Predicting English Premier League Fixture Results

ADIT PATEL aditpate@iu.edu
BEN JACOBS jacobben@iu.edu
AMAN MANOCHA amanocha@iu.edu

Abstract — We aimed to predict the outcomes of English Premier League (EPL) fixtures using the historical match data. By analyzing the past three seasons of data sourced through the API - Football API, we developed and evaluated predictive models based on team form, match statistics, home/away features, goals scored, and match date/time. We then preprocess this data using Python and Pandas and then implement K-Nearest Neighbors(K-NN) and Naïve Bayes Classifier we identify K-NN as the more effective model. Our analysis revealed key factors influencing EPL results, such as team forma and possession rates. With a focus on prediction high accuracy, we gain insights into the key factors that most strongly influence EPL results, contributing to sports analytics and match forecasting.

Keywords — English Premier League (EPL), API - Football, Data Analysis, Prediction, Pandas, Sports Analytics, Fixture Results, Outcomes

I. INTRODUCTION

The English Premier League (EPL) is one of the world's most competitive and unpredictable football leagues, captivating millions of fans and analysts each season. This unpredictability, driven by team form, player performance, and situational factors like match location and strategy, makes the EPL popular and challenging to analyze. Prediction models for EPL games could provide unique insights, aiding decision-making and enhancing fan engagement.

Advances in data analysis now allow for the processing of vast amounts of match data to uncover patterns that influence game outcomes. While predictive models have succeeded in other sports, EPL prediction presents unique challenges due to its competitive balance and evolving strategies. Recent EPL fixtures demonstrate these challenges: out of 10 games played between November 9 and 10, 2024, four ended in upsets, where the outcomes differed from those predicted by betting sites. This highlights the need for models that go beyond expected outcomes to capture these surprises.

To address this, we will use API Football to collect data on team form, possession, shots on target, and other key metrics. Using Python and Pandas, we'll analyze this data and apply machine learning techniques—K-Nearest Neighbors, Decision Trees, and Naive Bayes—to identify the most influential factors. Our goal is not only to achieve high prediction accuracy but also to shed light on why certain matches defy expectations, providing a deeper understanding of EPL match dynamics.

II. METHODS

A. Data Acquisition

We utilized the API-Football API to gather detailed match data from the past three EPL seasons, including team performance metrices, possession rates, goals scored, and match dates/locations. This API allows us to retrieve information over a specific time frame, making it easy to set up a dataset from the past three EPL seasons. By focusing on this recent timeframe, we can balance model accuracy with relevance, capturing the most current dynamics in team performance and league trends. API-Football was also a good choice because of its in-depth documentation, making it very easy to learn and extract all the data we need to build our predictive model. We integrated this data using Python environment, streamlining the acquisition process.

B. Data Preprocessing

Once we have acquired raw EPL data, our next step is to transform it into a clean and structured format suitable for model training. Preprocessing involved cleaning and organizing the raw data using Python and Pandas. Initially, we will perform exploratory data analysis to gain insights into key trends and identify any irregularities within the dataset. Missing values were addressed through interpolation, and statistical analysis was applied to detect and correct outliers. Then, we will standardize the selected attributes like team form, goal scored, possession, and shots on target to ensure compatibility between variables and facilitate comparisons across matches. These steps are important because they ensure that the data we have selected is clean, reliable, and optimized for input into our predictive models. Preprocessing of our data will be very important in our process, as we want to make sure we reduce all possibilities of overfitting in our model. Good data preprocessing with proper standardization and regularization will be key to success.

C. Predictive Model Development

After preprocessing our data, we will apply multiple data mining algorithms to determine the most effective model for predicting EPL match outcomes.

- **K-Nearest Neighbors(K-NN):**

K-NN classified match outcomes by comparing new games to the most similar historical matches based on selected features. We adjusted the parameters, including the number of neighbors and distance metrices, to optimize performance. The model highlighted team form and possession as critical determinants in prediction accuracy. Then for each new match the model calculates the distance to similar games in dataset and predict the result based on the most common outcome among the closest matches. Despite its simplicity, K-NN gives strong insights into recurring patterns.

- **Decision Tree:**

  We will use Scikit Learn (Sklearn) to create a decision tree model that splits data into groups based on the impurity of each attribute such as team form, team accuracy, and possession. This allows our model to focus on the most informative features for making predictions. To prevent overfitting which will occur when our model becomes too tailored to the training data and loses the accuracy of new data. We will use pruning to keep the tree from becoming too complex and overfitting, focusing on the features that make the biggest difference to the outcome.

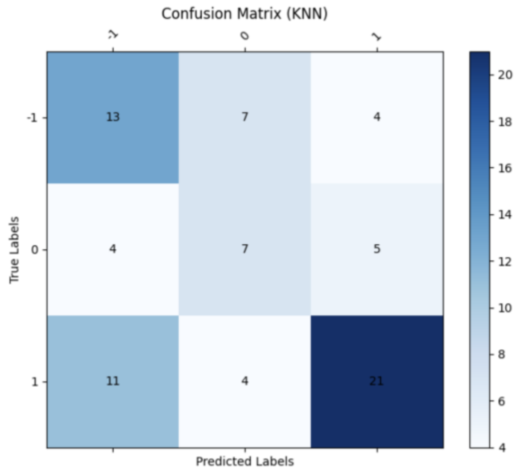
- **Naive Bayes Classifier:**

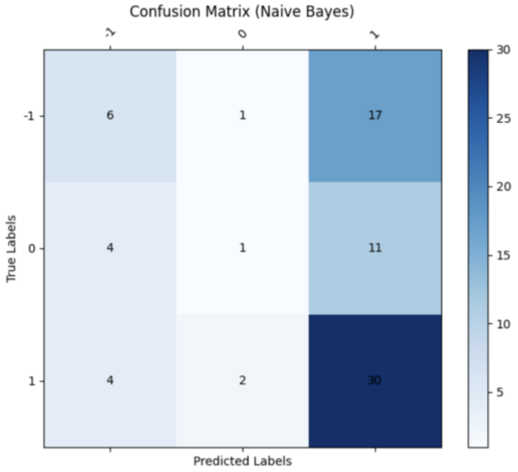
  Naïve Bayes calculated the likelihood of each outcome based on feature probabilities. It will use probabilities for each match feature and combine them to estimate which result is more likely. To keep these calculations stable, we used log transformations because they help to handle small probability values. We implemented it to handle continuous variable such as possession and goals.

After training both models we evaluated them by cross-validation to ensure generalizability. Metrics such as precision, recall, and F-1 score provided comprehensive insights into model performance along with basic accuracy. Once we identify the best-performing model, we will further analyze it to understand which attributes most significantly impact EPL match outcomes.

III. RESULTS

After completing the data preprocessing, we ran two methods on 2022 season data which is K-Nearest Neighbors and Naïve Bayes. We evaluated both the model's using metrics like precision, F-1 score, and recall. Among two models, K-NN gave better results in predicting match outcomes. For instance, it accurately predicted matches where teams with higher possession rates (above 55%) were more likely to win. Draws were harder to predict compared to home or away wins for us. We also noticed that matches involving stable and good team performances were predicted more accurately compared to the matches where there are more injuries or more player substitutions. Naïve Bayes struggled to give us the same accuracy as K-NN's accuracy because of its reliance on feature independence assumptions which are not true always. It is decent at predicting home wins but struggle for other predictions.





In our analysis, we observed that our K-NN implementation (Figure 1) achieved more balanced predictions across classes -1 and 1, with 13 and 21 correct predictions respectively, though it demonstrated some weakness with class 0, correctly identifying only 7 instances. In contrast, our Naive Bayes classifier (Figure 2) showed a stronger tendency to predict class 1, achieving 30 correct predictions for this class, but struggled significantly with classes -1 and 0, yielding only 6 and 1 correct predictions respectively. We found this comparison particularly insightful as it highlights how the two algorithms approach the classification task differently, with K-NN showing more balanced but potentially less confident predictions, while Naive Bayes displayed high confidence but potential bias toward class 1.

IV. DISCUSSION

The results we get illustrate the importance of feature selection and how data preprocessing is necessary step in it. The K-NN model excelled because it effectively used encoded team streaks and match data to identify patterns. Its ability to distinguish between home and away performances was a critical factor in getting higher accuracy than Naïve Bayes.

Our results saw the best performance through the K-NN algorithm but still showed struggles and complications with predicting a draw in for the game, which resulted in lower performance. One change that we would like to implement in the future is possibly eliminating the draw prediction and only predicting a home win or an away win. This is because in the context of sports betting there are double chance lines for soccer 1X for home win or game draw or X2 for away win or game draw. We can eliminate the draw prediction and reduce the complexity of our model and potentially result in better predictions.

An obstacle to this project that we would have liked to include and may potentially expand upon in the future, is the use of team statistics, specifically season statistics. The real challenge with this is that we would need to parametrize the date for the stats, as we want to get the season stats up to the game time so we do not include information that would not be available in real time. The API we used for data allows us to get this data in the parametrized form we need it, however a premium subscription is required if we want to get this data from previous seasons. As we continue to improve this model in the future, this is something we will consider implementing heavily.

All discussions have been oriented towards the K-NN algorithm because again it performed better than the naive bayes algorithm. We believe that this better performance will allow us to build off this algorithm more efficiently in the future. Additionally, we believed the higher performance is a result of K-NN better encapsulating the idea that some teams may perform better against another team's specific tactics. This is best represented in the situation where just because a team B beats team A and then loses to team C, it is not super indicative of whether team A will lose to team C when they play. Because of this important insight K-NN may be the best approach moving forward.

V. PREVIOUS WORK

The idea of building a predictive model for sports outcomes is not an under-attempt project. There have been many different data scientists who have created predictive sports models that are successful. For example, a paper titled "Exploiting sports-betting market using machine learning"[7] written by a professor and students at Czech Technical University, claims to achieve higher accuracy than most models and approached the problem but calculating averages for players based on season based on per-minute efficiencies. Using this they were able to form a model that generated profit over bookmaker odds for 7 NBA seasons (Hubáček et al., 2019). Another paper titled "A Systematic Review of Machine Learning in Sports Betting: Techniques, Challenges, and Future Directions" by researchers at Cornell University, discusses using machine learning in sports betting, which has a very insightful section on generalization and overfitting in a model, and how to mitigate this issue. It also addresses the fact that there is limited information (because of limited games in a season) so overfitting because even more of threat to be weary of (Galekwa et al., 2024). Finally, this paper titled "The Application of the Machine Learning Principles in the Sports Betting Systems" published by the Technical University of Kosice, gives great insight on ways to improve sports prediction model performance and has a special focus on a soccer fixture outcome prediction model. It also discusses the performance of different types of models, including logistic regression and random forest. The paper also concluded longer term data is better than short term data for accuracy (Chovanec & Ružička, 2019).

VI. CONCLUSION

We successfully used data analysis models to predict the English Premier League match outcomes with K-NN showing stronger performance. Our findings also emphasized the impact of factors such as home advantage and recent performance trends. These advancements could further enhance prediction accuracy and the practical application of sports analytics.

ACKNOWLEDGMENT

We acknowledge the API-Football for data access and Scikit-learn for data analysis tools. Along with that ChatGPT is used as a supportive tool in the proposal to help structure our proposal, refine language, and clarify technical explanations. All other project tasks including finding data, coding, data analysis, and model implementation were completed by team members as discussed among us. While Adit and Ben contributed to the GitHub repository, we delegated specific tasks to each member. Contributions were distributed as follows: Adit Patel lead in task management and delegation of work and essentially forming our plan and process including making proposal and final report. Ben Jacobs handled data acquisition, API integration, and coding part. Aman Manocha focused on model implementation.

REFERENCES

[1] API-Football: https://www.apifootball.com/documentation-v3

[2] Pandas: https://pandas.pydata.org/docs/

[3] Sklearn: https://scikit-learn.org/1.5/modules/tree.html

[4] C. Myson, M. Sisneros, (2024, November 7), “Premier League Match Predictions,” Opta Analyst. https://theanalyst.com/na/2024/11/premier-league-match-predictions

[5] M. Cox, M. Carey, A. Walid, T. Harris (2024, August 16), “Premier League Data and tactics roundtable: Expected outliers, setpiece invention and will the goal-rate stay high?” The Athletic. https://www.nytimes.com/athletic/5702819/2024/08/16/premier-league-data-tactics-roundtable/

[6] R. Best, (2023, May 30), “How Consistent was Every Premier League Lineup this Season?” FiveThirtyEight, ABC News. https://projects.fivethirtyeight.com/epl-consistency-2023/

[7] Hubáček, O., Šourek, G., & Železný, F. (2019). Exploiting sports-betting market using machine learning. *International Journal of Forecasting*, *35*(2), 783-796.

[8] Galekwa, R. M., Tshimula, J. M., Tajeuna, E. G., & Kyandoghere, K. (2024). A Systematic Review of Machine Learning in Sports Betting: Techniques, Challenges, and Future Directions. *arXiv preprint arXiv:2410.21484*.

[9] Marek, R. U., & CHOVANEC, M. (2019). The Application of the Machine Learning Principles in the Sports Betting Systems. *Acta Electrotechnica et Informatica*, *19*(3), 16-20.