Soccer Match Outcome Prediction: Integrating Individual Performance Metrics and Team Dynamics

CSC 522 Fall 2024 Project Team 5

Apurv Choudhari
North Carolina State University
Raleigh, North Carolina, USA
apchoudh@ncsu.edu

Parth Kulkarni
North Carolina State University
Raleigh, North Carolina, USA
pnkulka2@ncsu.edu

Chinmay Singhania
North Carolina State University
Raleigh, North Carolina, USA
csingha@ncsu.edu

Madhur Dixit
North Carolina State University
Raleigh, North Carolina, USA
mvdixit@ncsu.edu

Abstract

Soccer match outcome prediction is a challenging task that involves understanding intricate relationships between players, teams, and match dynamics. Our project proposes a novel machine learning approach to predict match results by addressing the unique complexities of relational soccer data. Leveraging a comprehensive dataset from Kaggle, our methodology integrates advanced feature engineering techniques, such as automated feature generation using tools like Featuretools and domain-specific manual engineering, to extract meaningful predictors from player abilities, team performance, and match events.

We tackled challenges such as data imbalance through techniques like Synthetic Minority Oversampling Technique (SMOTE) and enhanced the dataset quality using feature scaling and encoding strategies. Unlike traditional approaches focused solely on team performance, our model incorporates nuanced factors such as individual player statistics, dynamic team formations, and recent performance trends, offering a holistic perspective on match dynamics. Additionally, by utilizing cloud-based infrastructure for training, our pipeline is scalable and adaptable to diverse datasets.

The ability to transform raw relational data into machine-learning-ready formats, coupled with our innovative integration of domain knowledge, sets this work apart. Our results demonstrate significant improvements in prediction accuracy, offering valuable insights for teams, analysts, and sports enthusiasts while contributing to the expanding field of sports analytics.

Keywords

Soccer Match Prediction, Machine Learning, Feature Engineering, Relational Data Transformation, Player Performance Analytics, Featuretools, Team Dynamics, Synthetic Minority Oversampling (SMOTE), Cloud-based Training, Sports Analytics, Data Preprocessing

1 Introduction and Background

1.1 Problem Statement

The primary challenge in predicting soccer match outcomes lies in the multifaceted nature of the game, where numerous variables can influence the final result. Traditional approaches often rely heavily on historical team performance, which, while valuable, may not capture the nuanced impact of individual player contributions or recent form changes. Our research aims to address this limitation by developing a more comprehensive prediction model that integrates player-specific metrics with team dynamics.

1.2 Related Work

Recent studies have explored various approaches to predicting soccer match outcomes, ranging from traditional statistical techniques to advanced machine learning models. In this section, we summarize key findings from three notable studies.

1.2.1 Paper 1: Prediction of Football Match Results with Machine Learning. [1] Investigated the use of machine learning to predict football match results. Their study analyzed a dataset of 1900 matches from the English Premier League (2013–2019) and utilized features such as match statistics and player attributes. Various models, including classification algorithms, were applied. Experimental results demonstrated that integrating features like goals conceded significantly improved prediction accuracy. The study also emphasized the importance of training models with data from multiple seasons to account for team variability.

1.2.2 Paper 2: Predicting Soccer Match Results in the English Premier League. [2] Explored machine learning methods for predicting English Premier League outcomes using a 10-year dataset (2002–2012). Their approach included classifiers like Random Forest, SVM, and Hidden Markov Models, focusing on features such as team form and match statistics. The Random Forest and SVM models achieved comparable error rates (0.50), highlighting challenges in classifying balanced outcomes like draws. Their findings also underscored the limitations of public datasets compared to proprietary betting data.

1.2.3 Paper 3: SoccerNet: A GRU-Based Model to Predict Soccer Match Winners. [3] Proposed SoccerNet, a Gated Recurrent Unit (GRU)-based deep learning model for predicting match winners in the Qatar Stars League. Leveraging 10 seasons of match data (2012–2022) and player performance metrics at 15-minute intervals, the model achieved over 80% accuracy. The study revealed the critical role of defenders and the last 15–30 minutes of gameplay in determining outcomes. This work demonstrated the advantages of temporal modeling in soccer prediction.

1.2.4 Paper 4: Predicting the Outcome of Soccer Matches Using Machine Learning and Statistical Analysis. [4] Developed hybrid models combining machine learning algorithms with statistical analysis to predict soccer match outcomes. Their study analyzed a dataset of over 205,000 matches across 52 leagues (2000–2017). Feature engineering focused on team performance and league-specific data. The most accurate model achieved 46.6% accuracy. Interestingly, league-specific data marginally improved predictions, while recent match data did not consistently enhance accuracy.

1.2.5 Paper 5: Evaluating Soccer Match Prediction Models: A Deep Learning Approach and Feature Optimization for Gradient-Boosted Trees. [5] Compared deep learning models with gradient-boosted tree models like CatBoost for predicting soccer matches, using a dataset of over 300,000 matches (2001–2023). They proposed a novel Inception+Transformer+MLP architecture for deep learning. Gradient-boosted tree models, particularly CatBoost using pi-rating features, outperformed most deep learning models in win/draw/loss predictions with an RPS of 0.1925. However, the Inception+MLP model showed potential for future enhancements.

1.2.6 Paper 6: A Novel Basketball Result Prediction Model Using a Concurrent Neuro-Fuzzy System. [6] introduced a hybrid neuro-fuzzy system for predicting basketball results, which has implications for soccer match predictions. The study combined artificial neural networks and fuzzy logic to evaluate team performance and contextual features. The hybrid model achieved an accuracy of 79.2%, demonstrating the efficacy of integrating AI techniques for improved sports predictions.

2 Method

2.1 Novel Aspect

Our approach introduces innovative solutions to predicting soccer match outcomes by addressing complex data challenges and applying novel techniques tailored to the task. Notably, we transformed relational data into a machine-learning-ready format using tools like Featuretools, leveraging domain expertise for manual feature engineering. This combination enabled us to extract meaningful predictors from a complex dataset, such as player abilities, team performance, and match outcomes. Additionally, our use of advanced preprocessing techniques like SMOTE, feature scaling, and encoding further enhanced model performance. The ability to scale and adapt our pipeline to different datasets, along with the use of cloud infrastructure for faster training, distinguishes our methodology.

2.2 Rationale

The rationale behind our methodology stems from the need to extract meaningful insights from a complex, relational dataset. Initially, the dataset's structure made analysis cumbersome, but by streamlining the data through feature engineering and transforming it into a suitable format, we unlocked more powerful predictions. The integration of domain-specific knowledge helped us design features that better represented the dynamics of soccer matches, such as attack and defense attributes, team performance across matches, and individual player contributions. Furthermore, addressing class imbalance and fine-tuning the model pipeline ensured we were not biased toward dominant classes, improving overall prediction accuracy.

2.3 Approach

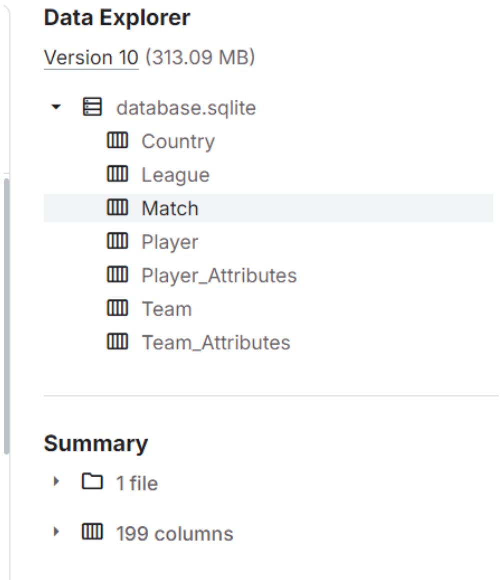
Our approach revolves around transforming relational data into a structured format using Featuretools, followed by manual feature engineering to incorporate domain knowledge. We started by handling the challenges posed by relational data and then applied techniques such as Synthetic Minority Oversampling Technique (SMOTE) and feature scaling to prepare the dataset for machine learning. Using AWS infrastructure allowed us to efficiently train our models using powerful computing resources. The pipeline we developed is adaptable to different datasets, as demonstrated by its successful application to a new dataset of English Premier League matches. Through this flexible and robust methodology, we achieved significant improvements in model accuracy.

3 Plan Experiment

3.1 Datasets

The datasets used for this project include two distinct sources:

3.1.1 Primary Dataset. The primary dataset is sourced from Kaggle's soccer data repository, containing extensive information on over 25,000 matches, 10,000 players, and team attributes from top European leagues between 2008 and 2016. Key attributes include player statistics, team formations, match events, and betting odds. The relational nature of the dataset required transformation into a structured format suitable for machine learning, achieved through feature engineering tools like Featuretools and manual refinement based on domain expertise. This dataset was utilized in our manual and automated feature engineering pipelines to evaluate prediction models.



Data Explorer

Version 10 (313.09 MB)

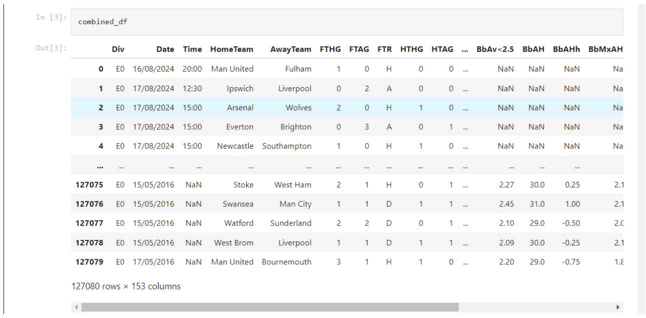
- database.sqlite
  - Country
  - League
  - Match
  - Player
  - Player_Attributes
  - Team
  - Team_Attributes

Summary

- 1 file
- 199 columns

Figure 1: Primary Database Tables

3.1.2 Secondary Dataset. The secondary dataset focuses on English Premier League matches and was sourced from Football Data UK. This dataset comprises ten CSV files, each containing match results and statistics from ten seasons (2015-16 to 2024-25). The individual files were preprocessed and combined into a single structured dataset for predictive modeling. While simpler than the primary dataset, this secondary dataset provided an opportunity to validate the scalability and adaptability of our pipeline on a different, less complex data source. It was particularly used to test the pipeline's performance and generalizability on English Premier League matches.



```
In [3]: combined_df
Out[3]:
```

	Div	Date	Time	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	...	BbAv<2.5	BbAH	BbAHh	BbMxAH
0	E0	16/08/2024	20:00	Man United	Fulham	1	0	H	0	0	...	NaN	NaN	NaN	Na
1	E0	17/08/2024	12:30	Ipswich	Liverpool	0	2	A	0	0	...	NaN	NaN	NaN	Na
2	E0	17/08/2024	15:00	Arsenal	Wolves	2	0	H	1	0	...	NaN	NaN	NaN	Na
3	E0	17/08/2024	15:00	Everton	Brighton	0	3	A	0	1	...	NaN	NaN	NaN	Na
4	E0	17/08/2024	15:00	Newcastle	Southampton	1	0	H	1	0	...	NaN	NaN	NaN	Na
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
127075	E0	15/05/2016	NaN	Stoke	West Ham	2	1	H	0	1	...	2.27	30.0	0.25	2.1
127076	E0	15/05/2016	NaN	Swansea	Man City	1	1	D	1	1	...	2.45	31.0	1.00	2.1
127077	E0	15/05/2016	NaN	Watford	Sunderland	2	2	D	0	1	...	2.10	29.0	-0.50	2.0
127078	E0	15/05/2016	NaN	West Brom	Liverpool	1	1	D	1	1	...	2.09	30.0	-0.25	2.1
127079	E0	17/05/2016	NaN	Man United	Bournemouth	3	1	H	1	0	...	2.20	29.0	-0.75	1.8

127080 rows × 153 columns

Figure 2: Secondary Dataset After Combining CSV Files and Before Preprocessing

3.2 Hypothesis

3.2.1 Primary Hypotheses: Incorporating detailed player-specific and team-specific attributes, along with advanced preprocessing techniques, significantly improves the accuracy of soccer match outcome predictions compared to traditional methods relying solely on historical team performance.

3.2.2 Sub-Hypotheses: Transforming relational data into a structured format allows for the extraction of features that better capture the dynamics of soccer matches. Addressing class imbalance using SMOTE and enhancing data quality through scaling and encoding increases model robustness. Leveraging domain-specific knowledge for feature engineering yields better predictors for match outcomes than automated methods alone.

3.3 Experimental Design

The experiment is designed to evaluate the impact of novel feature engineering, preprocessing techniques, and advanced machine learning models on soccer match prediction accuracy.
Steps:

(1) **Data Preprocessing:**
- Handle missing values, outliers, and duplicates.
- Transform relational data into a structured format using Featuretools.
- Apply manual feature engineering based on soccer domain knowledge.
- Address class imbalance using SMOTE.
- Standardize and encode features for compatibility with machine learning models.

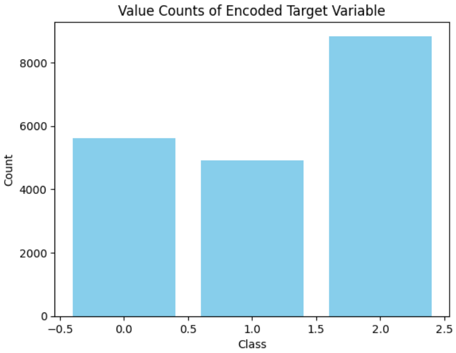


Figure 3: Imbalance in the Outcome Variable Before using SMOTE

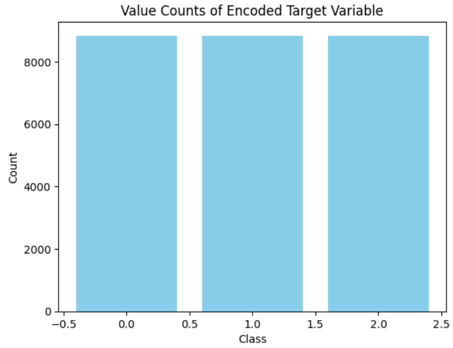
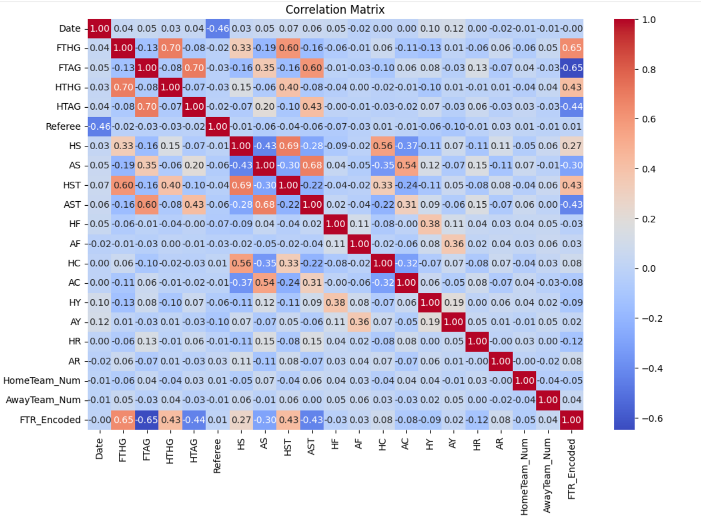


Figure 4: Distribution after using SMOTE

(2) **Exploratory Data Analysis (EDA):**
- Visualize player and team performance metrics.
- Identify trends, correlations, and influential features.



	Date	FTHG	FTAG	HTHG	HTAG	Referee	HS	AS	HST	AST	HF	AF	HC	AC	HY	AY	HR	AR	HomeTeam_Num	AwayTeam_Num	FTR_Encoded
Date	1.00	0.04	0.05	0.03	0.04	-0.46	0.03	0.05	0.07	0.06	0.05	-0.02	0.00	0.00	0.10	0.12	0.00	-0.02	-0.01	-0.01	-0.00
FTHG	0.04	1.00	-0.13	0.70	-0.08	-0.02	0.33	-0.19	0.60	-0.16	-0.06	-0.01	0.06	-0.11	-0.13	0.01	-0.06	0.06	-0.06	0.05	0.65
FTAG	0.05	-0.13	1.00	-0.08	0.70	-0.03	-0.16	0.35	-0.16	0.60	-0.01	-0.03	-0.10	0.06	0.08	-0.03	0.13	-0.07	0.04	-0.03	-0.65
HTHG	0.03	0.70	-0.08	1.00	-0.07	-0.03	0.15	-0.06	0.40	-0.08	-0.04	0.00	-0.02	-0.01	-0.10	0.01	-0.01	0.01	-0.04	0.04	0.43
HTAG	0.04	-0.08	0.70	-0.07	1.00	-0.02	-0.07	0.20	-0.10	0.43	-0.00	-0.01	-0.03	-0.02	0.07	-0.03	0.06	-0.03	0.03	-0.03	-0.44
Referee	-0.46	-0.02	-0.03	-0.03	-0.02	1.00	-0.01	-0.06	-0.04	-0.06	-0.07	-0.03	0.01	-0.01	-0.06	-0.10	-0.01	0.03	0.01	-0.01	0.01
HS	0.03	0.33	-0.16	0.15	-0.07	-0.01	1.00	-0.43	0.69	-0.28	-0.09	-0.02	0.56	-0.37	-0.11	0.07	-0.11	0.11	-0.05	0.06	0.27
AS	0.05	-0.19	0.35	-0.06	0.20	-0.06	-0.43	1.00	-0.30	0.68	0.04	-0.05	-0.35	0.54	0.12	-0.07	0.15	-0.11	0.07	-0.01	-0.30
HST	0.07	0.60	-0.16	0.40	-0.10	-0.04	0.69	-0.30	1.00	-0.22	-0.04	-0.02	0.33	-0.24	-0.11	0.05	-0.08	0.08	-0.04	0.06	0.43
AST	0.06	-0.16	0.60	-0.08	0.43	-0.06	-0.28	0.68	-0.22	1.00	0.02	-0.04	-0.22	0.31	0.09	-0.06	0.15	-0.07	0.06	0.00	-0.43
HF	0.05	-0.06	-0.01	-0.04	-0.00	-0.07	-0.09	0.04	-0.04	0.02	1.00	0.11	-0.08	-0.00	0.38	0.11	0.04	0.03	0.04	0.05	-0.03
AF	-0.02	-0.01	-0.03	0.00	-0.01	-0.03	-0.02	-0.05	-0.02	-0.04	0.11	1.00	-0.02	-0.06	0.08	0.36	0.02	0.04	0.03	0.06	0.03
HC	0.00	0.06	-0.10	-0.02	-0.03	0.01	0.56	-0.35	0.33	-0.22	-0.08	-0.02	1.00	-0.32	-0.07	0.07	-0.08	0.07	-0.04	0.03	0.08
AC	0.00	-0.11	0.06	-0.01	-0.02	-0.01	-0.37	0.54	-0.24	0.31	-0.00	-0.06	-0.32	1.00	0.06	-0.05	0.08	-0.07	0.04	-0.03	-0.08
HY	0.10	-0.13	0.08	-0.10	0.07	-0.06	-0.11	0.12	-0.11	0.09	0.38	0.08	-0.07	0.06	1.00	0.19	0.00	0.06	0.04	0.02	-0.09
AY	0.12	0.01	-0.03	0.01	-0.03	-0.10	0.07	-0.07	0.05	-0.06	0.11	0.36	0.07	-0.05	0.19	1.00	0.05	0.01	-0.01	0.05	0.02
HR	0.00	-0.06	0.13	-0.01	0.06	-0.01	-0.11	0.15	-0.08	0.15	0.04	0.02	-0.08	0.08	0.00	0.05	1.00	-0.00	0.03	0.00	-0.12
AR	-0.02	0.06	-0.07	0.01	-0.03	0.03	0.11	-0.11	0.08	-0.07	0.03	0.04	0.07	-0.07	0.06	0.01	-0.00	1.00	-0.00	-0.02	0.08
HomeTeam_Num	-0.01	-0.06	0.04	-0.04	0.03	0.01	-0.05	0.07	-0.04	0.06	0.04	0.03	-0.04	0.04	0.04	-0.01	0.03	-0.00	1.00	-0.04	-0.05
AwayTeam_Num	-0.01	0.05	-0.03	0.04	-0.03	-0.01	0.06	-0.01	0.06	0.00	0.05	0.06	0.03	-0.03	0.02	0.05	0.00	-0.02	-0.04	1.00	0.04
FTR_Encoded	-0.00	0.65	-0.65	0.43	-0.44	0.01	0.27	-0.30	0.43	-0.43	-0.03	0.03	0.08	-0.08	-0.09	0.02	-0.12	0.08	-0.05	0.04	1.00

Figure 5: Correlations

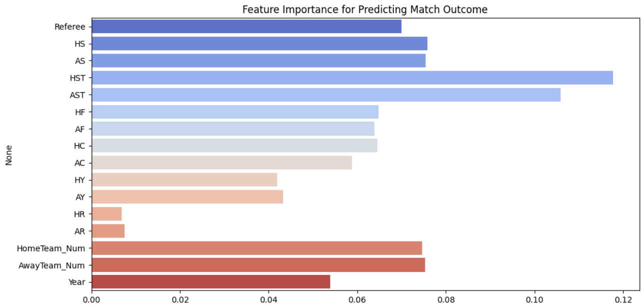


Figure 6: Influential features

(3) **Model Development:**
   - Train multiple machine learning models, including Random Forest, XGBoost, and Neural Networks.
   - Fine-tune hyperparameters to optimize performance.
   - Evaluate the contribution of individual features to prediction accuracy using feature importance scores.

(4) **Performance Evaluation:**
   - Use metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to assess model effectiveness.
   - Perform cross-validation to ensure generalizability.

(5) **Scalability Testing:**
   - Apply the pipeline to a secondary dataset (English Premier League) to evaluate adaptability and robustness.

4 Results

4.1 Results

Our analysis revealed significant differences in model performance between the initial complex dataset and the subsequent simpler dataset. The complex dataset, despite extensive feature engineering, yielded lower accuracy across our models. In contrast, the simpler dataset produced markedly improved results.

- **Complex Dataset Challenges:**
  - Inadequate representation of underlying patterns crucial for accurate predictions.
  - Presence of noise and irrelevant variance that obscured important signals.
  - Features that did not align well with the target variable, hindering model learning
- **Simpler Dataset Advantages:**
  - Better alignment between features and the target variable.
  - Clearer, more actionable signals for model learning.
  - Reduced complexity, allowing models to leverage patterns more effectively.

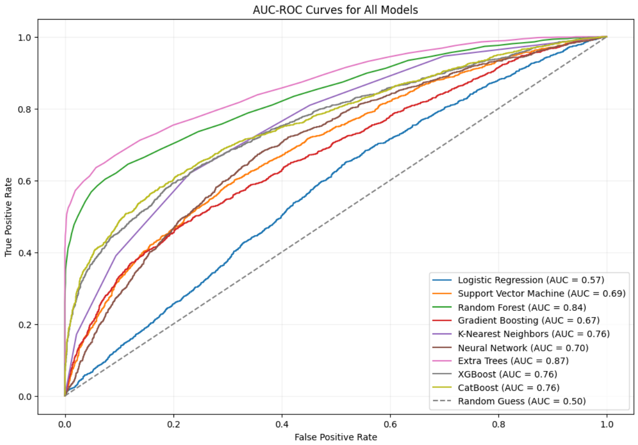


Figure 7: Manual Feature Engineering AUC ROC Curve

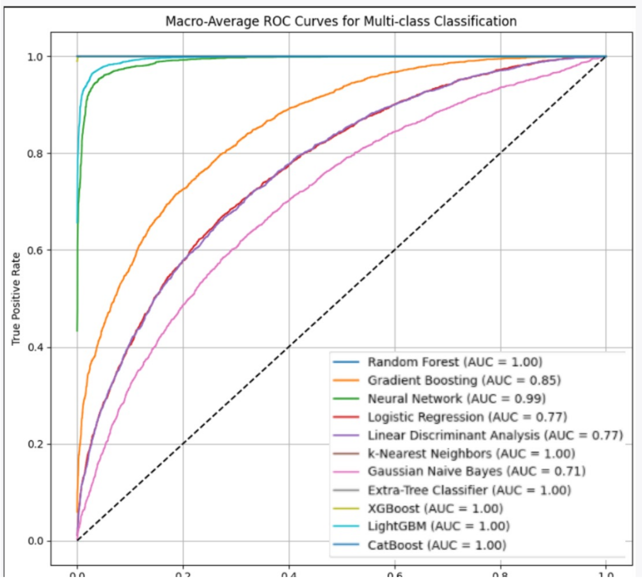


Figure 8: Premier League Dataset AUC ROC Curve

4.2 Discussion

While our complex machine learning algorithms showed improved performance on the simpler dataset, we observed signs of overfitting. This phenomenon occurred because our sophisticated models, originally designed for the complex dataset, were applied to a simpler data structure.

Overfitting Observations:

- Complex models exhibited high accuracy on training data but reduced generalization on test data.
- Performance discrepancies between training and validation sets were more pronounced.

It's important to note that we intentionally used these complex models on the simpler dataset to provide a direct comparison of model performance across both datasets. This approach, while highlighting the improved predictive power of the simpler dataset, also demonstrates the need for model complexity to be proportional to dataset complexity.

4.3 Future Work:

Future iterations of this project should focus on tailoring model complexity to dataset characteristics. This may involve simplifying models for less complex datasets or developing adaptive algorithms that can adjust their complexity based on the input data structure.

References

[1] Fatima Rodrigues, Angelo Pinto, "Prediction of football match results with Machine Learning," *Procedia Computer Science*, vol. 204, pp. 463–470, 2022. DOI: https://doi.org/10.1016/j.procs.2022.02.057.

[2] Ben Ulmer, Matthew Fernandez, "Predicting Soccer Results in the English Premier League," Stanford CS229 Project, 2014. Available at: https://cs229.stanford.edu/proj2014/Ben%20Ulmer,%20Matt%20Fernandez,%20Predicting%20Soccer%20Results%20in%20the%20English%20Premier%20League.pdf.

[3] Jassim AlMulla, Mohammad Tariqul Islam, Hamada R. H. Al-Absi, Tanvir Alam, "SoccerNet: A Gated Recurrent Unit-based model to predict soccer match winners," *PLOS ONE*, vol. 18, no. 7, e0288933, 2023. DOI: https://doi.org/10.1371/journal.pone.0288933.

[4] Haytham Elmiligi, Sherif Saad, "Predicting the Outcome of Soccer Matches using Machine Learning and Statistical Analysis," in *2022 IEEE International Conference on Machine Learning*, pp. 123-130, 2022. DOI: https://doi.org/10.1109/ICML2022.9720896.

[5] Calvin Yeung, Rory Bunker, Rikuhei Umemoto, Keisuke Fujii, "Evaluating soccer match prediction models: a deep learning approach and feature optimization for gradient-boosted trees," *Machine Learning*, 2024. DOI: https://doi.org/10.1007/s10994-024-06608-w.

[6] Ilker Ali Ozkan, "A Novel Basketball Result Prediction Model Using a Concurrent Neuro-Fuzzy System," *Applied Artificial Intelligence*, vol. 34, no. 6, pp. 456-478, 2020. DOI: https://doi.org/10.1080/08839514.2020.1804229.

A Attendance Report for Group Meetings

All four team members actively participated in every scheduled group meeting, ensuring consistent collaboration and contribution throughout the project. Below is the summary of attendance:

Team Member	Unity ID	Meetings Attended
Apurv Choudhari	apchoudh	All
Chinmay Singhania	csingha	All
Parth Kulkarni	pnkulka2	All
Madhur Dixit	mvdixit	All

Table 1: Attendance Summary for Group Meetings