

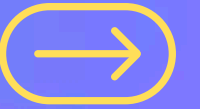
 Final Project Data Science



# MATCH PREDICTION ANALYSIS

PREDICTING EPL MATCH OUTCOMES USING MACHINE  
LEARNING

 Muhammad Umar Abdurrahman



Description



Data Understanding



Data Pre-Processing



Exploratory Data Analysis



Model Preparation



Machine Learning Model



Recommendations

# TABLE OF CONTENTS





## Description

### Fokus

- Analisis & prediksi hasil pertandingan EPL (2020–2023).

### Tujuan

- Bangun model ML untuk klasifikasi home win, away win, draw.

### Target

- Akurasi >60% + insight faktor kemenangan + rekomendasi bisnis.

### Pentingnya Proyek

- EPL = liga bernilai miliaran dolar.
- Bermanfaat bagi perusahaan taruhan (Bet365), fantasy sports (FanDuel), dan klub EPL.

Final Project Data Science

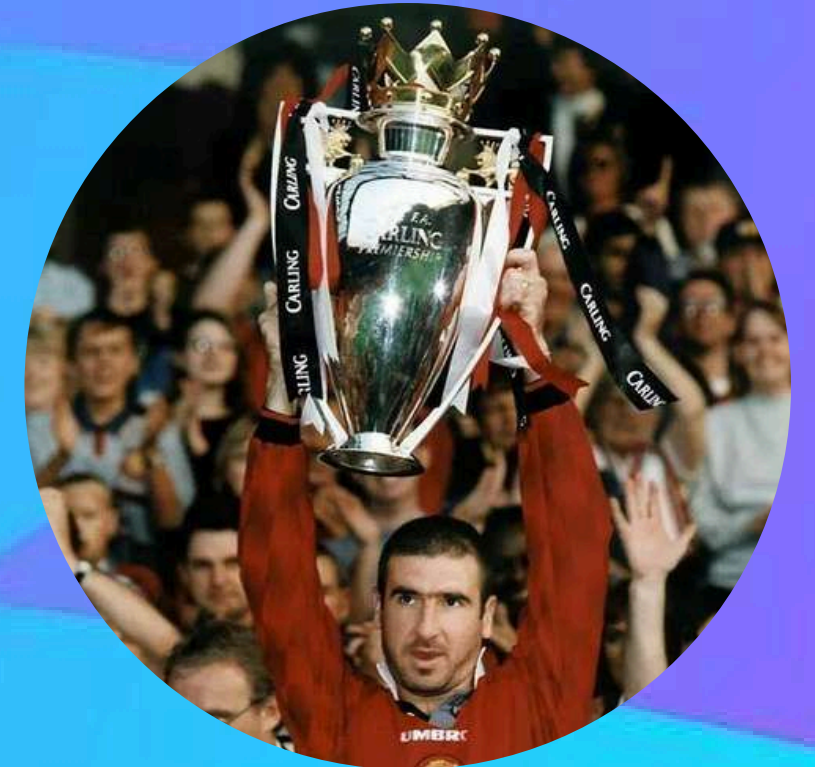


### Problem Statement

- **Problem:** Prediksi hasil pertandingan EPL hanya akurat ~50%, mendekati random guess.
- **Objective:** Membangun model prediksi dengan akurasi  $\geq 60\%$ .
- **Impact:**
  - 📈 Sportsbook → mengurangi risiko finansial.
  - ⚽ Football Clubs → mendukung strategi tim.
  - 📱 Fans & Fantasy Sports → meningkatkan engagement.

# DESCRIPTION

# DATA UNDERSTANDING



## Match prediction Analysis



### Sumber

- English Premier League (2020–2023).

### Ukuran

- 1,140 baris × 40 kolom.

### Fitur Utama

- Match Info: Date, Home Team, Away Team, Attendance.
- Performance Metrics: Goals, Shots, Shots on Target, Possession, Passes, Corners, Fouls.
- Target: Hasil pertandingan → Home Win, Away Win, Draw.



# DATA PRE-PROCESSING

## Missing Values



## Duplicated Data



## Data Manipulated



## Feature Engineering

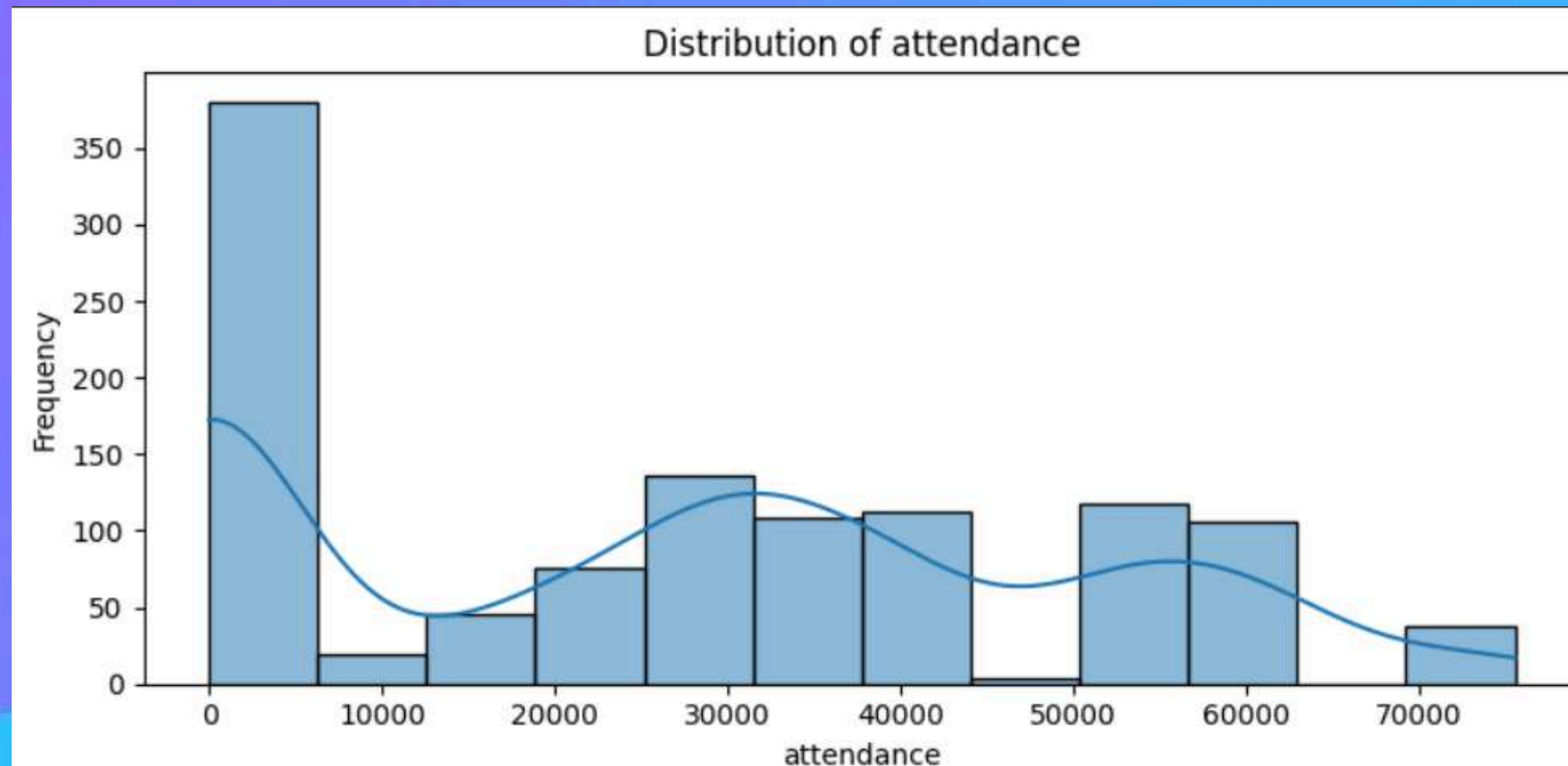
Tidak ada data null atau missing value pada dataframe

Tidak ada data duplikat pada dataframe

Merubah format date yang sebelumnya object menjadi datetime

Memilih feature yang akan digunakan untuk EDA dan model machine learning

## Attendance Distribution – Majority of Matches <10K



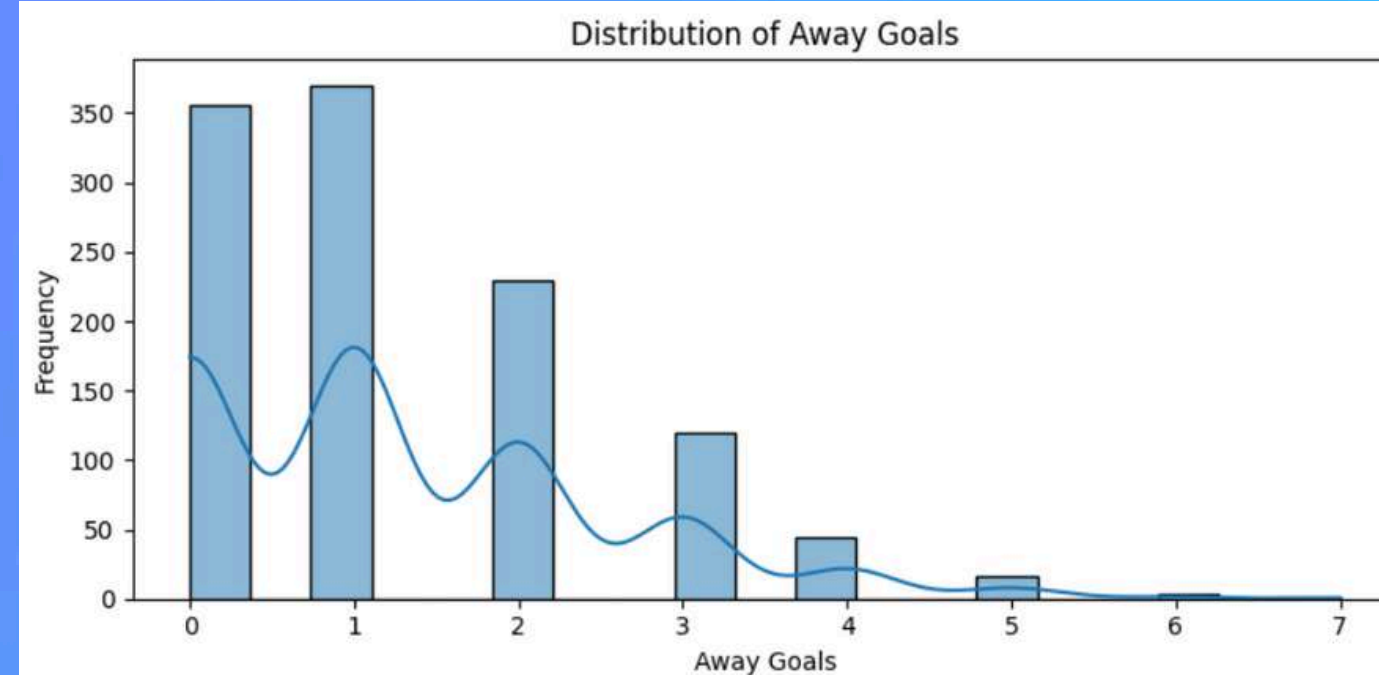
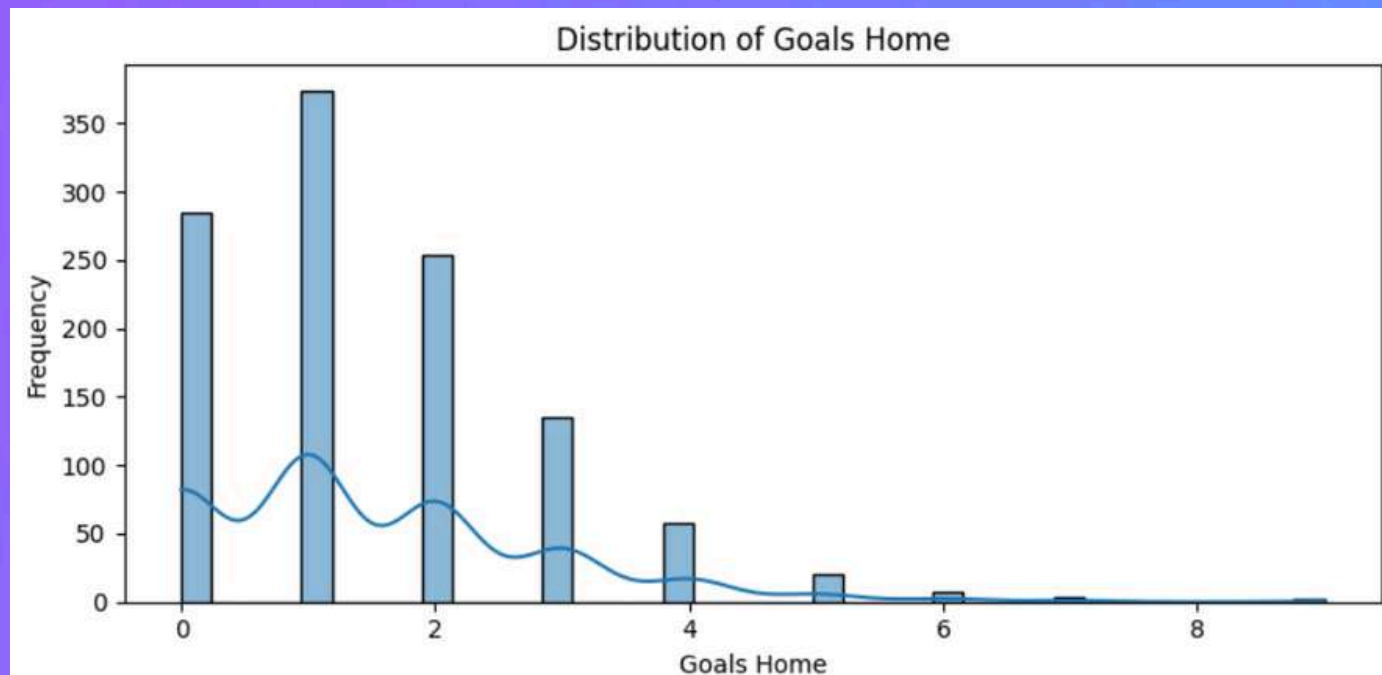
 Distribution of Attendance

- Mayoritas pertandingan <10.000 penonton
- Laga bergengsi tertentu >50.000 penonton
- Konsentrasi tinggi pada big match

EXPLORATORY  
DATA ANALYSIS



## Goals Distribution – Home Advantage Evident



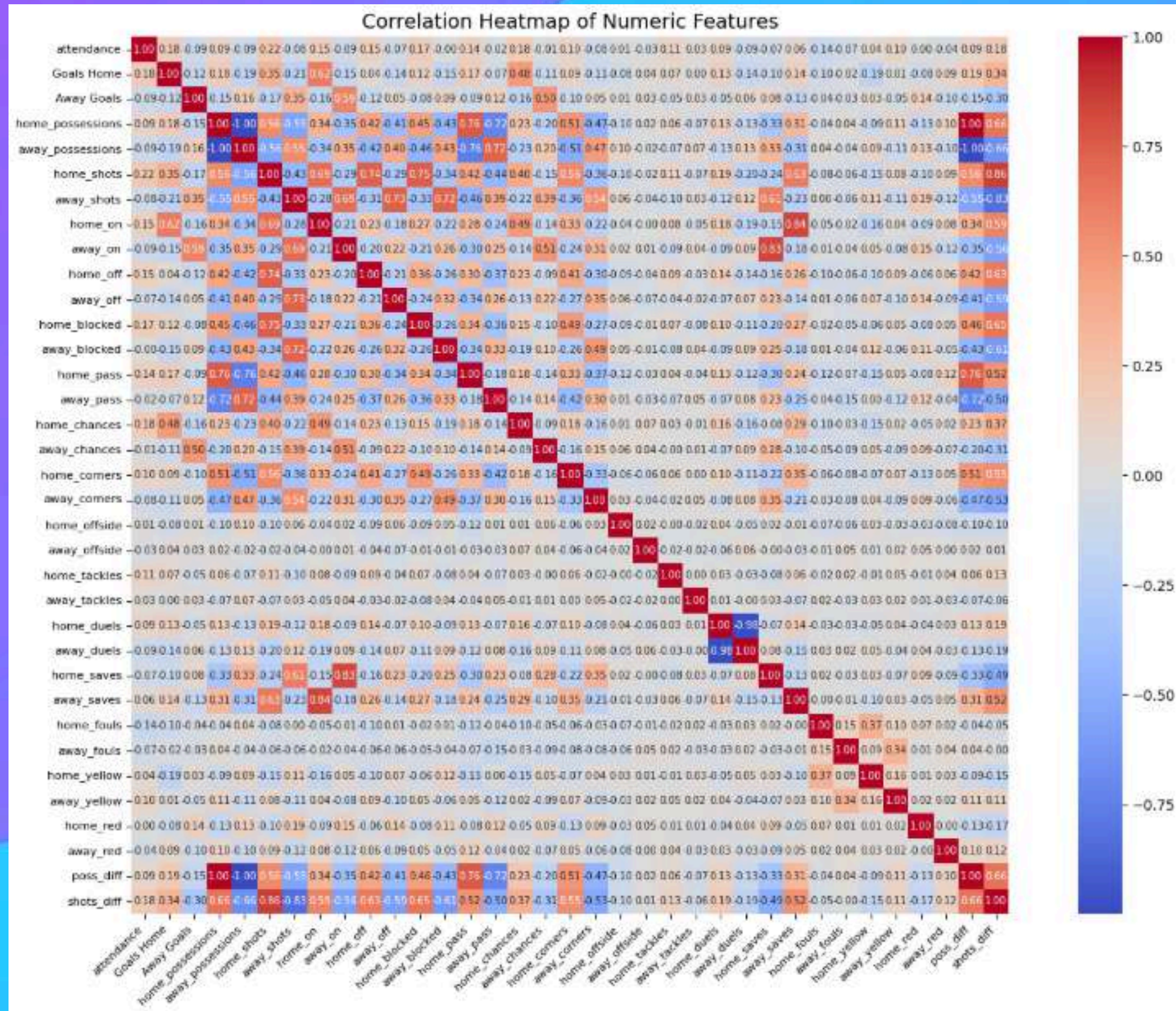
### ⚽ Distribution of Home & Away Goals

- Umumnya 0–2 gol per pertandingan
- 1 gol paling sering dicetak
- Laga dengan >3 gol jarang terjadi
- Tim tuan rumah sedikit lebih produktif → bukti home advantage

EXPLORATORY  
DATA ANALYSIS







## Correlation Heatmap

### Key Insights from Correlation

- Goals dipengaruhi oleh shots & shots on target
- Possession lebih terkait ke passing volume, bukan goals langsung
- Corners & chances mencerminkan aktivitas menyerang
- Defensive metrics & fouls → lebih lemah dalam prediksi hasil
- Selisih gol & selisih tembakan = indikator kuat hasil pertandingan

# EXPLORATORY DATA ANALYSIS





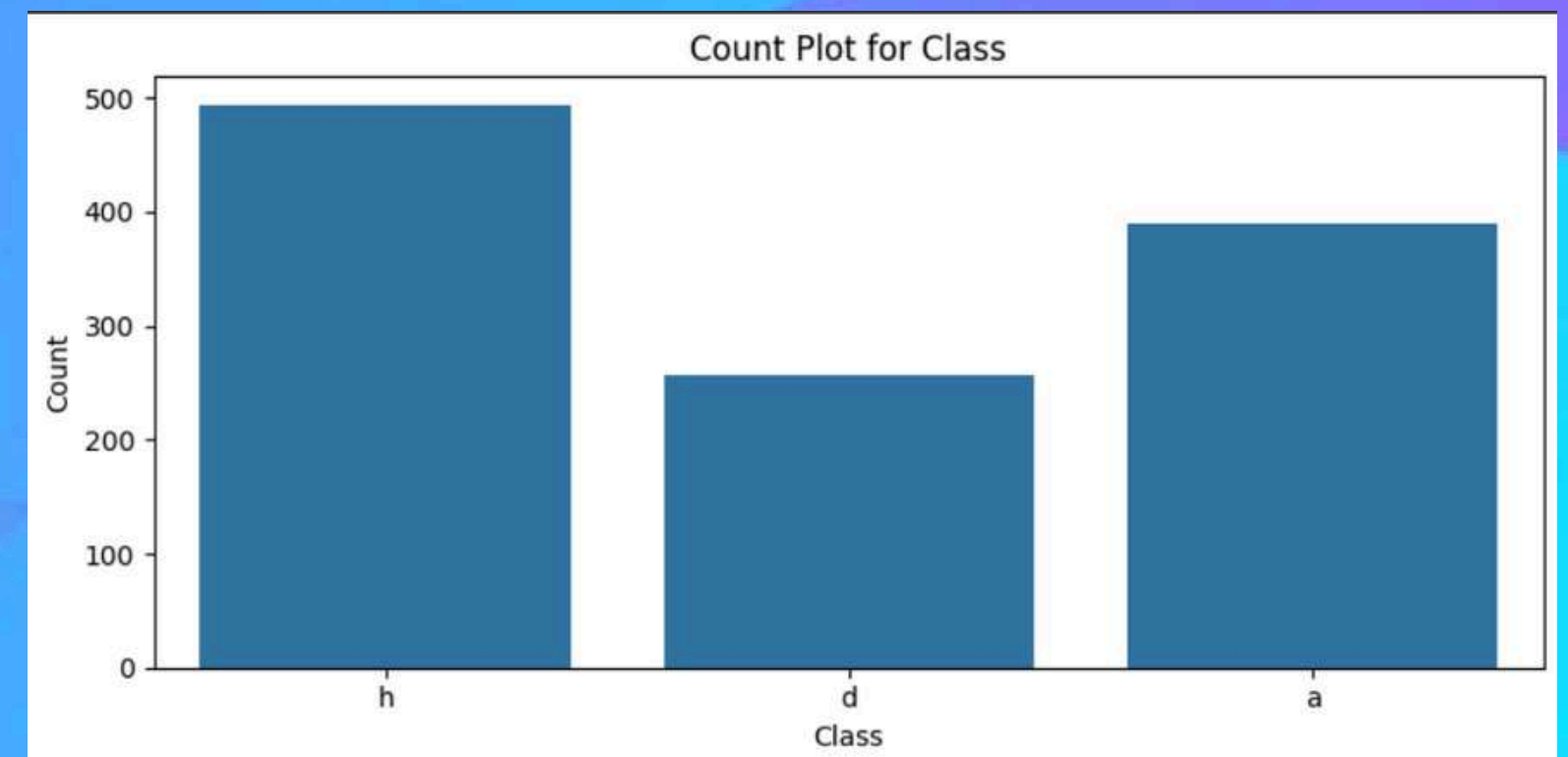
# EXPLORATORY DATA ANALYSIS

## Home/Away Win, & Draw Class

- Class h  $\approx 500$
- Class d  $\approx 250$
- Class a  $\approx 390$

Total  $\approx 500 + 250 + 390 = 1,140$

- Class h:  $500 / 1140 \approx 43.9\%$
- Class d:  $250 / 1140 \approx 21.9\%$
- Class a:  $390 / 1140 \approx 34.2\%$



# MODEL PREPARATION



## Split Train & Test Data

Melakukan split data dan test data

Training Accuracy: 0.61  
Test Accuracy: 0.60



## Build Multinomial Logistic Regression Model

Membangun model Multinomial Logistic Regression



## Evaluate The Model Using Accuracy Score

Memastikan Accuracy score dari model sesuai yang diharapkan sehingga model bisa digunakan

Accuracy of the Multinomial Logistic Regression model: 0.60

Classification Report:				
	precision	recall	f1-score	support
0	0.62	0.67	0.65	91
1	0.17	0.07	0.10	45
2	0.64	0.78	0.71	92
accuracy			0.60	228
macro avg	0.48	0.51	0.48	228
weighted avg	0.54	0.60	0.56	228



# MACHINE LEARNING MODEL

## Plotting The Confusion Matrix

### 1. Away Win (Class 0)

- Precision: ~65%
- Recall: ~65%
- F1-score: ~65%

Model cukup stabil, meski masih ada kecenderungan salah mengklasifikasi sebagai home win..

### 2. Draw (Class 1)

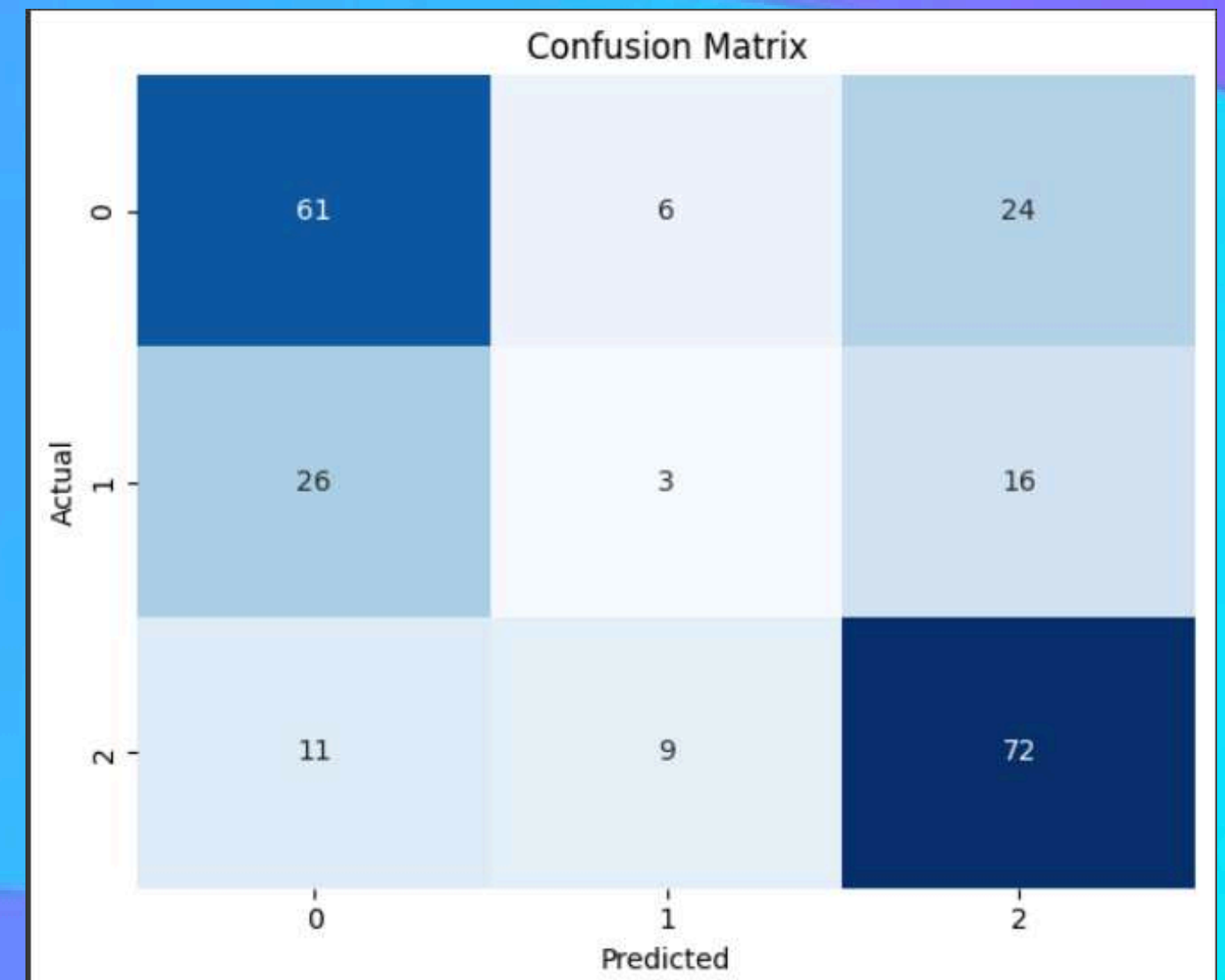
- Precision: ~20%.
- Recall: sangat rendah (~5–10%)
- F1-score: rendah (~8–12%)

Model hampir gagal menangkap pola draw. Ini jelas dampak dari class imbalance dan fitur yang kurang representatif untuk hasil imbang.

### 3. Home Win (Class 2)

- Precision: ~70%
- Recall: ~75%
- F1-score: ~72%

Model paling baik di kelas ini, selaras dengan fenomena home advantage dalam sepakbola.

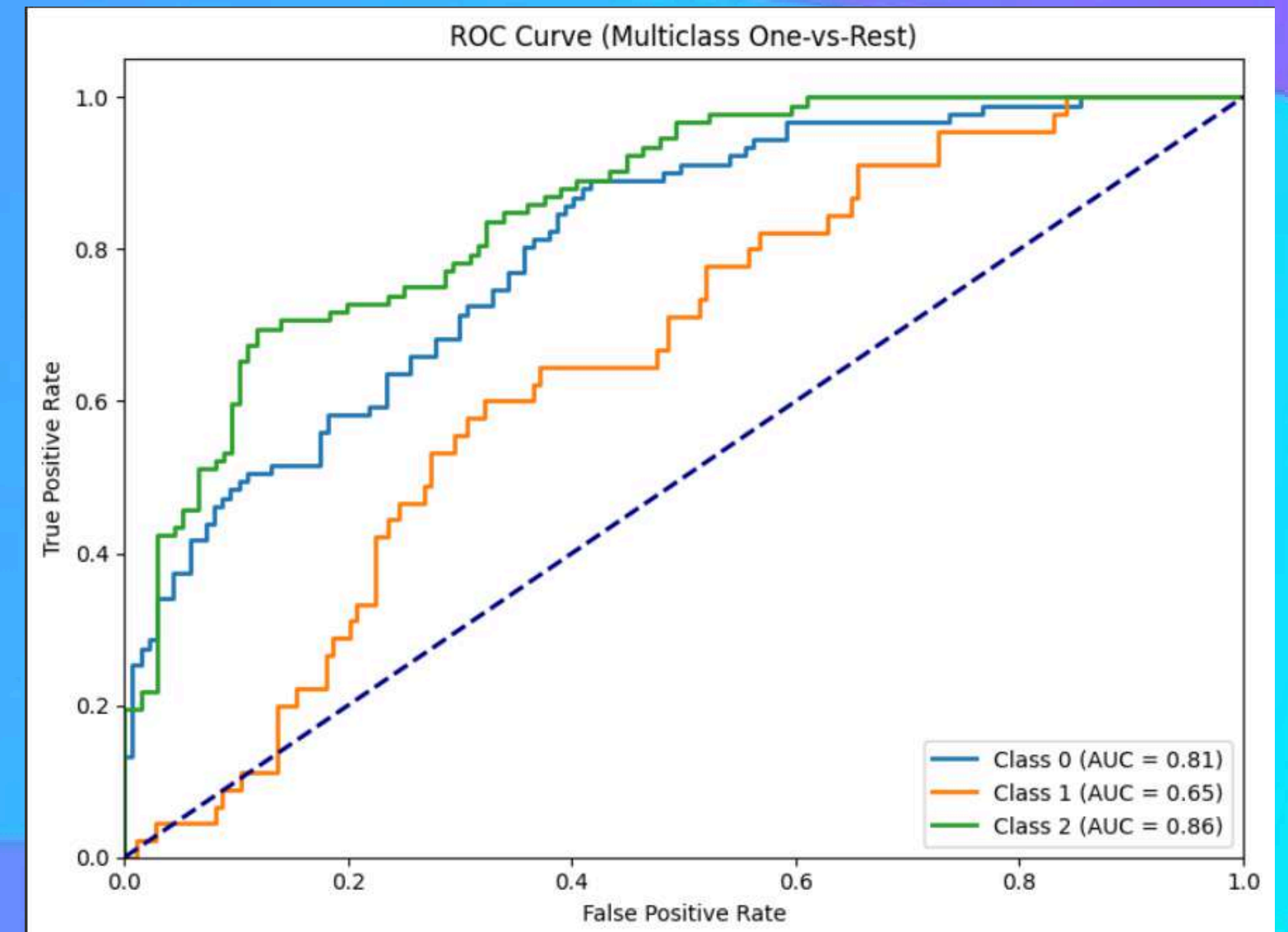


# MACHINE LEARNING MODEL

## Plot ROC Curve

### ROC Curve (Multiclass One-vs-Rest)

- Class 0 – Away Win (AUC = 0.81)
  - Model cukup andal membedakan kemenangan tim tamu.
  - $AUC > 0.8 \rightarrow$  prediksi relatif reliabel, meski masih ada ruang perbaikan.
- Class 1 – Draw (AUC = 0.65)
  - Performa terlemah; mendekati random guess (0.5).
  - Sering salah diklasifikasikan sebagai home win atau away win.
  - Menegaskan bahwa hasil draw adalah yang paling sulit diprediksi.
- Class 2 – Home Win (AUC = 0.86)
  - Performa sangat baik, mendekati AUC 0.9.
  - Konsisten dengan fenomena home advantage, sehingga lebih mudah dikenali oleh model.





# RECOMMENDATIONS

## Sportsbook & Betting

- Optimalkan odds home/away win (akurasi tinggi).
- Gunakan odds konservatif untuk draw (akurasi rendah).

## Sponsorship & Marketing

- Tunjukkan home advantage dalam kampanye sponsor.
- Gunakan data prediksi kemenangan kandang untuk promosi tiket.





# RECOMMENDATIONS

## Fan Engagement 📱

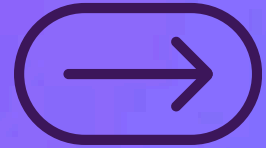
- Publikasikan prediksi model sebagai match preview.
- Jadikan draw sebagai faktor “unpredictable” untuk meningkatkan rasa penasaran fans.

## Future Development 🚀

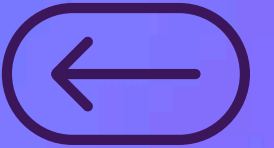
- Tambah variabel taktis (fouls, cards, expected goals).
- Coba model lebih kompleks (XGBoost, Random Forest, Neural Networks).







# THANK YOU



Thank you for your time! Football is more than just a game; it's a global phenomenon that unites people across cultures. Enjoy the game and keep the passion alive!



<https://wa.me/6282215082801>



[www.linkedin.com/in/umarabdurrahman](http://www.linkedin.com/in/umarabdurrahman)



[muhammadumarabdurrahman10@gmail.com](mailto:muhammadumarabdurrahman10@gmail.com)