Match Outcome Prediction Serie A

Ohad Krispin , Nadav Shamir

August 9, 2025

Abstract

In this project, we use machine learning to predict the results of football matches in the Italian Serie A from the 2020–2025 seasons. The goal is to classify each match as a home win, away win, or draw. We used post match stats such as goals for (GF), goals against (GA), and expected goals (xG), we also engineered historical features like recent team form and Elo rating system to better capture team level over time. After cleaning and preparing the data, we tested three models: Naive Bayes, Gradient Boosting, and SVM. We compared them using accuracy, precision, recall, and weighted F1 score with cross validation and a separate test set. Gradient Boosting achieved the best performance, with a weighted F1 score of about 0.6 in cross validation and 0.58 accuracy on the test set. While these results show that machine learning can provide useful predictions, predicting draws remains particularly challenging, which limits overall predictive performance.

1 Introduction

Football is the most popular sport in the world, attracting millions of fans and generating huge financial interest. Predicting match outcomes is very important for coaches, analysts, and betting companies, as it can help in making tactical decisions, understanding team performance, and improving fan engagement. However, football is highly unpredictable due to many factors, such as team form, player injuries, random events, and outcomes during the match, making accurate prediction a challenging task. Traditional prediction methods often rely on expert opinion or basic statistics, but recent research shows that machine learning can find deeper patterns by combining multiple features, such as historical team performance, head to head records, and new advanced statistics like expected goals (xG). Some studies have used logistic regression or simple rating systems, while others explored tree based models and neural networks, showing that data driven methods can outperform manual predictions. In this work, we focus on predicting Italian Serie A matches from the 2020–2025 seasons. We use post match statistics (goals, xG, shots) and engineered historical features such as recent form and Elo ratings to improve predictions. We compare three machine learning models—Naive Bayes, Gradient Boosting, and Support Vector Machine (SVM) to identify which performs best for this task.

2 Related Work

Several recent studies have applied machine learning to pre match football prediction. Below we summarize three representative works and highlight how they relate to our work.

2.1 Rodrigues & Pinto (2022) — "Prediction of Football Match Results with Machine Learning"

- **Dataset:** 1900 Premier League matches from 2013/14 to 2018/19.
- **Features:** Post-match box-score counts (goals, shots, corners, fouls, yellow/red cards), betting odds, referee, plus static player and team ratings (from sofifa.com).
- **Feature Engineering:**
  - Averages of goals conceded at home/away and cumulative home/away wins.
  - Pre-match expanding and rolling averages of goals, shots, cards..., computed separately for home and away.
  - Correlation analysis to drop highly redundant variables, followed by Boruta and RFE to select a final subset.
- **Models:** Naïve Bayes, KNN, Random Forest, SVM, C5.0, XGBoost, Multinomial Logistic Regression, and ANN.
- **Results:** The classifier outcome task was to predict Home team win, Draw , Away team win. The top Random Forest model reached about 65.3% accuracy.

- **Comparison:** They rely heavily on raw post match stats and odds. We build on this by incorporating expected goals metrics (xG/xGA), rest days and fatigue indicators, head to head form, and an online Elo rating to better capture latent team strength before each fixture.

2.2 Capobianco et al. (2019) — "Can Machine Learning Predict Soccer Match Results?"

- **Dataset:** 378 Serie A matches (2017–18) with 98 attributes from match reports.
- **Key Features:**
  - Possession splits, pitch area, and team center of gravity (attacking/defensive).
  - Goals scored/conceded and other team specific indicators.
  - Best first search + PCA reduced attributes to 20 (player distances, team speed, ball recoveries).
- **Tasks:** Binary *match result* (win/lose) and *goal count in wins* (< 2 vs. ≥ 2).
- **Models:** J48, SMO (SVM), RepTree, RandomTree, RandomForest, MLP.
- **Results:**
  - Result prediction: RandomForest precision=0.857, recall=0.750.
  - Goal prediction: RandomForest precision=0.879 (< 2), 0.800 (≥ 2); avg. precision=0.862, recall=0.868.
- **Comparison:** Similar to our work, the main task was outcome classification; however, they used in play spatial/possession data, while we rely only on post match stats enriched with xG, Elo, and rest fatigue indicators.

2.3 Zaveri et al. (2018) — "Prediction of Football Match Score and Decision Making Process"

- **Dataset:** Spanish La Liga matches from five seasons (2012–2017), combining:
  - *Match History:* 12 attributes (shots, shots on target, corners, yellow/red cards).
  - *Team vs. Team:* Head to head records.
  - *Goals History:* Goals, shots, and shots on target across seasons.
  - *Player/Team Stats:* FIFA 18 player attributes ,tactical ratings.
- **Feature Engineering:**
  - Team IDs (1-29, ranked by performance).
  - Home win percentage vs. specific opponent.
- **Tasks:** Multiclass outcome (Home Win, Away Win, Draw) plus secondary predictions (scoreline, probable scorers, team selection).
- **Models:** Logistic Regression, Random Forest, ANN, Linear SVM, Naïve Bayes.
- **Results:**
  - Match History only: best Logistic Regression accuracy = 63.9%.
  - Head to head feature: Logistic Regression = 71.6% (best), RF = 69.9%, ANN = 69.2%.
  - Exact score (Goals History): Logistic Regression = 69.9%.
- **Comparison:** Similar to our work, the main task was match outcome classification (w,d,l), using ranking system to the teams like our elo system; however, they extended it with tactical suggestions and scoreline prediction.

3 Method

For our database, we chose the Serie A Matches Dataset (2020–2025), compiled via web scraping by Mr. Marcel Biezunski from the Poznan University of Technology in Poland. The dataset contains 3,800 match records, each described by 29 attributes. The label distribution is relatively balanced: 1,386 home wins (34.5%), 1,386 away wins (34.5%), and 1,028 draws (27%), ensuring fair training and a reliable evaluation of our models.

3.1 Pre-processing

We started by removing attributes irrelevant to our predicting task, specifically: `Notes`, `Match Report`, `Comp`, `Captain`, `Referee`, and `Round`.

The missing values were then addressed. The attribute `Attendance` had 684 missing entries (18%), which we filled in using the mean attendance for that team and the venue (home or away). If such information was unavailable, we used the global mean. The other missing attributes (`xG`, `xGA`, and `FK`, each missing twice, and `Dist` missing four times) were imputed with their respective column medians to mitigate the effect of outliers.

Finally, the data was reformatted into a ML structure. The dates were converted into standard datetime format, sorted chronologically per team, and encoded into new derived features such as Month and a three stage SeasonStage (early, mid, late). Strings were converted into numerical or categorical encodings, such as binary encoding for Venue (home = 1, away = 0) and one-hot encoding for Team and Opponent. Formation strings (like "4-3-3") were split into five numeric positional features per side.

3.2 Feature Engineering

All engineered features were based solely on information available before kickoff to prevent data leakage. Rolling averages were computed for key performance metrics (xG, xGA, SoT, Sh, Poss, Dist, GF, GA), shifted by one match to ensure causality. Team form was captured through a five match rolling win rate and explicit win/draw/loss streak counters.

Rest and fatigue were modeled via RestDays (capped at 14) and the number of matches played in the previous 10 days. An Elo rating system was implemented for each team, updated match by match, producing both raw Elo differences (Elo_Diff) and their five match rolling means.

Head to head performance indicators were calculated as expanding averages for each team opponent pair. Additionally, we engineered features to capture venue specific advantages (e.g., differences between a team's home averages and its overall averages for xG and GA). Interaction terms (such as Elo difference multiplied by rest days) and goal related features (e.g., total goals rolling frequencies, and attacking/defensive trend indicators) were also included.

In total, after preprocessing and feature engineering, we retained and constructed 21 additional attributes beyond the original dataset.

3.3 Feature Selection

To identify the most informative variables, a Random Forest classifier was initially trained on the complete feature set with balanced class weights to account for class imbalance. Each feature's importance was extracted from the trained model and ranked accordingly. The top 20 features were selected based on their contribution to predictive performance. This selection process aimed to improve model generalization, reduce overfitting, and enhance interpretability by eliminating redundant or less relevant variables.

3.4 Models

We used three different machine learning models. Two of them which we learned in class - Support Vector Machine (SVM) and Naive Bayes, while the third which we studied on our own - Gradient Boosting.

SVM: SVM finds the optimal boundary that maximizes the margin between classes. It focuses on support vectors and can model non-linear patterns using kernels. We selected it for its robustness in high-dimensional and well-separated data.

Naive Bayes: Naive Bayes is a probabilistic classifier based on Bayes' Theorem, assuming feature independence. It is fast, interpretable, and performs well on small or text-based datasets, making it a strong baseline.

Gradient Boosting: Gradient Boosting builds decision trees one after another, where each new tree tries to fix the mistakes of the previous ones. It focuses more on the difficult cases and slowly improves the predictions. Unlike Random Forest, where all trees are built independently, Gradient Boosting adds trees step by step and uses a method called gradient descent to reduce errors. We chose it because it usually works very well on structured data like this football dataset.

4 Evaluation

4.1 Experimental Setup

The dataset was split into an 80/20 stratified train/test partition, maintaining class balance across splits. All numeric features were standardized to zero mean and unit variance. The three chosen models SVM, Naive Bayes, and Gradient Boosting were configured with balanced class weights where applicable to account for the equal importance of win, draw, and loss outcomes.

4.2 Cross-Validation and Hyperparameter Tuning

We used 5-fold stratified cross-validation on the training set to check how stable the models are, measuring accuracy and weighted F1. For tuning, we used 3-fold GridSearchCV. For SVM, we tested different kernels, C values, and gamma. For Gradient Boosting, we tried different numbers of trees, learning rates, and depths.

4.3 Final Test Evaluation

After tuning, each model was retrained on the entire training set and evaluated on the hold out test set. We reported Accuracy, weighted Precision, Recall, and F1 scores. Additionally, a confusion matrix and full classification report were generated to better understand the misclassification patterns and overall predictive reliability of each model.

5 Results

5.1 Cross-Validation Performance

Table 1: Cross-Validation Performance (Training Set)

Model	F1 (± std)	Acc (± std)
Gradient Boosting	**0.609 ± 0.017**	**0.611 ± 0.016**
SVM	0.506 ± 0.018	0.507 ± 0.018
Naive Bayes	0.456 ± 0.016	0.493 ± 0.015

The best hyperparameters after tuning were: **SVM** – C = 1, gamma = auto, kernel = rbf; **Gradient Boosting** – learning rate = 0.05, max depth = 3, n estimators = 100; **Naive Bayes** – default settings.

5.2 Test Set Performance

Table 2: Test Set Performance

Model	Acc	Prec	Rec	F1
Gradient Boosting	**0.589**	**0.591**	**0.589**	**0.586**
SVM	0.510	0.508	0.510	0.509
Naive Bayes	0.494	0.479	0.494	0.456

5.3 Confusion Matrices

Table 3: Confusion Matrices (0 = Loss, 1 = Draw, 2 = Win)

Gradient Boosting			SVM			Naive Bayes		
151	59	64	152	57	65	197	18	59
25	153	27	63	80	62	101	24	80
84	51	141	67	56	153	108	16	152

5.4 Result Analysis

Gradient Boosting gave the best results, with the highest F1 score (0.586) and accuracy (0.583). This makes sense because it can learn complex interactions between features, which is important in football where many factors (form, rest days, Elo ratings, head-to-head stats) interact and influence outcomes. Its sequential learning also allows it to focus on harder to predict matches, improving balance across classes.

SVM performed moderately well (F1 = 0.509). It was good at predicting draws but struggled to clearly separate wins and losses. This is likely because SVM, can have difficulty modeling non linear relationships in multi class tabular data.

Naive Bayes performed worst (F1 = 0.456). This was expected because it assumes all features are independent, which is unrealistic for football data (for example, shots, possession, and xG are strongly correlated). Still, it did reasonably well at predicting losses, likely because a few strong features (like low Elo or poor recent form) are reliable indicators for that class.

We used F1 score as the main metric because it balances precision and recall, which is important when missclassifying a win as a draw or the opposite has similar consequences. Accuracy is shown for completeness, but F1 is preferred because it better reflects balanced performance when class distributions are slightly uneven.

6 Conclusions and Future Work

6.1 Conclusions

This work addressed the task of predicting football match outcomes (win, draw, loss) as a multiclass classification problem. Gradient Boosting achieved the best results (F1 = 0.586), likely because it can capture complex, non linear relationships, which are crucial in football where many factors interact simultaneously.

Predicting football outcomes remains challenging because the game is influenced by many unpredictable elements, such as individual player performance, tactical changes during the match, fatigue of players, and random events like penalties or red cards. While the models performed reasonably well, some classes were harder to predict than others.

We found that multiclass prediction in football is particularly difficult because, unlike sports such as basketball where scoring (points) is frequent and performance is more consistent, football results can be decided by a single goal or moment. A stronger team can still lose due to one mistake or unexpected event, which makes patterns less stable and harder for models to learn reliably.

6.2 Future Work

Future work we should focus on improving win prediction and overall performance. This could be done by using more advanced versions of Gradient Boosting (XGBoost), better hyperparameter fine tuning, and, most importantly, improved feature engineering and adding external features like injuries, possession, player stats, or match importance. Ensemble methods or time series approaches may also help capture long term trends and improve classification balance.

REFERENCES

[1] Fátima Rodrigues and Ângelo Pinto. 2022. Prediction of Football Match Results with Machine Learning. In *International Conference on Industry Sciences and Computer Science Innovation*. Interdisciplinary Studies Research Center (ISRC), Polytechnic of Porto - School of Engineering, Porto, Portugal. Available at: https://www.sciencedirect.com/science/article/pii/S1877050922007955.

[2] Giovanni Capobianco, Umberto Di Giacomo, Francesco Mercaldo, Vittoria Nardone, and Antonella Santone. 2019. Can Machine Learning Predict Soccer Match Results? In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence*. Department of Bioscience and Territory, University of Molise, Italy; Institute for Informatics and Telematics, National Research Council of Italy (CNR), Pisa, Italy; Department of Engineering, University of Sannio, Benevento, Italy.: https://www.scitepress.org/Papers/2019/73075/73075.pdf.

[3] Nilay Zaveri, Utkarsh Shah, Shubham Tiwari, Pramila Shinde, and Lalit Kumar Teli. 2018. Prediction of Football Match Score and Decision Making Process. *International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC)*, 6(2), 162–165. ISSN: 2321-8169. Mumbai, India. : https://ijritcc.org/index.php/ijritcc/article/view/1441/1441.

[4] Marcel Biezunski. 2025. Serie A Matches Dataset (2020–2025). Compiled via web scraping. : https://www.kaggle.com/datasets/marcelbiezunski/serie-a-matches-dataset-2020-2025/data.