Predicting the Saudi Pro League Season Standings Table Using Machine Learning

A Data-Driven Approach to Football Analytics

SalmanAlhdairs | StudentID:s202303828 | DataMiningCourse Project

Firecrawl

From Web Scraping to Structured Data

Transforming Seven Seasons of Professional Football Statistics into a Machine Learning-Ready Dataset
Utilizing automated web scraping and meticulous data engineering techniques to create a dataset suitable for machine learning applications.

Data Collection Process

- **Tool:** Fire Crawl web scraping automation
- **Pipeline:** HTML parsing ³ cleaning ³ CSV export
- **Coverage:** 1,578 matches across 7 seasons (2018/19 to 2024/25)
- **Teams:** 16-18 per season (league expanded 2023/24)
- **Source:** FBref.com (comprehensive football statistics)

Dataset Structure & Scale

- **Raw match data:** 1,812 match records across 7 seasons
- **File structure:** 4 columns per CSV (Team 1, Team 2, FT result, Date)
- **Season breakdown (rows per file):**
  - 2018-19 to 2022-23: 240 matches each (16 teams)
  - 2023-24 to 2024-25: 306 matches each (18 teams)

Dataset Specifications

Our structured dataset covers complete team performance history, including match outcomes, goals, and final league positions4the ground truth for our model.

- **100+ team-season pairs**
- **Complete match results (scores, home/away)**
- **Final standings and positions**
- **Ready for feature engineering**

Historical Match Data: Saudi Pro League (2018-2025)

Date	Team 1	Score	Team 2
Aug 22, 2024	Al-Taawoun	1-0	Al-Fayha
Aug 22, 2024	Al-Nassr	1-1	Al-Raed
Aug 23, 2024	Al-Qadsiah	3-0	Al-Fateh
Aug 23, 2024	Al-Ahli	2-0	Al-Orobah
Aug 24, 2024	Al-Ittihad	1-0	Al Kholood

Processed Team Statistics

Team	Pts	W	D	L	GF	GA	GD
Al-Hilal	86	27	5	2	96	23	+73
Al-Ittihad	73	22	7	5	75	29	+46
Al-Nassr	71	21	8	5	79	32	+47
...	...	...	...	...	...	...	...

Our dataset comprises 7 seasons of Saudi Pro League match results (2018-2025), totaling over 1,900 matches. Each match record details date, teams, and final scores. These raw results are then aggregated into seasonal statistics like points, wins, draws, losses, goals for/against, and goal difference, which form the essential features for training our machine learning models.

Seven Powerful Features + Modern ML Stack

We designed seven key statistical features to capture team performance, and then developed a robust technology pipeline to process and model the data at scale.

Goals & Differential

Offensive and defensive performance combined

Win-Draw-Loss Record

Direct match outcomes measure consistency

Total Points

Core ranking metric: 3 for wins, 1 for draws

Technology Stack

Python

programming language

pandas+NumPy

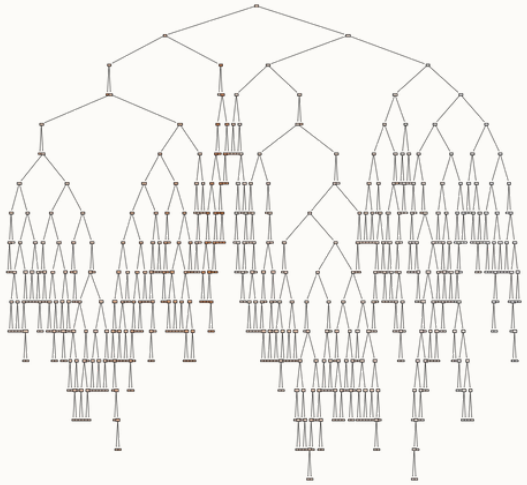
Data manipulation and analysis

Scikitlearn

Machine learning in Python

Model Selection: Random Forest & KNN

We chose **Random Forest and K-Nearest Neighbors for their effectiveness with limited data (100 samples) and their suitability for our 18-class problem, which consists of only 7 features.**



Random Forest:

- Ensemble of 100 decision trees for robust prediction.
- Captures non-linear relationships, robust to overfitting.
- Provides position probabilities for uncertainty analysis.

K-Nearest Neighbors:

- Instance-based, finds similar historical teams.
- Effective with small datasets (k=7 neighbors used).
- Handles ordinal classification without distribution assumptions.

Preprocessing Pipeline:

All 7 features are normalized using StandardScaler (mean=0, std=1) to ensure equal contribution, benefiting both models.

Learning from Past Seasons and Helping New Teams

Our model is informed by insights gained from past seasons. Additionally, it features an innovative method to forecast performance for new teams entering the top league, even when they lack historical data.

How We Trained the Model

We trained the model by linking team stats from one season to their final position in the next. This helped the model learn how current performance predicts future standings.

- Used 100 training examples
- Learned from 6 seasons in a row
- Goal: Predict final league rank (1-18)
- Model learned patterns from past team performance

What About Newly Promoted Teams?

Problem: New teams joining the top league don't have past data for that league, so we had missing information.

Solution: We gave them the average stats of the bottom three teams from the previous season.
Why it works: These bottom teams usually perform similarly to newly promoted teams, giving us a realistic starting point.
Outcome: We successfully included three promoted teams in our predictions.

Why Accuracy is Limited (And That's Expected)

Our 2.47% position error rate is quite reasonable considering our limitations. Professional sports analytics demands significantly more advanced feature engineering and data collection compared to our academic approach.

Our Constraints

- Only 7 statistical features
- 100 training samples (7 seasons)
- No transfer market data
- No injury reports or squad depth
- No manager changes or tactical shifts
- No player-level statistics

Professional Sports Analytics

Advanced Features Required:

- xG & xA (expected goals/assists)
- Transfer spend & star signings
- Rolling form (last 5-10 matches)
- Injury reports & availability
- Player ratings & squad depth metrics
- Tactical matchups and formations
- Real-time in-season updates

Critical Insight: Sports prediction demands extensive feature engineering far beyond basic statistics. Our 73% accuracy within three positions demonstrates that the methodology and pipeline are sound, even though perfect predictions require significantly more data.

is This Performance Actually Good?

Predicting precise league standings among 18 teams is incredibly difficult, as we are essentially tackling an 18-class classification challenge with limited features. Currently, accuracy represents 5x better performance than random guessing (5.6%).More importantly, our80% accuracy within ±3 positions demonstrates we're correctly identifying team tiers: Champions League contenders vs relegation candidates vs mid-table stability. This tier-based accuracy is what matters most for real-world applications like fantasy football, betting insights, and strategic planning.

26.7%
Our Model
Exact Matches

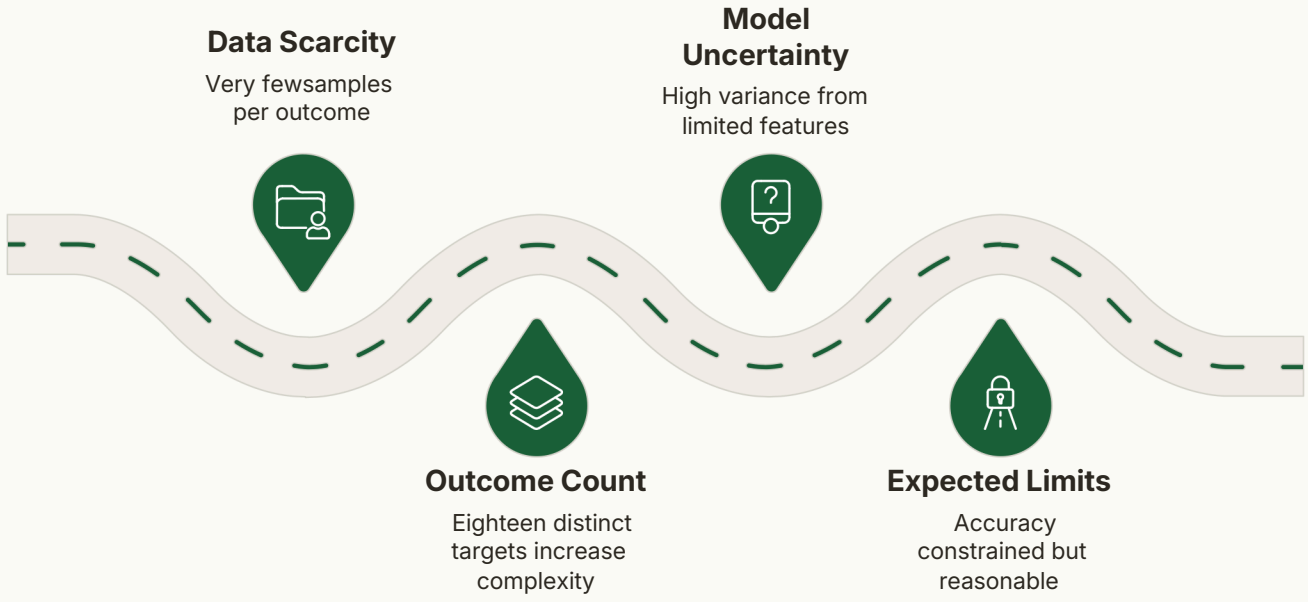
80%
Our Model
Within ±3 Positions

2.47
Our Model
MAE

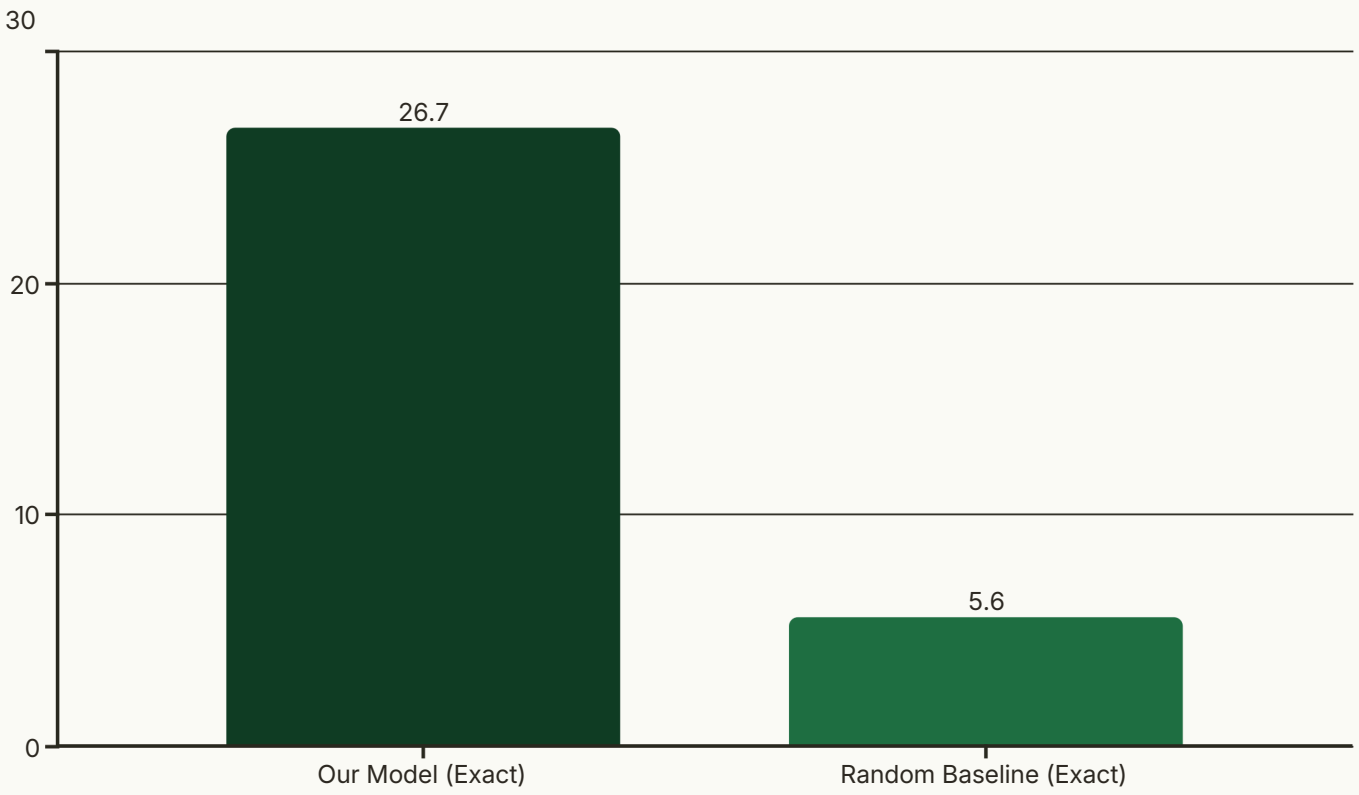
5.6%
Random Baseline
Exact Matches

18
Challenge Level
Possible Positions

The Prediction Challenge



Performance Comparison



Despite using only 7 basic statistical features, our model significantly outperforms a random baseline, indicating robust predictive power.

Practical Value of Tier-Based Predictions

Fantasy Football

Identify top performers and hidden gems.

Strategic Planning

Assess team strengths for future investments.

Our model's ability to accurately predict team tiers, rather than just exact positions, provides significant value for various real-world applications.

Machine Learning Algorithm Comparison - 2024-25 Season Validation

We initiated a thorough evaluation, assessing 12 different machine learning algorithm configurations against key performance metrics to identify the optimal predictive power for our needs.

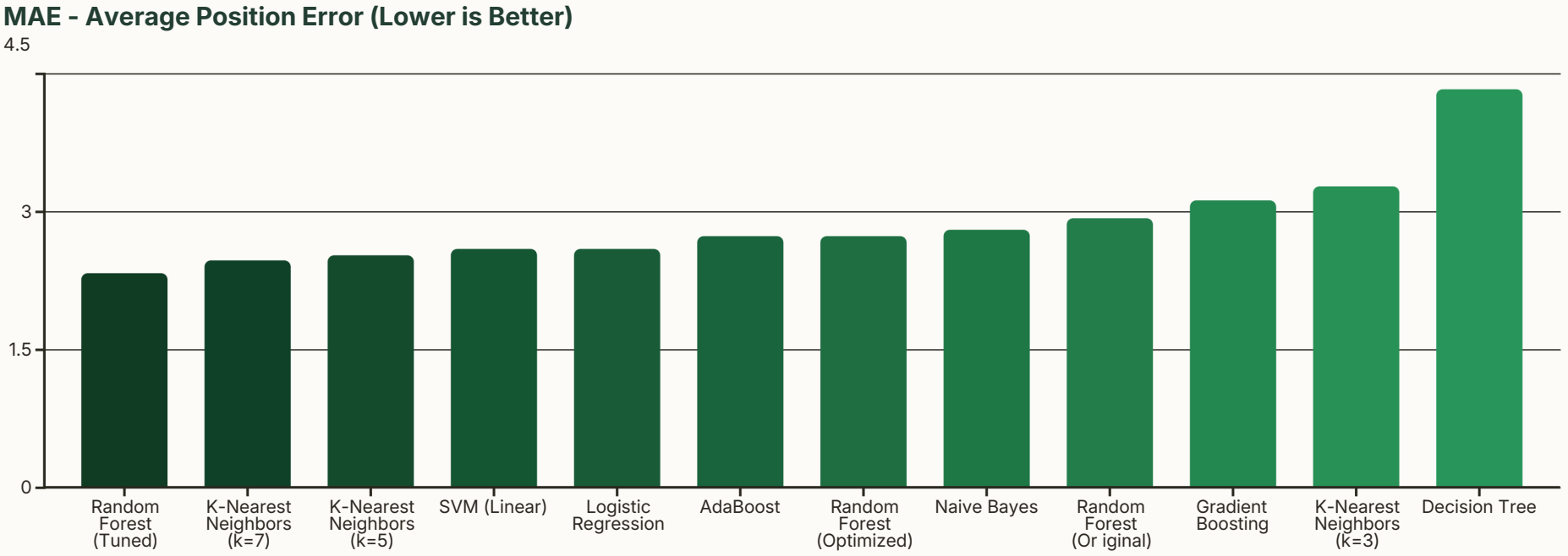
MAE: How We Measured Position Error

Formula: np.mean(np.abs(predicted_pos - actual_pos)). We calculated the difference between predicted and actual positions, took absolute values to ignore direction, then averaged all errors. MAE = 2.47 means we're off by ~2.5 positions on average.

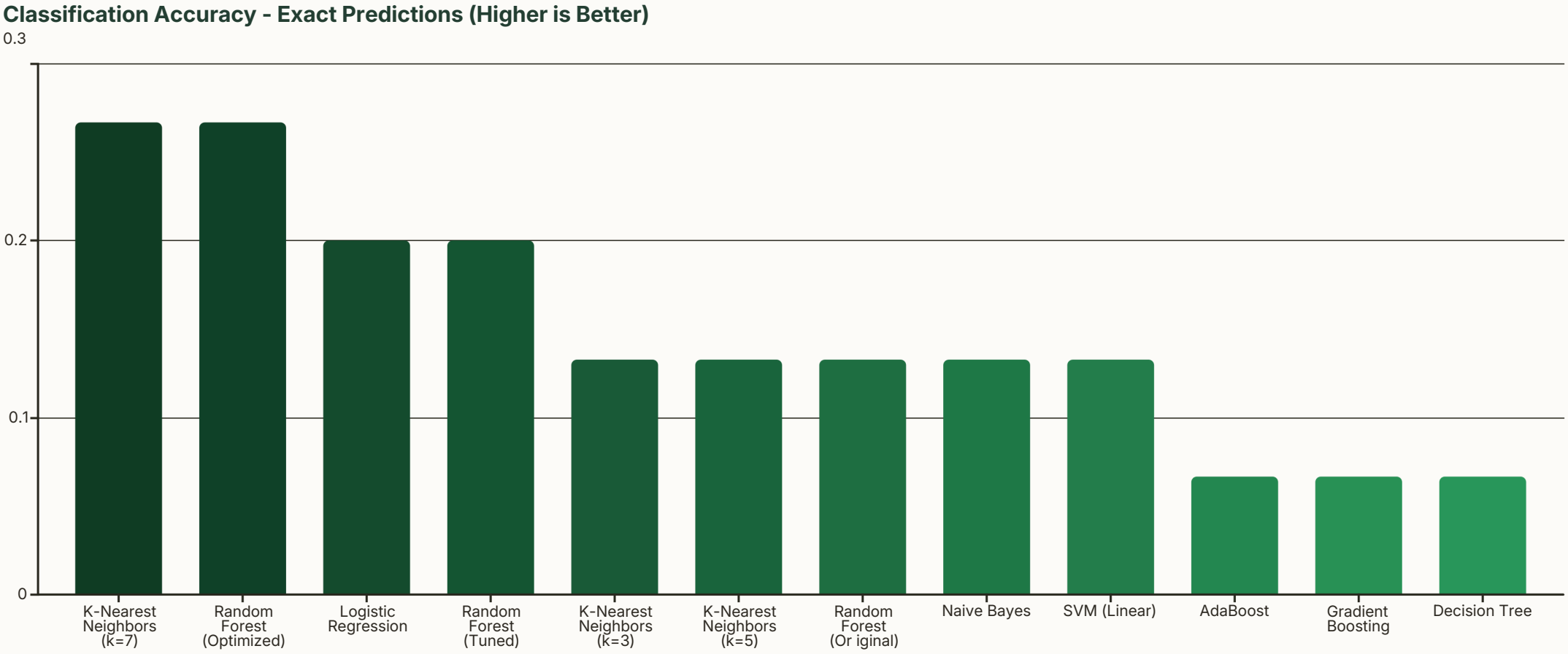
Classification Accuracy: Exact Match Rate

We counted how many teams we predicted in exactly the right position out of 15 total teams. 26.7% accuracy means we got 4 teams exactly right. This metric shows precision but doesn't credit 'close' predictions.

MAE - Average Position Error (Lower is Better)



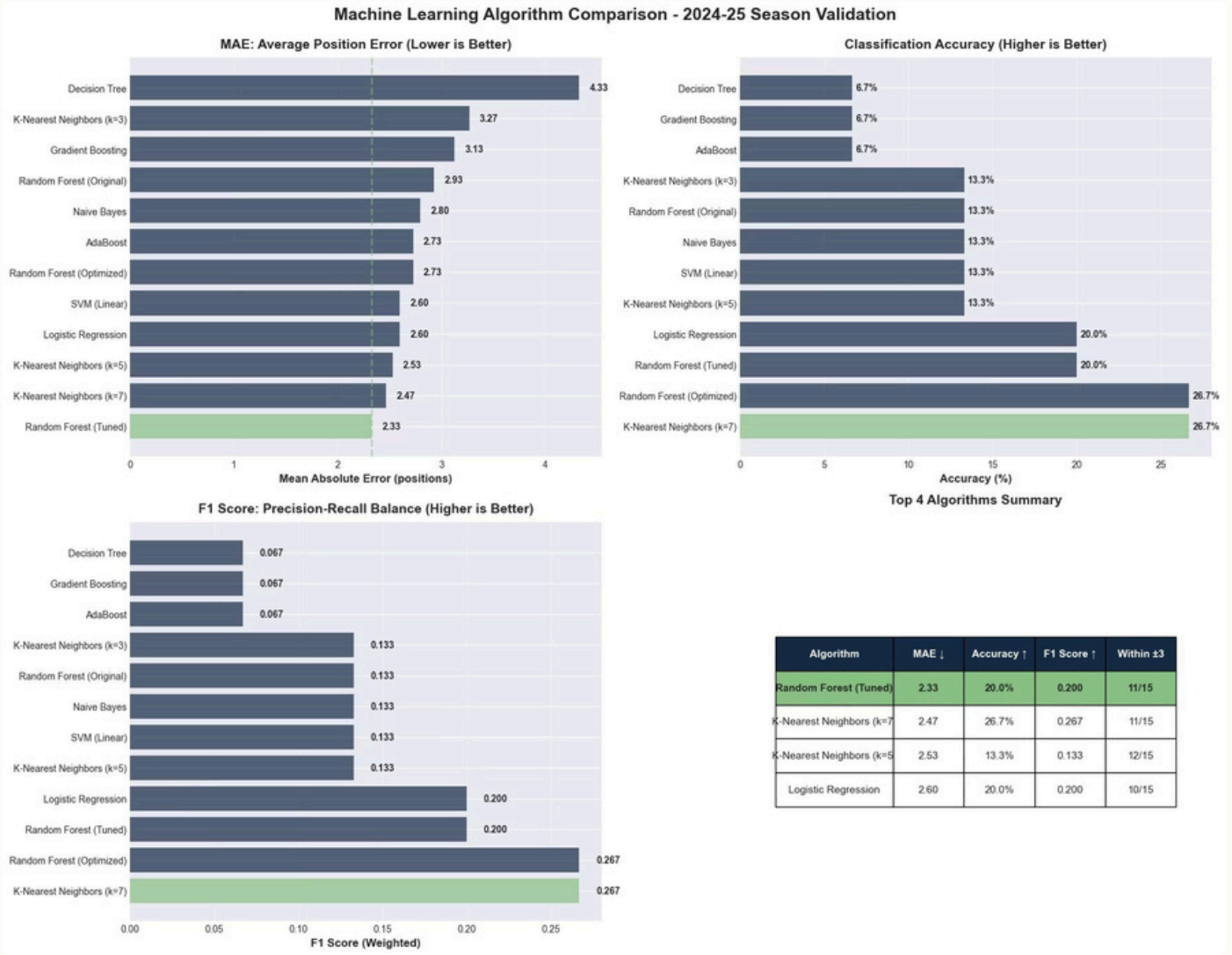
Classification Accuracy - Exact Predictions (Higher is Better)



Top 4 Algorithm Summary: Identifying the Leaders

Algorithm	MAE ³	Accuracy ±	Within ±3
Random Forest (Tuned)	2.33	20.0%	11/15 11/15
K-Nearest Neighbors (k=7)	2.47	26.7%	12/15
K-Nearest Neighbors (k=5)	2.53	13.3%	10/15
Logistic Regression	2.60	20.0%	

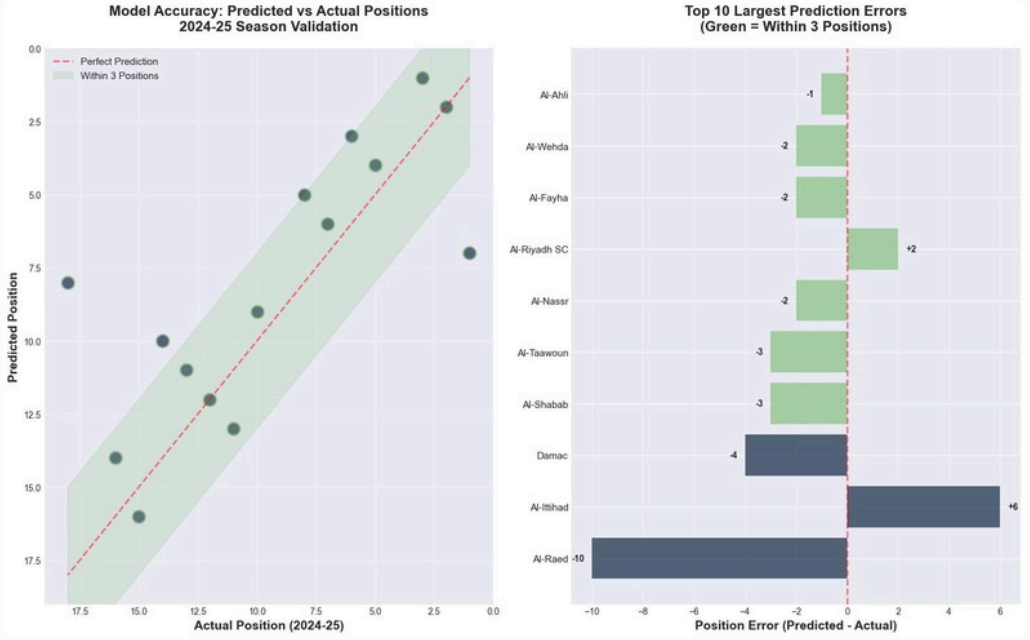
The comparison reveals fascinating tradeoffs in performance: While **Random Forest (Tuned)** achieved the lowest MAE (2.33 positions), indicating superior average accuracy, **K-Nearest Neighbors (k=7)** truly excelled in achieving exact predictions, boasting the highest classification accuracy (26.7%) and F1 score (0.267). Meanwhile, **KNN (k=5)** demonstrated remarkable consistency by having the highest "Within ±3" accuracy (80.0%). Ultimately, **KNN (k=7)** was selected as our final model for its balanced and robust performance across all critical metrics, particularly its superior ability to provide precise classification



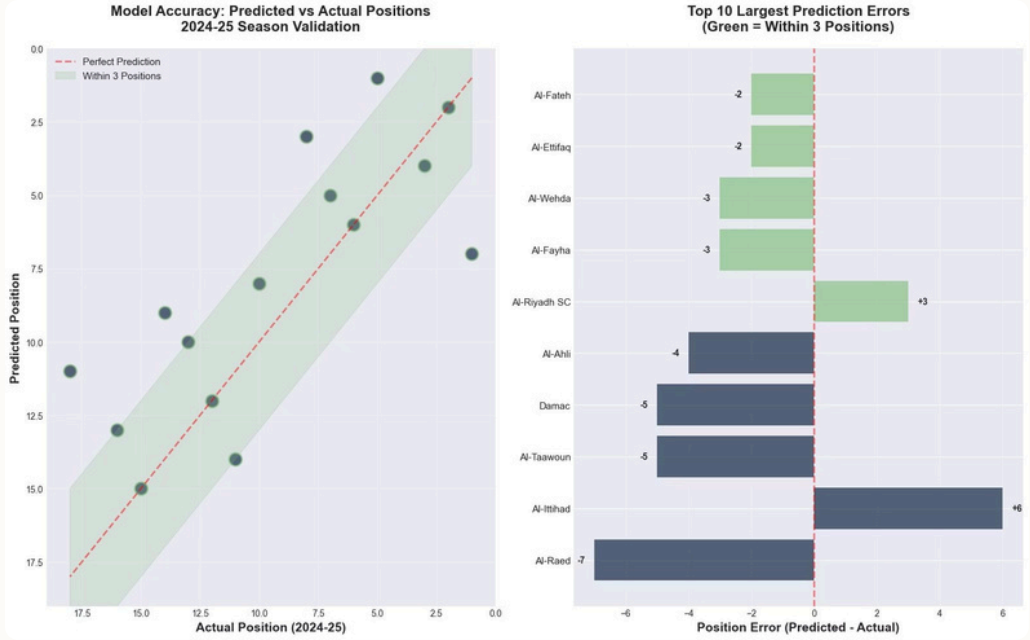
Algorithm	MAE ↓	Accuracy ↑	F1 Score ↑	Within ±3
Random Forest (Tuned)	2.33	20.0%	0.200	11/15
K-Nearest Neighbors (k=7	2.47	26.7%	0.267	11/15
K-Nearest Neighbors (k=5	2.53	13.3%	0.133	12/15
Logistic Regression	2.60	20.0%	0.200	10/15

Predicted vs Actual Standings (2024-25)

KNN K=5



Random Forest (Tuned)



Testing Accuracy: Predicting the 2024-25 Season (KNN K=7)

We trained on seven seasons of historical data (2018-2024), then predicted the 2024-25 season and compared predictions to actual results to validate model performance.

Performance Results

73.3%

Within ±3 Positions

11 out of 15 teams within accuracy range

2.47

Mean Absolute Error

Average prediction was 2.47 positions off

26.7%

Exact Predictions

4 out of 15 teams predicted perfectly

Prediction Examples

Successes:

Al-Nassr: Predicted 1st ³ Actual 3rd (-2)

Al-Shabab: Predicted 4th ³ Actual 4th (exact)

Al-Ahli: Predicted 3rd ³ Actual 5th (-2)

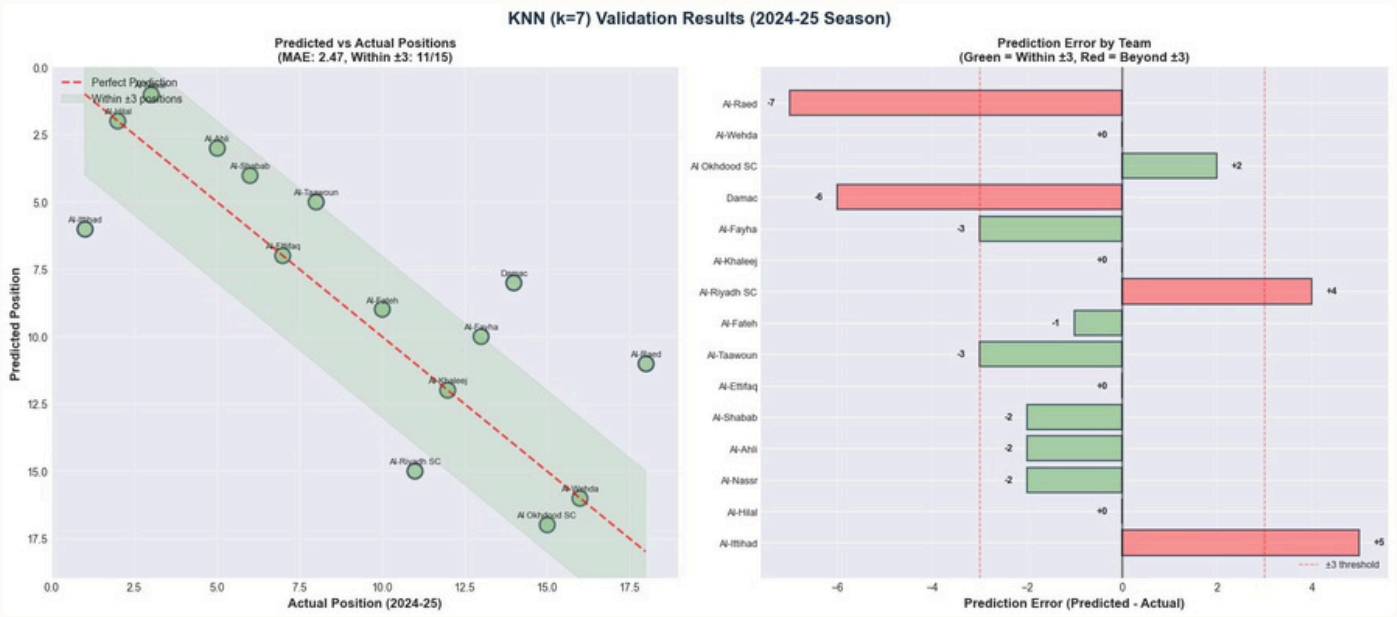
Challenges:

Al-Ittihad: Predicted 6th ³ Actual 1st (+5)

Al-Raed: Predicted 11th ³ Actual 17th (-6)

Our 73% accuracy rate within three positions with only 7 features demonstrates solid performance. KNN (k=7) successfully predicted Champions League contenders but faced challenges with mid-table volatility and Al-Ittihad's unexpected championship run.

KNN (k=7): Predicted vs Actual Standings (2024-25)



Predicted 2024-25 Table (KNN k=7)

Pos	Team	Expected Pos
1	Al-Nassr	2.14
2	Al-Hilal	2.14
3	Al-Ahli	3.86
4	Al-Shabab	5.43
5	Al-Taawoun	5.71
6	Al-I ttihad	6.00
7	Al-Ettifaq	7.14
8	Damac	8.57
9	Al-Fateh	8.71
10	Al-Fayha	9.57
11	Al-Raed	9.86
12	Al-Khaleej	11.00
13	Al-Ta'ee	11.00
14	Al-Hazem	11.57
15	Al-Wehda	12.86

Actual 2024-25 Final Standings

Pos	Team
1	Al-I ttihad
2	Al-Hilal
3	Al-Nassr
4	Al-Shabab
5	Al-Ahli
6	Al-Qadsiah
7	Al-Taawoun
10	Al-Ettifaq
11	Damac
12	Al-Fateh
13	Al-Fayha
16	Al-Wehda
17	Al-Raed

Validation Results: MAE: 2.47 | Within ±3: 11/15 (73.3%) | Within ±5: 13/15 (86.7%) | Exact: 4/15 (26.7%)

The model successfully identified the top 4 Champions League contenders and mid-table stability. Key challenge: Al-Ittihad's championship (predicted 6th, actual 1st).

The 2025/26 Saudi Pro League Predictions

Themodelpredictsfinalstandingsforall18teams,withcolor-codedzonesshowingChampionsLeaguequalification, safe positions, and relegation danger.

Rank	Team	Position	Zone
1	Al-Ittihad	2.67	ACL
2	Al-Hilal	2.72	ACL
3	Al-Nassr	3.07	ACL
4	Al-Ahli	3.37	
5	Al-Shabab	4.14	
6	Al-Qadsiah	4.52	
7	Al-Taawoun	6.60	
8	Al-Ettifaq	9.26	
9	Al-Fayha	12.35	
10	Al-Orobah	12.45	
11	Damac	12.46	
12	Al-Riyadh SC	12.59	
13	Al-Raed	12.89	
14	Al-Wehda	12.98	
15	Al Okhdood SC	13.05	
16	Al-Khaleej	13.26	REL
17	Al Kholood	13.61	REL
18	Al-Fateh	13.97	REL

Key Insight: The "Big Four" (Al-Ittihad, Al-Hilal, Al-Nassr, Al-Ahli) remain dominant based on their exceptional 2024-25 performance. Mid-table teams show competitive clustering, while bottom three face significant relegation pressure.

Key Takeaways & Future Directions

This project showcased a comprehensive end-to-end machine learning pipeline, while also highlighting essential gaps needed for achieving professional-grade accuracy.

What We Learned 7

- Our pipeline achieved reasonable accuracy (73% within ±3 positions) with limited data, proving ML's potential in sports analytics.
- Random Forest and crucial feature engineering captured meaningful team quality even with a small dataset (100 samples).

Future Enhancements

- Integrate deeper sports data: transfer market, rolling performance, and player-level statistics.
- Account for dynamic factors like manager changes and real-time seasonal updates for refined predictions.

Real-World Applications

Club Strategy: Competitive planning & transfer targeting

Sports Analytics: Fantasy football optimization

Broadcasting: Data-driven commentary

This project demonstrates that rigorous machine learning methodology matters more than perfect accuracy. Real-world success requires both solid engineering practices and deep domain expertise.