Predicting English Premier League Fixture Results

Adit Patel, Ben Jacobs, Aman Manocha

Introduction

English Premier League is one of the world's most unpredictable football leagues

It draws of millions of fans and analysts due to its unpredictability and intense matches.

Prediction is difficult due to factors like player form, tactics and external influences.

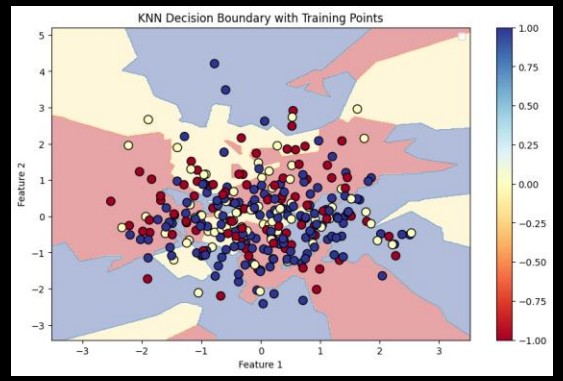
Our Goal

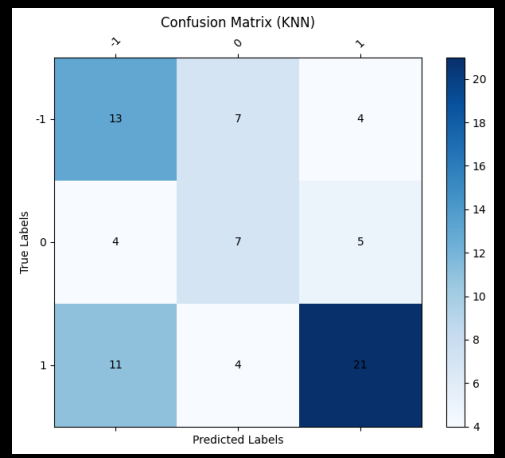
- Create a robust Data Analysis model to predict match results along with higher accuracy
- Identify influential factors such as home advantage, recent match, and key statistics.
- Provide deeper insights into the key factors driving outcomes beyond traditional prediction methods

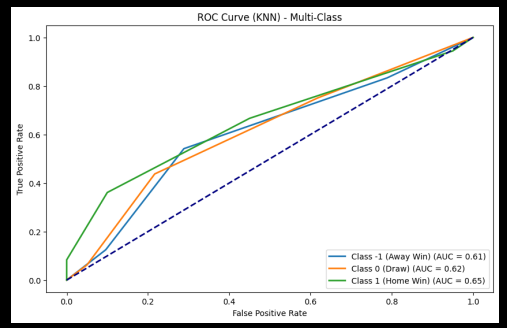
Methods

- Data Acquisition:
  - API-Football API is our data source
  - Highlight key features like team form, goals, possession, shots on target, match timing/location

- Data Preprocessing:
  - Steps: Handling missing data, remove outliers, feature standardization and adding relevant features not directly provided
  - We used Python and Pandas





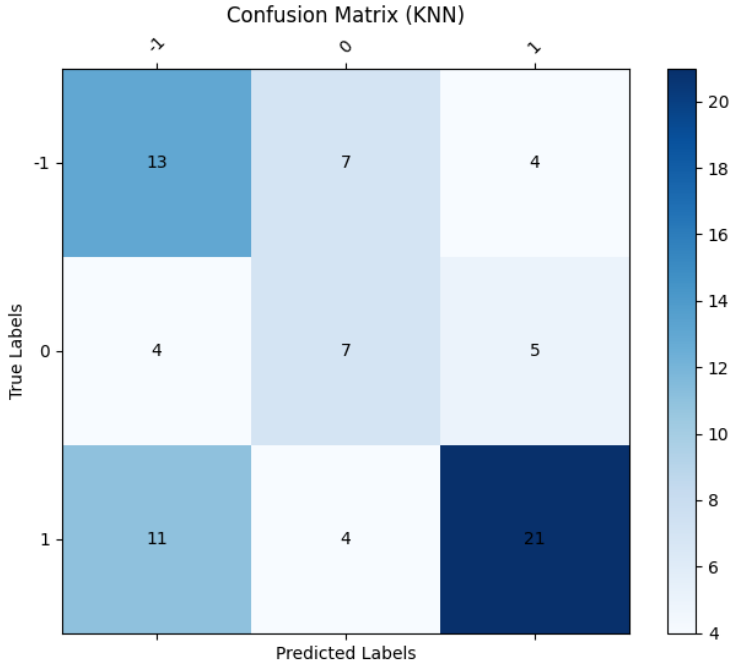


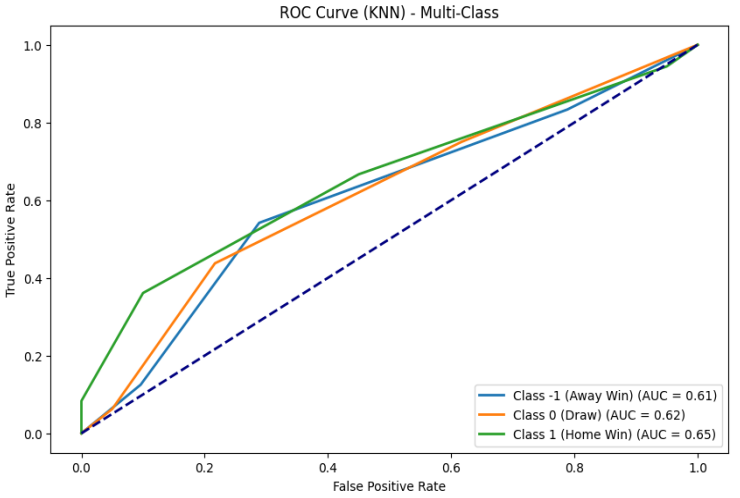
Predictive Model Approach:

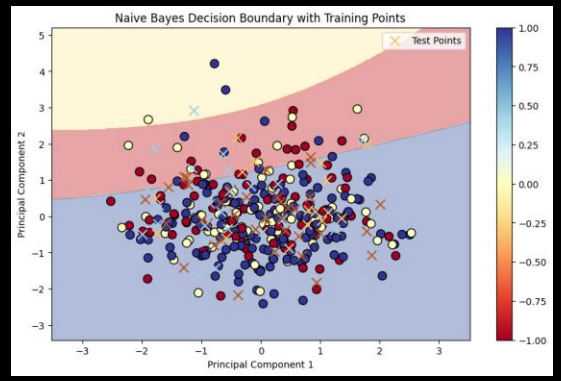
1. K-Nearest Neighbors (K-NN)
   - Classify Outcomes by finding similar historical matches
   - Standardize key statistics
   - Plot ROC curve and Confusion Matrix for each class
   - Achieved moderate accuracy, but struggled with complex patterns

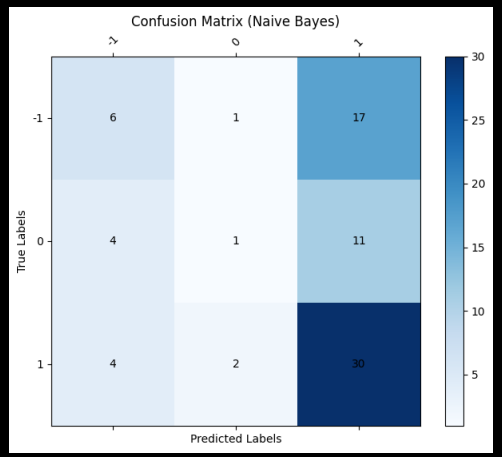
K-NN Results

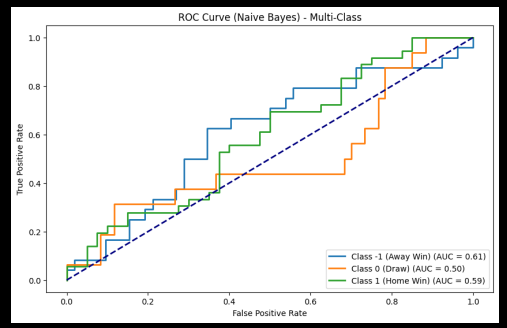
- Decent at predicting home wins (1) and slightly worse at predicting away wins (-1)
- Struggles in predicting draws (0)
- We achieved fair performance when analyzing our ROC curve
  - The AUC (area under curve) values for KNN are not amazing with ranges of 0.61 - 0.65 for each prediction (-1, 0, 1)
- Predicting draws may be a hinderance in our results, and is not exactly relevant in the context of sports betting (more on discussion slides)
- Plots generated with matplotlib











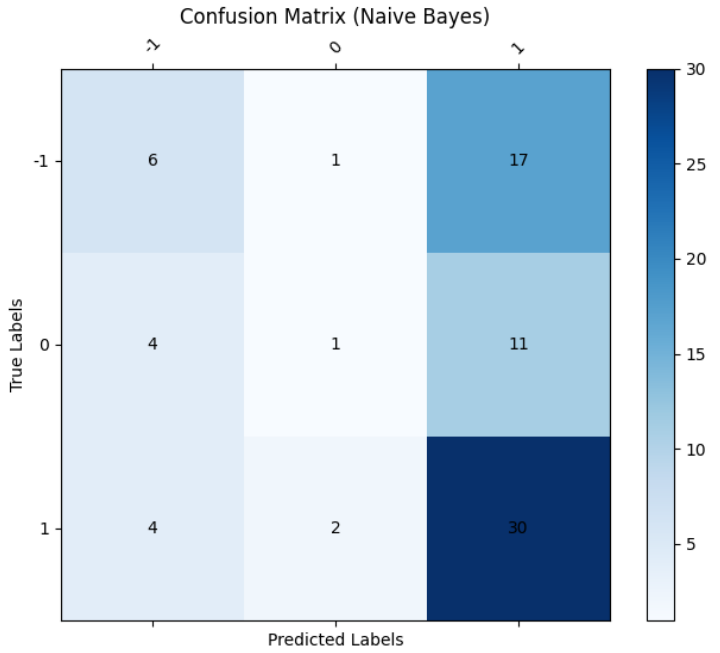
Predictive Model Approach:

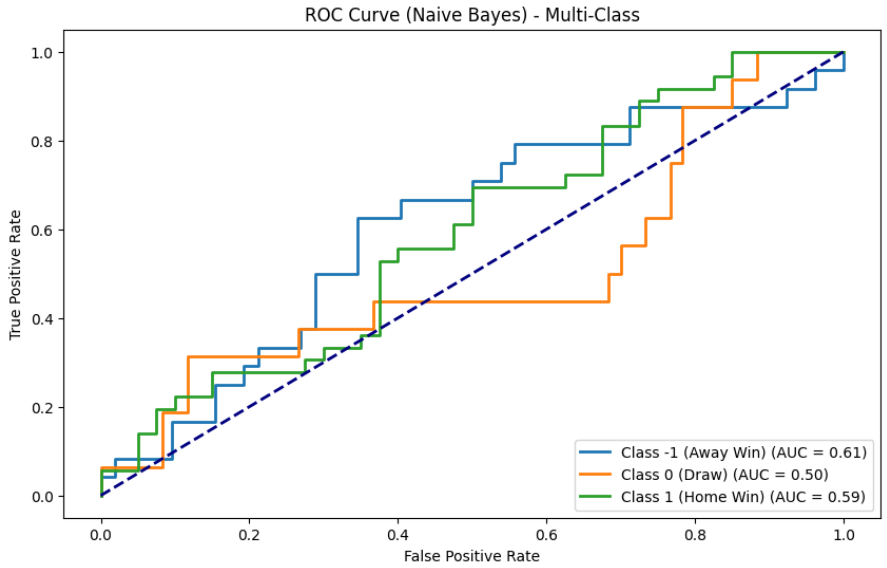
2. Naïve Bayes Classifier

- Estimate win, loss, or draw likelihood
- Applies probabilistic methods for classification
- Plot ROC curve and Confusion Matrix for each class

Naïve Bayes Results

- Much poorer performance than K-NN approach
  - Possibly due to probabilities for results not as relevant because teams match up better against other teams
- Decent at predicting home wins, but struggles for other predictions
- AUC ranges from 0.5 - 0.61 for predicted classes (poor performance)
- Plots generated with matplotlib





Discussion and Future Works

- We saw much better results from K-NN approach, so refining this process are goals for future work
  - Removing the draw prediction and only having a home win or away win prediction may be effective
  - In the context of sports betting, you can take 1X (home win or game draw) or X2 (away win or game draw) lines (read more here)
  - Utilizing this we would only have to worry about misclassifying a home win as an away win and vice versa
- There is much more analysis needed for sports betting algorithms to determine profitability


Q&A

References

- API-Football:
  - https://www.apifootball.com/dpcumentation-v3
- Pandas:
  - https://pandas.pydata.org/docs/
- Sklearn:
  - https://scikit-learn.org/1.5/modules/tree.html