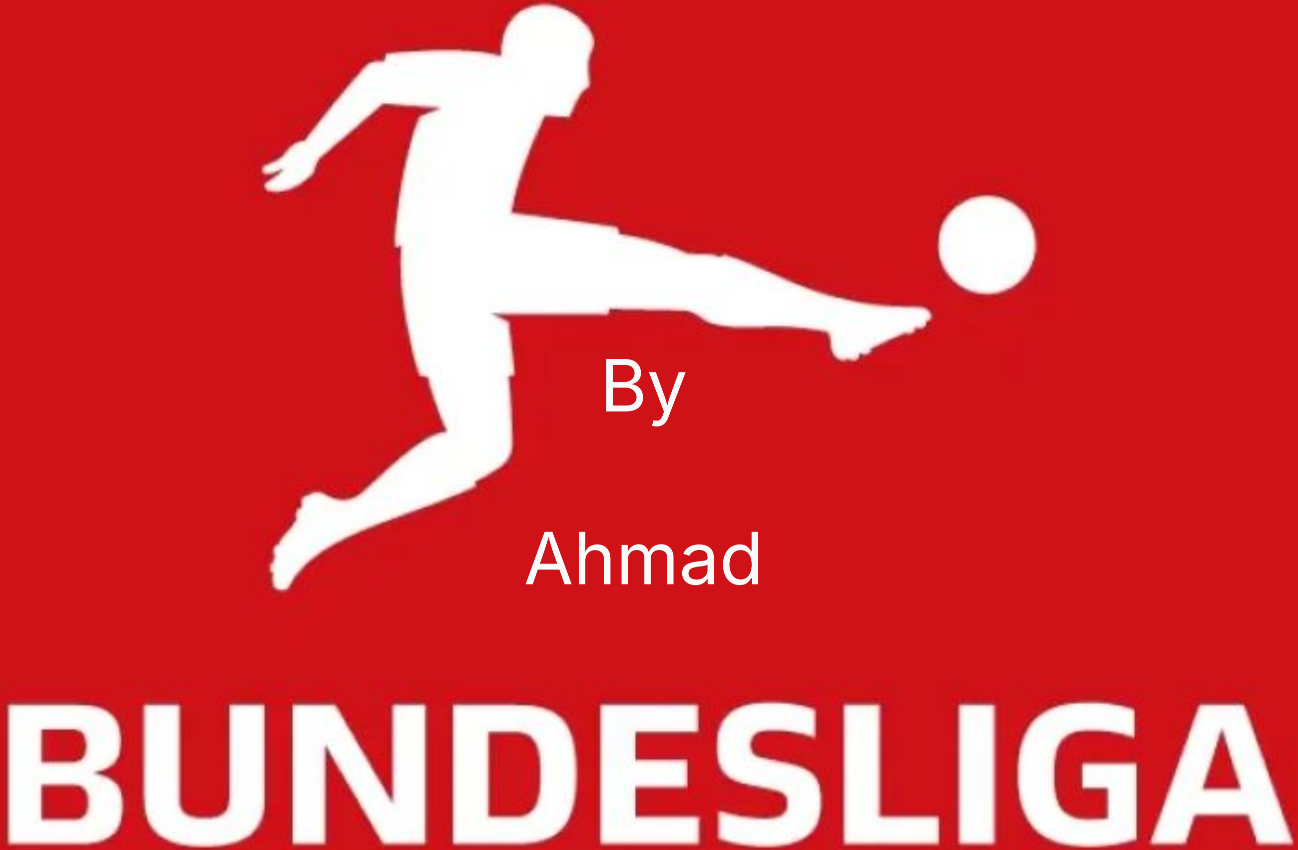
Machine Learning-Based Prediction of Bundesliga 2023/24 Match Outcomes



By

Ahmad

Project Overview

This project applies supervised machine learning techniques to predict both **match outcomes** (win, loss, draw) and **goal differences** based on historical team performance data in Bundesliga for the season year 2023/2024..

We explore classification models (e.g., Logistic Regression, Random Forest, KNN, AdaBoost) to predict categorical match results, and regression models (e.g., Linear Regression, Gradient Boosting, LightGBM) to estimate continuous variables like the number of goals separating two teams.

Primary Goal (Classification): Predict whether the home team will win, lose, or draw using structured match and team performance data.

Secondary Goal (Regression): Estimate the expected goal difference (goals scored minus goals conceded) for more granular match insights.

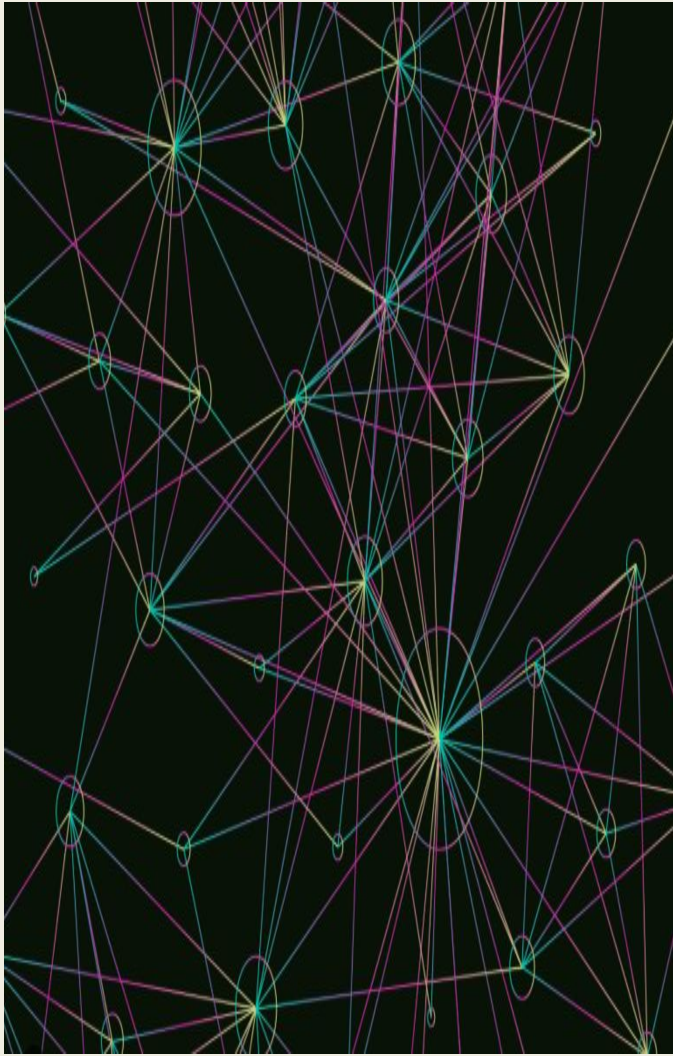
Data Selection and Preparation

📁 **Dataset Overview:**

- **Source:** Transfermarkt, Bundesliga.com, Kaggle

  Small dataset: 18 teams* 17 opponent*2/2 =306 matches

- **Target Variables:**
  - ***Classification**:* `result` → `-1` (Loss), `0` (Draw), 1 (Win)
  - ***Regression:*** `goals_difference` → Numeric value representing goal margin



Data Preprocessing Steps:

- Removed non-numeric and irrelevant columns (e.g., date, team names)

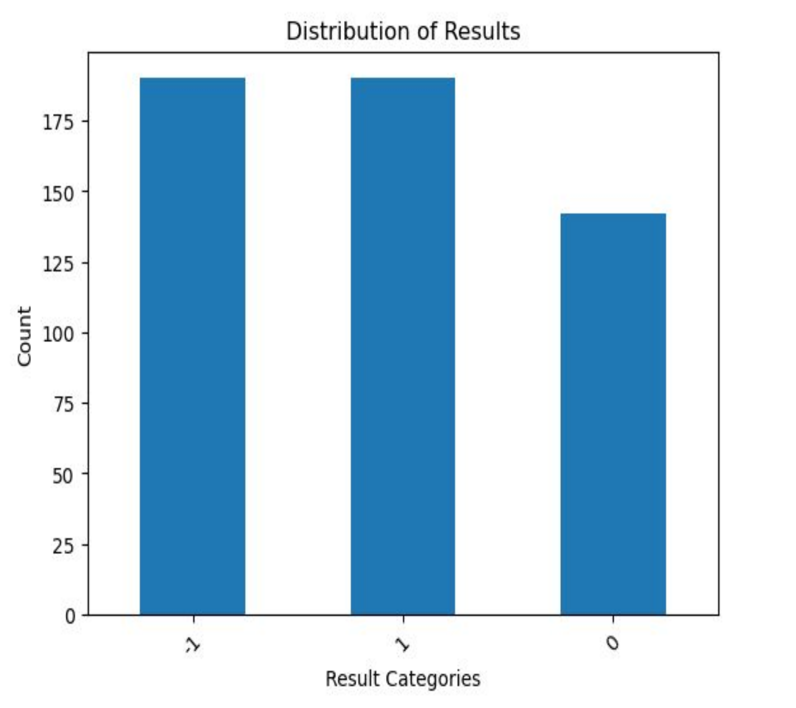
- Handled missing values and class imbalances (via **SMOTE** for classification)

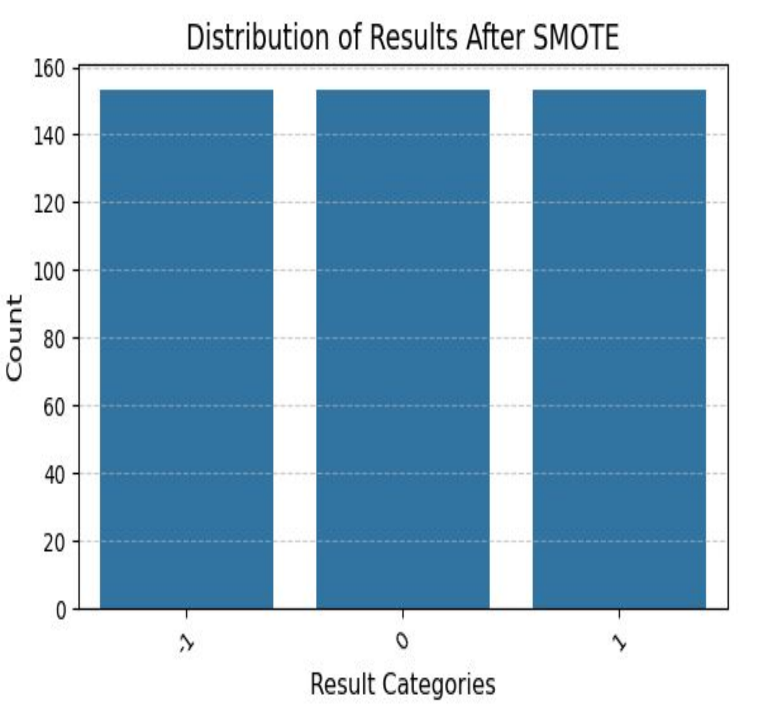
- Feature scaling using **StandardScaler** (for models sensitive to feature scale)

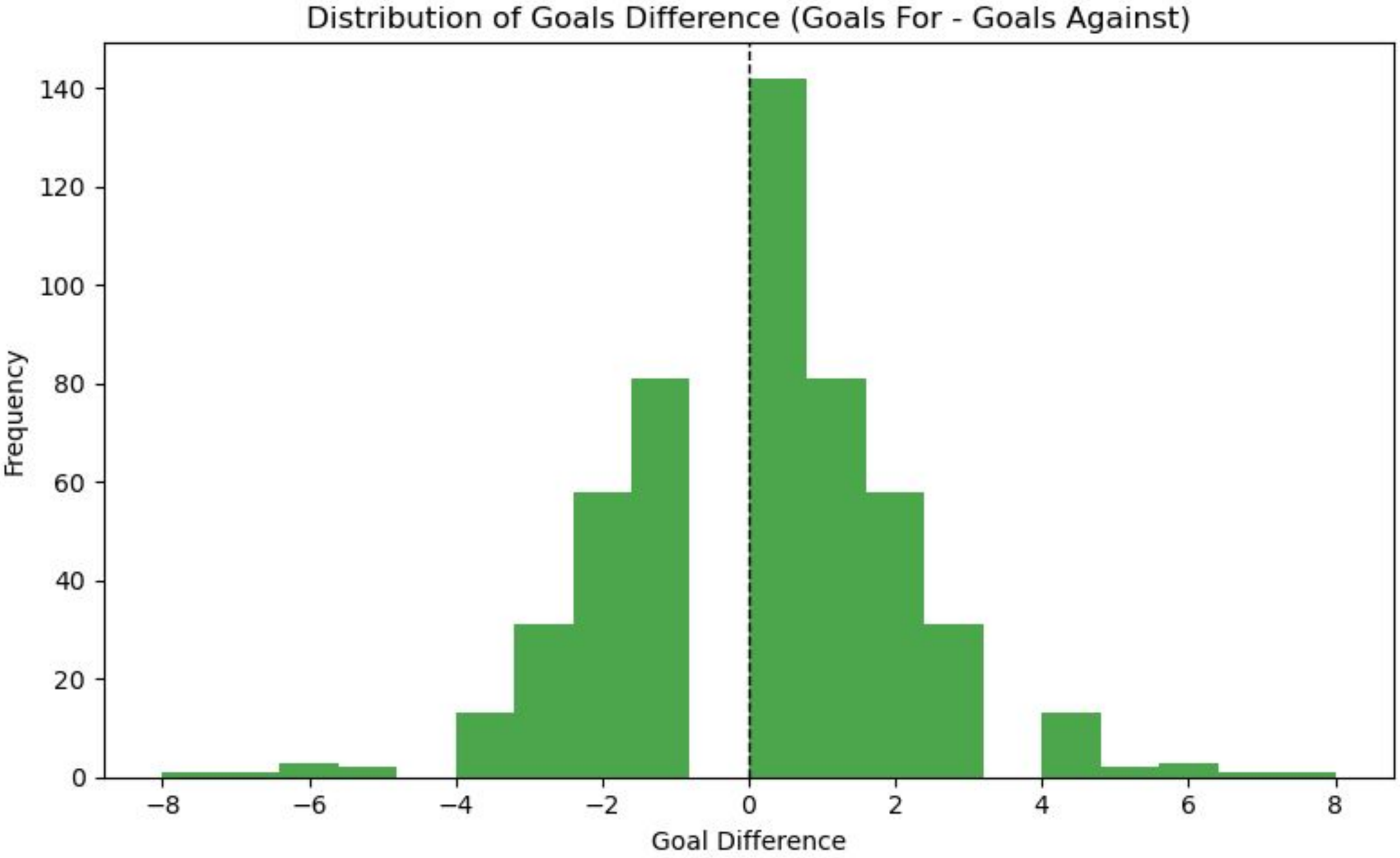
- Feature engineering to add insightful metrics:

    - **Defensive pressure ratio:** interceptions made vs. conceded. Ratio of interceptions made to those conceded (team vs opponent)

    - **Attack efficiency:** goals scored per shot on target
    - **Decipline_gap**: Difference in penalties conceded (team minus opponent)
    - **Last_5_maches result:** WWDLD


Distribution of Results
Count
0
25
50
75
100
125
150
175
-1
1
0
Result Categories


Distribution of Results After SMOTE
Count
0
20
40
60
80
100
120
140
160
-1
0
1
Result Categories


Distribution of Goals Difference (Goals For - Goals Against)
Frequency
0
20
40
60
80
100
120
140
−8
−6
−4
−2
0
2
4
6
8
Goal Difference

Model Building and Evaluation – Classification (Match Result)

Model Used:

- **Logistic Regression** (Multinomial): (Win, Draw, Loss).

Techniques Applied:

- **Pipeline**: StandardScaler → SMOTE → LogisticRegression.
- **Hyperparameter Tuning**: GridSearchCV on `C` (regularization strength).
- **Validation**: 5-fold Cross-Validation for robust performance estimation.

Evaluation Metrics

- **Overall Accuracy**
- **Recall, Precision, F1**

Model Building and Evaluation – Regression (Goals Difference)

Model Used:

- **Linear Regression: Goals_difference**

Techniques Applied:

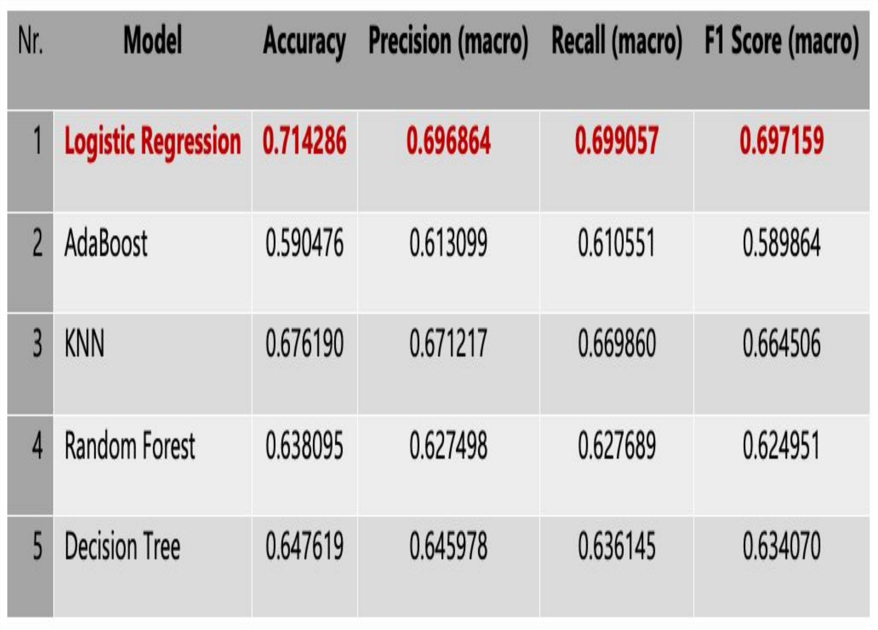
- Scaled data
- Default hyperparameters
- Separate train-test split for evaluation (80/20).

Evaluation Metrics:

- **Mean Squared Error (MSE)** .
- **R^2 Score**

Models Performance:

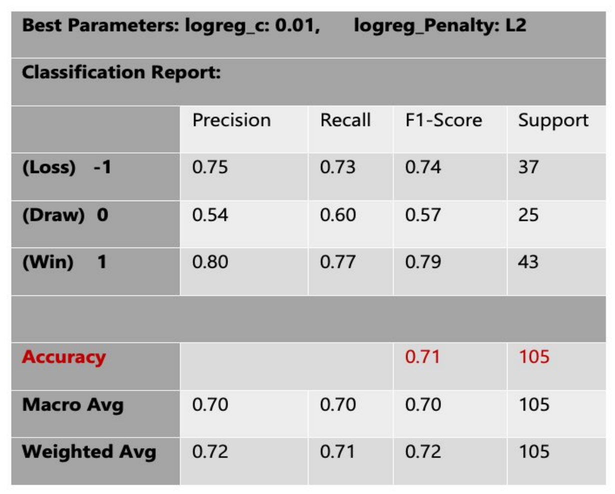
Model that have been used to analysis the Primary goals are:



Nr.	Model	Accuracy	Precision (macro)	Recall (macro)	F1 Score (macro)
1	Logistic Regression	0.714286	0.696864	0.699057	0.697159
2	AdaBoost	0.590476	0.613099	0.610551	0.589864
3	KNN	0.676190	0.671217	0.669860	0.664506
4	Random Forest	0.638095	0.627498	0.627689	0.624951
5	Decision Tree	0.647619	0.645978	0.636145	0.634070

Best Model: *Logistic Regression*

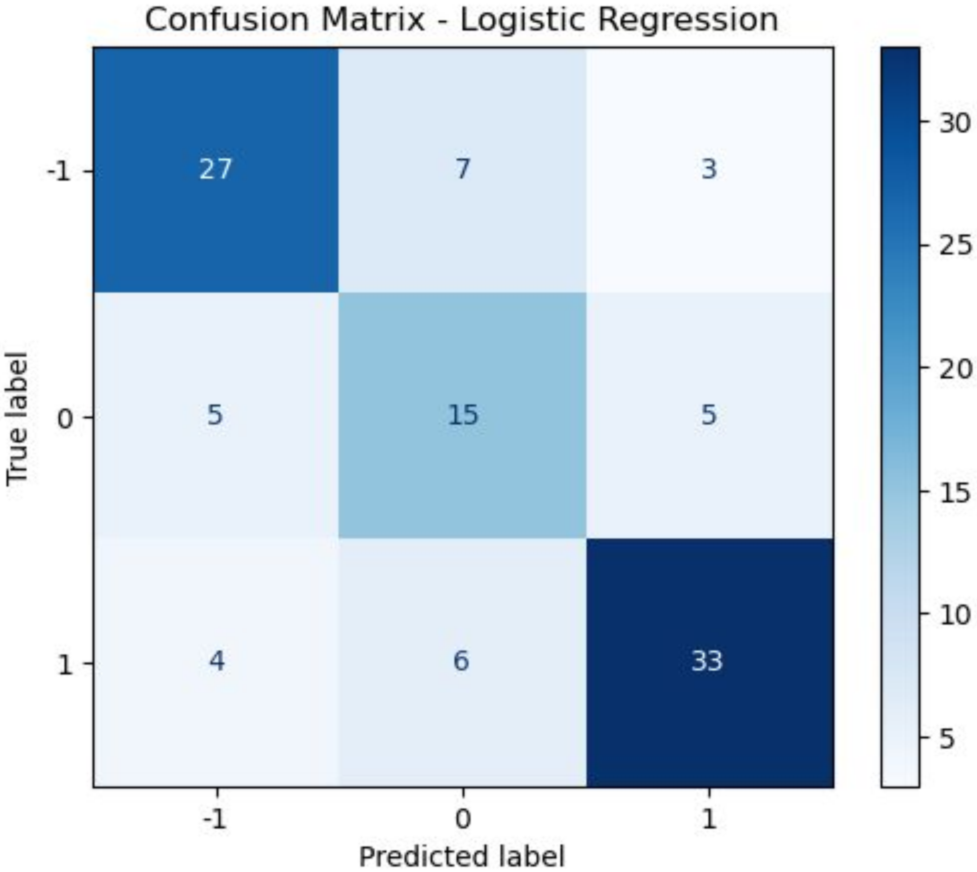
Balanced performance across all three outcome classes (Win, Draw, Loss)



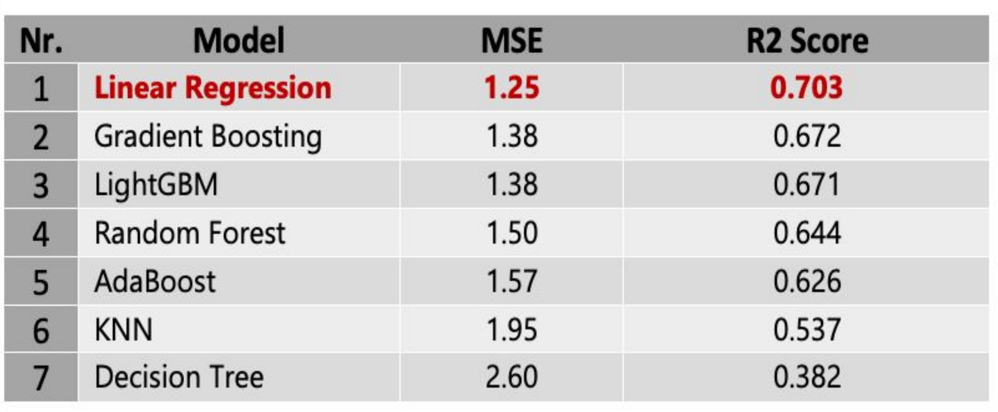
Best Parameters: logreg_c: 0.01, logreg_Penalty: L2

Classification Report:

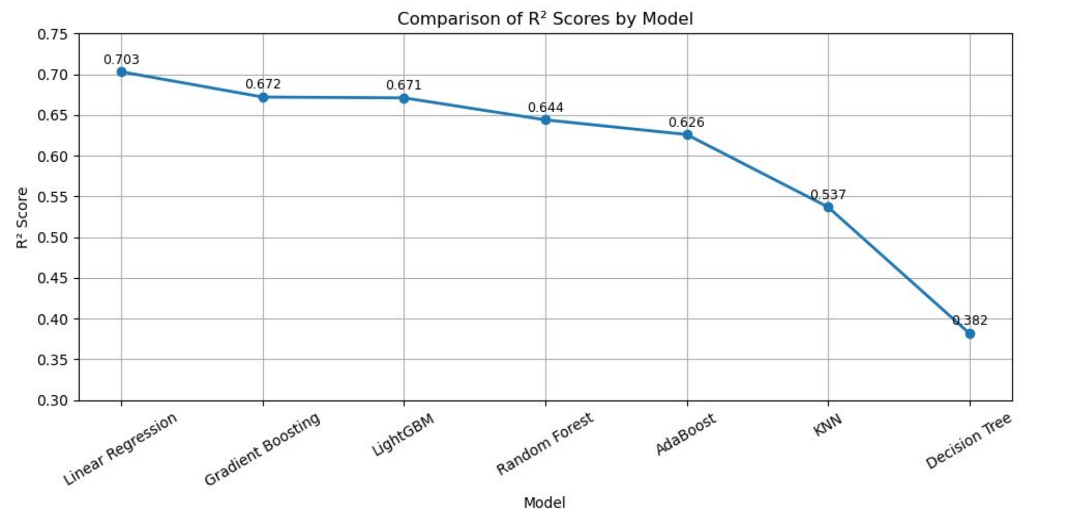
	Precision	Recall	F1-Score	Support
(Loss) -1	0.75	0.73	0.74	37
(Draw) 0	0.54	0.60	0.57	25
(Win) 1	0.80	0.77	0.79	43
Accuracy			0.71	105
Macro Avg	0.70	0.70	0.70	105
Weighted Avg	0.72	0.71	0.72	105


Confusion Matrix - Logistic Regression
True label
-1
0
1
27
7
3
5
15
5
4
6
33
-1
0
1
Predicted label
30
25
20
15
10
5

Models used to find the outcome of a match with goal_difference:

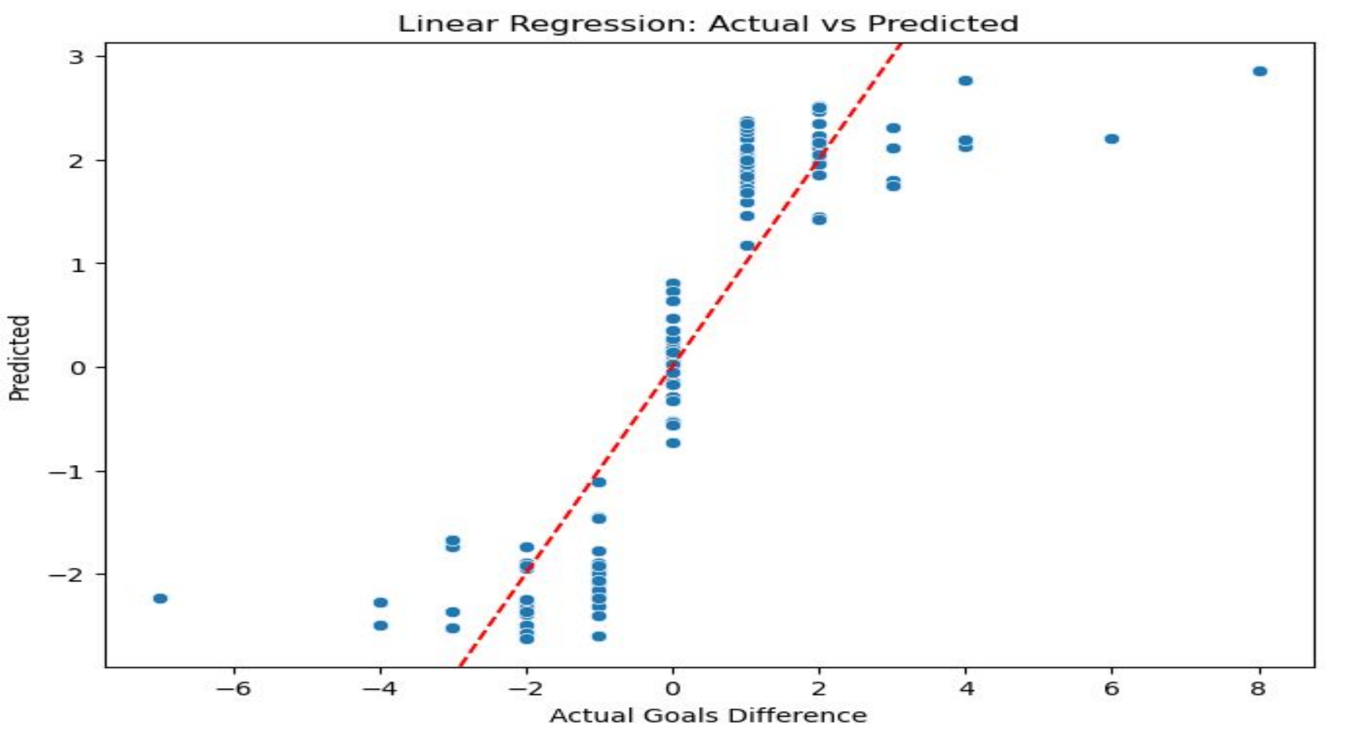


Nr.	Model	MSE	R2 Score
1	**Linear Regression**	**1.25**	**0.703**
2	Gradient Boosting	1.38	0.672
3	LightGBM	1.38	0.671
4	Random Forest	1.50	0.644
5	AdaBoost	1.57	0.626
6	KNN	1.95	0.537
7	Decision Tree	2.60	0.382



Best Model: *Linear Regression*

- Highest R² Score: 0.703
- Indicates ~70% of variability in goal difference is explained by the model.



Real-World Application & Impact

- **Match Outcome Prediction:**
  Teams, analysts, and betting agencies can use the classification model to predict match outcomes based on recent performance metrics and team characteristics.

- **Goals Difference Estimation:**
  The regression model helps estimate score margins, useful for:

  - Pre-game strategy planning

  - Broadcast graphics and match previews

Challenges Faced

- **Class Imbalance in Classification:**
  - The match outcome dataset had more instances of certain results (e.g., draws or wins).
  - Solution: Used SMOTE to balance the target classes and improve model generalization.

- **Feature Relevance:**
  - Identifying which features had the most predictive power required experimentation.
  - Solution: Performed feature engineering (e.g., last 5 match result, `defensive_pressure_ratio`, `attack_efficiency`) and inspected model coefficients/feature importance.

key Learnings

- **Model Choice Matters:**
  - No one-size-fits-all model — performance varies by task (classification vs. regression).
- **Preprocessing is Crucial:**
  - Scaling and data balancing significantly affect model outcomes, especially in KNN and logistic regression.
- **Iterative Development Helps:**
  - Combining feature engineering with model evaluation in cycles led to meaningful performance gains.

Future Work and Improvements

1. Expand Dataset

- Incorporate **more seasons**, **international leagues**, or **player-level statistics**.

2. Advanced Modeling Techniques

- Explore **deep learning architectures** (e.g., RNNs for sequential match data).

Thank you