KING SAUD UNIVERSITY
COLLEGE OF COMPUTER AND INFORMATION SCIENCES
DEPARTMENT OF SOFTWARE ENGINEERING

جامعة الملك سعود
King Saud University

كلية علوم الحاسب والمعلومات
قسم هندسة البرمجيات

SUPERVISED LEARNING

PHASE 2

#	Student Name	Student ID
1	Faisal AlBader	443102460
2	Monther Mohsen Batais	443106846
3	Khaled Mohammed Alawi	443106841

2025
SWE485

Table of Contents

Introduction:

In this phase, we aim to build supervised machine learning models using the Saudi Premier League match dataset to predict the winning team. The models will analyze match data such goals, and other related features to predict whether the home team wins, away team wins, or draw.

We will use at least two supervised learning algorithms, such as XGBoost and Linear Regression, to compare their performance and understand which approach works best for this dataset.

Our Strategy:

Our dataset allows both classification (predicting win, loss, or draw) and regression (predicting the number of goals).
We decided to build two models — one for classification and one for regression — to compare their strengths, weaknesses, and overall performance.

Models used:

XGBoost:

XGBoost is a powerful machine learning algorithm based on decision trees. It works by combining many small trees to make strong and accurate predictions. It is fast, handles large datasets well, and can deal with both classification and regression problems.

Random Forest Classifier:

We chose Random Forest as our supervised learning model. It is a powerful and modern ensemble model that works by building hundreds of individuals "Decision Trees" and combining their predictions through a voting process. We selected it because it is non-linear, meaning it can capture complex patterns and interactions between features (e.g., home_team vs. away_team) that simpler models might miss.

Methodology & Data Preparation:

XGBoost:

we prepared the Saudi Premier League dataset for model training by performing several feature engineering steps. We calculated key statistics for each team, such as

1. Home_Attack_Strength – Average goals scored by the home team
2. Home_Defense_Strength – Average goals conceded by the home team.
3. Away_Attack_Strength – Average goals scored by the away team.
4. Away_Defense_Strength – Average goals conceded by the away team.
5. H2H_Win_Ratio – The home team's win ratio in past matches against the same opponent.
6. Home_Net_Strength – The difference between the home team's attack and the away team's defense.
7. Away_Net_Strength – The difference between the away team's attack and the home team's defense.
8. Strength_Ratio – The overall strength comparison between the home and away teams.
9. Home_Advantage – A constant value representing the general benefit of playing at home.
10.

XGBoost (Extreme Gradient Boosting): This is a powerful machine learning model that builds hundreds of decision trees sequentially. It's known for being highly accurate.

The Settings (Hyperparameters)

Hyperparameters are the model's **control dials** that we set before training. We check a range of options (**param_grid**) for key settings:

- **max_depth (Tree Complexity):** Controls how complex each individual decision tree is.
- **n_estimators (Number of Trees):** Determines how many trees are built in total.
- **learning_rate (Learning Speed):** Controls how fast the model learns from its mistakes.

Random Forest Classifier:

To prepare the data for training, we followed a specific pipeline:

Target Variable (y):

The target column, match_outcome, was not in the original dataset. We created it by defining a function that compared the home_score and away_score for each match.

Feature Selection (X):

As identified in our Phase 1 analysis, several features (stadium, city, attendance, referee_name) were unusable due to having 85-95% of their data missing. Therefore, we selected the following high-quality features with complete data:

1. home_team
2. away_team
3. season_start_year

Feature Encoding:

The machine learning model requires numeric inputs, so we could not use the text-based home_team and away_team columns directly. We used One-Hot Encoding (pd.get_dummies) to convert these two categorical columns into 72 new numeric columns (one for each unique team).

Data Splitting:

We used the train_test_split function from scikit-learn to divide our dataset.

Test Size: We held back 20% of the data (652 matches) as a test set to evaluate the model on unseen data.

Training Size: 80% of the data (2605 matches) was used for training.

Stratification: We used stratify=y to ensure that the class distribution (% of Wins, Draws, and Losses) was identical in both the training and test sets.

Evaluation & Results

We trained the Random Forest model on the training data and evaluated it on the test data. The results, as required by the handbook [1], are shown in the classification report below.

classification results:

Classification Report (Random Forest Classifier)						
	precision	recall	f1-score	support	macro avg	weighted avg
Away Win	0.53	0.50	0.51	208	0.46	0.49
Draw	0.30	0.25	0.27	165	0.46	0.50
Home Win	0.56	0.64	0.60	279	0.46	0.49
accuracy	0.50					

1. Home Win

- **Precision (0.56):** the model predicts a Home Win, it is correct **56%** of the time.
- **Recall (0.64):** The model successfully identifies **64%** of all actual Home Wins in the test set.
- **F1-Score (0.60):** This is the balanced measure, indicating that Home Win is the **best-predicted outcome**.
- **Support (279):** Home Wins are the most frequent outcome in the test data.

2. Away Win

- **Precision (0.53):** the model predicts an Away Win, it is correct **53%** of the time.
- **Recall (0.50):** The model correctly identifies **50%** of all actual Away Wins.
- **F1-Score (0.51):** The performance for Away Win is **strong and reliable**.

3. Draw

- **Precision (0.30):** the model predicts a Draw, it is only correct **30%** of the time.
- **Recall (0.25):** The model misses most of the actual Draws, only identifying **25%** of them.

- **F1-Score (0.27): Predicting a Draw is the model's main weakness.**

The overall accuracy of **50%** is mainly driven by the good performance in predicting Home Wins and Away Wins, compensating for the poor performance on Draws.

Regression results:

Regression Report (XGBoost)

	MAE	RMSE	R²	precision	recall	f1-score
Home GoalsWin	0.9555	1.2817	0.0326	0.58	0.53	0.56
Away Goals	0.8374	1.1683	0.0262			
accuracy	0.5061 (50.61%)					

Confusion matrix (XGBoost)

	Home win	Draw	Away win
Home win	166	66	40
Draw	74	50	42
Away win	56	44	114

- The **Mean Absolute Error (MAE)** for home goals is **0.9555** and for away goals is **0.8374**, meaning the predicted goals are usually off by less than one goal on average.
- The **R² values** are quite low (0.03 for home and 0.02 for away), showing a small part of the variation in goals.
- This means goal prediction is accurate.

Classification results (match outcome):

- The **accuracy** is **50.61%**, which means the model correctly predicts the result in about half of the matches.
- **Precision (0.58):** the model predicts a team will win, it is correct about **58% of the time**.

- **Recall (0.53):** the model correctly finds **53% of all actual wins**.
- **F1-score (0.56):** A score of **0.56** shows a balance between making accurate predictions and finding all true wins, but there's still room to improve.

Confusion Matrix (XGBoost):

- The model correctly predicted **166 home wins**, **50 draws**, and **114 away wins**.

Comparative Analysis of Model Performance

The analysis presents a trade-off between minimizing error and maximizing classification accuracy.

Model 2 (XGBoost) is the **better and more robust model** for this dataset.

Although Model 1 achieved a slightly higher final classification **Accuracy (50.61% vs. 44.56%)**, Model 2 is superior in its primary goal, **regression performance**, by achieving a lower MAE (0.9330).

This indicates that XGBoost has a better understanding of the magnitude of team strengths and goal-scoring tendencies.

Therefore, **the Tuned XGBoost Regressor is recommended** as the optimal model for providing data-driven goal predictions for the Saudi Pro League."