Predicting Game Outcomes Using Machine Learning in Sports Analytics

Emerson Hernandez Toufeeq Sharieff Mani Vutukuri
Michael Nguyen

April 6, 2025
Professor: Professor Fang Yi Yu

Contents

Abstract

This project aims to predict the outcome of sports matches (win, loss, or draw) by analyzing historical game data. We build a machine learning pipeline that leverages key performance metrics to generate reliable predictions. This work is intended to provide actionable insights for coaches, enhance fan engagement, and support data-driven decision making.

1 Introduction

1.1 Problem Statement

The goal of this project is to develop a predictive model that forecasts match outcomes by analyzing historical game data. By using features such as goals scored, possession percentages, and recent form, our model seeks to reliably predict whether a match will result in a win, loss, or draw.

1.2 Motivation

- **Strategic Insights:** Coaches and managers can adjust their game plans and predict opponents' tactics.
- **Fan Engagement:** Accurate predictions create a more engaging experience for fans, especially in fantasy leagues and sports betting.
- **Data-Driven Decision Making:** Using machine learning provides objective insights, reducing guesswork.

1.3 Proposed Methods

Our methodology involves the following steps:

1. **Data Collection and Preprocessing:** Use a Kaggle-sourced dataset, clean and preprocess the data, and perform exploratory data analysis (EDA).
2. **Feature Engineering:** Create new features such as goal difference and rolling averages of performance metrics.
3. **Model Development:** Implement baseline models including Logistic Regression, Decision Trees, and K-Nearest Neighbors.
4. **Model Evaluation:** Evaluate models using accuracy, precision, recall, F1-score, and confusion matrices.

2 Related Work

2.1 Key Studies

- **NBA Game Outcome Prediction:** Previous research has demonstrated that decision trees and logistic regression can predict outcomes with high accuracy by using features like field goal percentages and rebounds.

- **Soccer Match Prediction:** Studies using historical match data (goals scored, possession, etc.) have shown significant predictive accuracy for soccer outcomes.

These studies underscore the potential of machine learning models in sports analytics and highlight the importance of careful feature selection and model tuning.

3 Data and Methodology

3.1 Dataset Description

The dataset, obtained from Kaggle, contains historical match data with features such as goals scored, goals conceded, possession percentage, number of games won, and recent form. The target variable is the match outcome (win, loss, or draw).

3.2 Methodology

1. **Exploratory Data Analysis:** Summary statistics and visualizations (histograms, correlation matrices) to understand the data.
2. **Preprocessing:** Handling missing values, encoding categorical variables, and normalizing numerical features.
3. **Modeling:** Implementation of Logistic Regression, Decision Trees, and K-Nearest Neighbors. Hyperparameter tuning is performed to optimize model performance.
4. **Evaluation:** Models are evaluated using validation metrics and the best model is tested on a hold-out test set.

4 Results and Discussion

The best performing model is identified based on evaluation metrics such as accuracy and F1-score. A detailed analysis of feature importance and model performance is provided, highlighting the strengths and limitations of our approach.

5 Conclusion and Future Work

This project demonstrates the viability of using machine learning to predict sports match outcomes. Future work may include refining the model with additional features, exploring ensemble methods, and extending the analysis to different sports leagues.