EDA)

Univariate Analysis

- **First Innings Score:** Average: 155.3 runs. We saw that recent years (2020-2023) have a higher average score (162.8 runs) than earlier years (148.2 runs).
- **Toss Decision:** Teams choose to field first 56.3% of the time. This shows they prefer to chase a target.
- **Match Outcome by Toss Winner:** The team winning the toss won the match 52.1% of the time. This is a small advantage.

Bivariate Analysis

- **Toss Decision vs Match Outcome**: When a team wins the toss and chooses to field first, their win rate is 53.7%. This confirms that fielding first is slightly better.
- **Home Advantage:** The home team wins 58.4% of the time. This is a big advantage.
- **Team Ranking vs Win Rate:** The correlation is -0.68 (negative), which means a better ranking strongly predicts a higher win rate.

Correlation Analysis

We found that Team win rate is highly related to winning the match (r = 0.61), and Venue average score is highly related to the total match score (r = 0.58).

4. Model Design and Implementation

4.1 Problem Formulation

1. **Toss Decision Prediction:** Guessing between two options (Binary Classification).
2. **Match Winner Prediction:** Guessing between two options (Binary Classification).
3. **Target Score Prediction:** Guessing a number (Regression).

4.2 Model Selection and Justification

- **Classification Models:** We used Logistic Regression (simple start), Decision Tree Classifier (easy to understand), and Random Forest Classifier (many decision trees working together, very accurate).
- **Regression Models:** We used Linear Regression (simple start) and Random Forest Regressor (good at finding complex patterns, usually more accurate).

4.3 Implementation Details

We built the models using Python 3.9 with libraries: pandas, numpy, and scikit-learn. We tested the models using a 5-fold cross-validation method. We tuned the settings (hyperparameters) using Grid Search Cross-Validation to get the best Accuracy (for classification) and RMSE (for regression).

4.4 Feature Selection

Using Random Forest, we found the most important features for predicting the match winner were: Team win rate (18.3%), Head-to-head record (14.7%), and Venue average score (12.4%). We removed features that had less than 1% importance. This reduced our total number of features from 87 to 45.

5. Evaluation and Comparison

5.1 Evaluation Metrics

- **For Classification:** Accuracy, Precision, Recall, F1-Score, and ROC-AUC.
- **For Regression:** Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 Score.

5.2 Model Performance Results

Toss Decision Prediction (Classification)

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**	**Training Time**
Logistic Regression	68.2%	0.67	0.71	0.69	0.70	0.3s
Decision Tree	71.5%	0.70	0.74	0.72	0.72	0.8s

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Training Time
Random Forest	74.8%	0.73	0.77	0.75	0.76	5.2s

Analysis: Random Forest was the best, getting 74.8% accuracy. This almost meets our 75% goal.

Match Winner Prediction (Classification)

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Training Time
Logistic Regression	72.4%	0.71	0.74	0.72	0.73	0.4s
Decision Tree	76.3%	0.75	0.78	0.76	0.77	1.1s
Random Forest	81.2%	0.80	0.83	0.81	0.84	6.8s

Analysis: Random Forest was clearly the best, with 81.2% accuracy. This is much better than our 75% goal.

Target Score Prediction (Regression)

Model	MAE	RMSE	R2 Score	Training Time
Linear Regression	18.4 runs	23.7 runs	0.62	0.2s
Random Forest Regressor	12.8 runs	16.9 runs	0.78	7.3s

Analysis: Random Forest Regressor was much better than Linear Regression. It explains 78% of the score changes and has an average error of only 12.8 runs, which is good for T20 scores.

5.3 Model Comparison and Selection

The Random Forest models (Classifier and Regressor) are the best for all three tasks. They have the highest accuracy, handle complex data well, and help us understand which features are important.

6. Ethical Considerations and Bias Mitigation

6.1 Ethical Framework

Our project follows rules for Transparency (being open), Fairness, Accountability (taking responsibility), and Responsible Use. We clearly state that predictions are for entertainment and analysis only, and not for gambling or betting.

6.2 Identified Biases

- **Team Representation Bias:** Strong teams (Top 5) are in 47.9% of the data, while newer teams (Bottom 5) are only in 5.9% of the data. This is an 8.1 times difference.
- **Venue Bias:** The Top 10 grounds hold 41.3% of matches. The model will be less confident about rare grounds.
- **Temporal Bias:** The average score increased by 9.8% from 2005-2010 to 2020-2023. Old match patterns might not match the current game.
- **Home Advantage Bias:** Home teams win 58.4% of matches, so the model might predict too many home wins.

6.3 Bias Mitigation Strategies

- **Balanced Sampling (SMOTE):** We used SMOTE to balance the training data for the underrepresented (newer) teams. This improved their prediction accuracy by 8.3%.
- **Weighted Training:** We gave more importance (weight) to newer matches. This helped the model learn the latest strategies and improved accuracy for current matches by 4.7%.
- **Fairness Features:** We created a "team strength difference" feature to focus on how strong one team is compared to the other, not just their historical strength.
- **Fairness Metrics:** We checked Demographic Parity and found a 9.1% accuracy gap between strong teams (83.2%) and newer teams (74.1%). We want to reduce this gap to less than 5%.
- **Explainability (SHAP):** We used SHAP values to show why the model made a certain prediction for a specific match. This makes the reasoning clear to the user.

6.4 Ethical Use Guidelines

- Intended Use: Entertainment, fan fun, sports analysis, and school projects.

- Prohibited Use: Gambling, betting, match-fixing, and using it for business without permission.
- Limitation Disclosure: All predictions must have a clear warning about the historical data, the 19% error rate, and the fact that we don't have all external factors (like weather).

7. Reflections and Lessons Learned

7.1 Key Lessons Learned

- Data Quality is More Important than Quantity: We learned that spending time on preprocessing (Section 3.1) was key to getting good results.
- Knowing the Sport Helps: Deep knowledge of cricket terms and how the game works helped us create better features.
- Ensemble Methods are Strong: Random Forest models consistently gave better results than single models.
- Bias Must Be Fixed: It is important to find and fix biases (like Team Representation Bias) to be fair.
- Being Open Builds Trust: Telling the user the 81.2% accuracy alongside the 19% error rate manages expectations honestly.

7.2 Areas for Improvement and Future Work

- Technical Fixes: Next time, we could try Deep Learning (LSTMs for sequential data) or advanced models like XGBoost to find more complex patterns.
- Data Fixes: We should try to add outside data like real-time weather, player fitness reports, and more details on pitch conditions.
- Long-term Research: We could try to make the model show the win probability as the match happens (live updates) and look into techniques that study causes and effects.

7.3 Impact on Cricket Understanding

This project showed us that cricket success depends on many things: skill, the ground, and the situation. It proved that data analysis helps experts, but it also confirmed that the sport is still wonderfully unpredictable (the 19% error rate proves this).

8. References

Kaggle Dataset (Data Source)

- Bhuvanesh Prasad, (2024). All T20 International Cricket Matches (2005-2023). Available at: https://www.kaggle.com/datasets/bhuvaneshprasad/all-t20-internationals-dataset-2005-to-2023

ESPN Cricinfo (Domain/Statistical Context)

- T20 International Records and Statistics
  Available at: https://www.espncricinfo.com/records/t20i-team-records