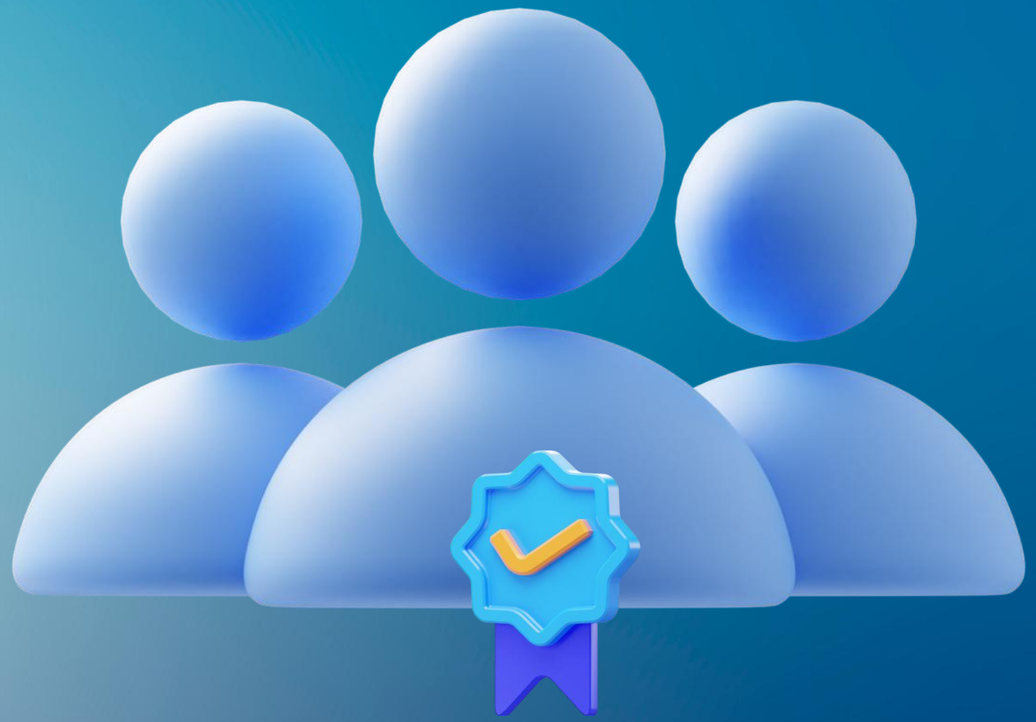
# Analisis Performa Pemain di 5 Liga Top Eropa

# KELOMPOK 14

1. 11423017-Anisetus Bambang Manalu
2. 11423043-Jonathan Prima Tamba
3. 11423044-Nathanael T.J Tampubolon



# Tujuan & Pertanyaan Riset

Tujuan:

- Memahami distribusi metrik performa kunci (khususnya per-90).
- Membandingkan performa antar posisi & liga.
- Menguji hubungan Expected Goals (xG) dengan Goals aktual.
- Mengevaluasi model sederhana untuk memprediksi Goals_per_90.

# Dataset & Data Collection

Dataset Diperoleh dari Kaggle dengan judul “All Football Players Stats in Top 5 Leagues 2023–2024

Rincian dataset:

- Jumlah baris: 2.852 pemain
- Jumlah kolom: 37 fitur
- Format: CSV
- Fitur utama: Goals_per_90, Expected_Goals_per_90, Assists_per_90, Minutes_Played, Position, Competition, Nation.

# Data Cleaning & Preprocessing (Advance)

Yang dilakukan:

1. Penghapusan Duplikat

Pada data kami tidak ada data yang duplikat

```
print("\nCek data duplikat:")
print(df.duplicated().sum())
```

```
Cek data duplikat:
0
```

2. Imputasi Missing Values

• Data numerik seperti Goals, Assists, xG, dan Minutes_Played diisi menggunakan nilai median agar tidak terpengaruh oleh outlier.

• Data kategorikal seperti Position dan Competition diisi menggunakan modus (mode) untuk menjaga konsistensi label.

```
for col in df.columns:
    if df[col].dtype in ['int64', 'float64']:
        df[col] = df[col].fillna(df[col].median())
    else:
        df[col] = df[col].fillna(df[col].mode()[0])
```

```
Jumlah missing value setelah cleaning:
0
```

3. Feature Scaling

Fitur numerik seperti Goals, Assists, xG, Minutes_Played, dan Goals_per_90 dinormalisasi menggunakan StandardScaler agar seluruh variabel berada dalam skala yang sebanding saat digunakan dalam model regresi.

4. Encoding categorical variables

• C(Position_group) dan C(Competition) di model OLS

• OneHotEncoder di pipeline sklearn untuk model LinearRegression dan LassoCV.

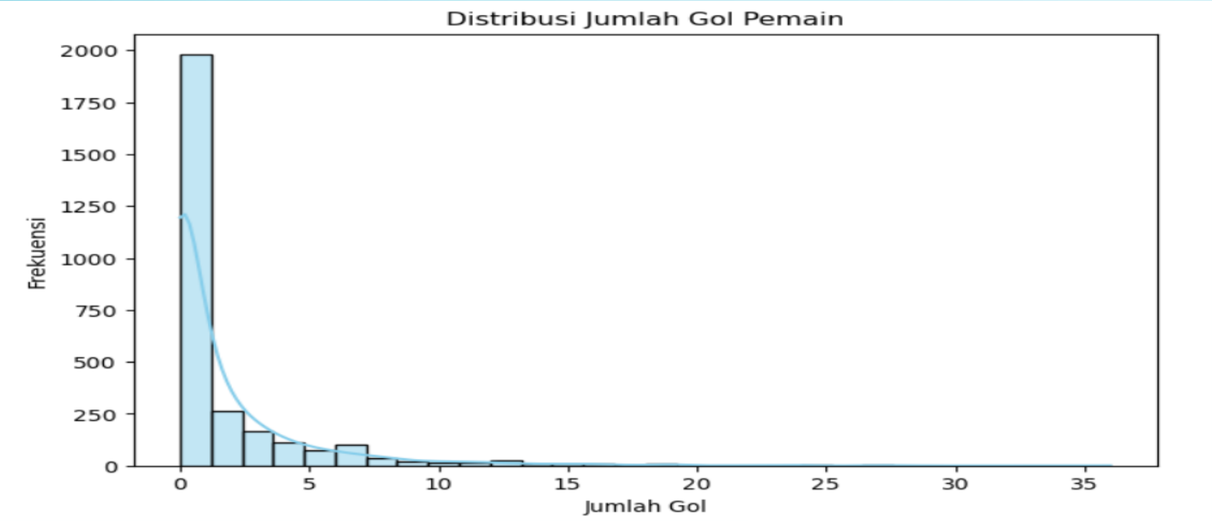
5. Handling outliers

```
df['Expected_Goals_per_90_clipped'] = df['Expected_Goals_per_90'].apply(lambda x: x if x > 0.1 else 0.1)
```

Clipping pada xG/90 untuk perhitungan efisiensi menghindari inflasi metrik

# VISUALISASI

1. a. Histogram Distibusi Jumlah Gol Pemain

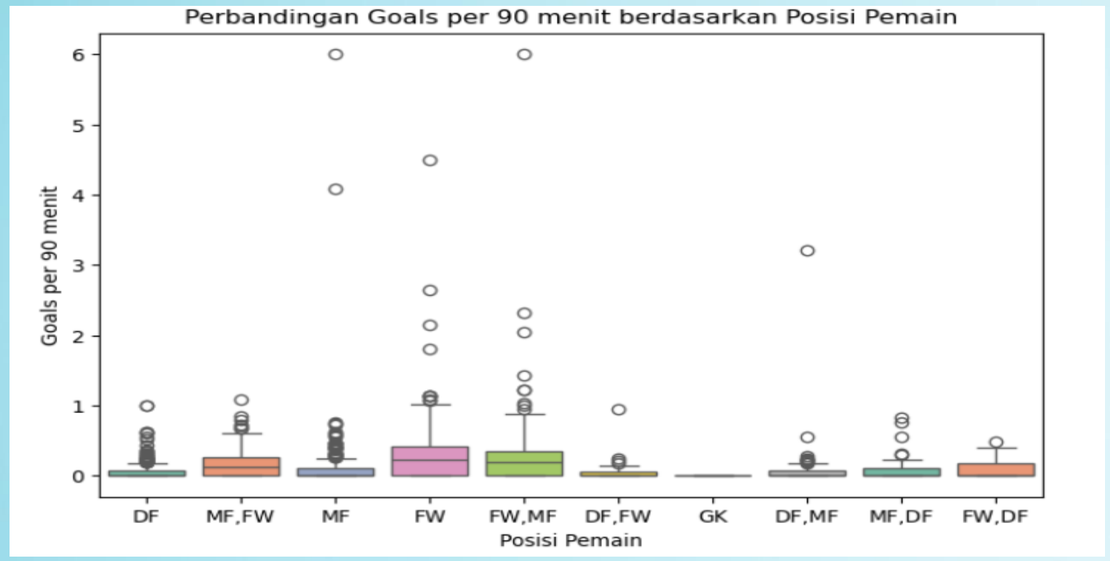



Insight:

- Distribusi right-skewed: mayoritas Goals_per_90 rendah; sedikit outlier tinggi.
- hanya segelintir pemain (terutama penyerang top) yang menyumbang jumlah gol sangat tinggi. Pola ini realistis dan umum pada kompetisi profesional.

# VISUALISASI

2. Boxplot Goals per 90 Menit Berdasarkan Posisi Pemain

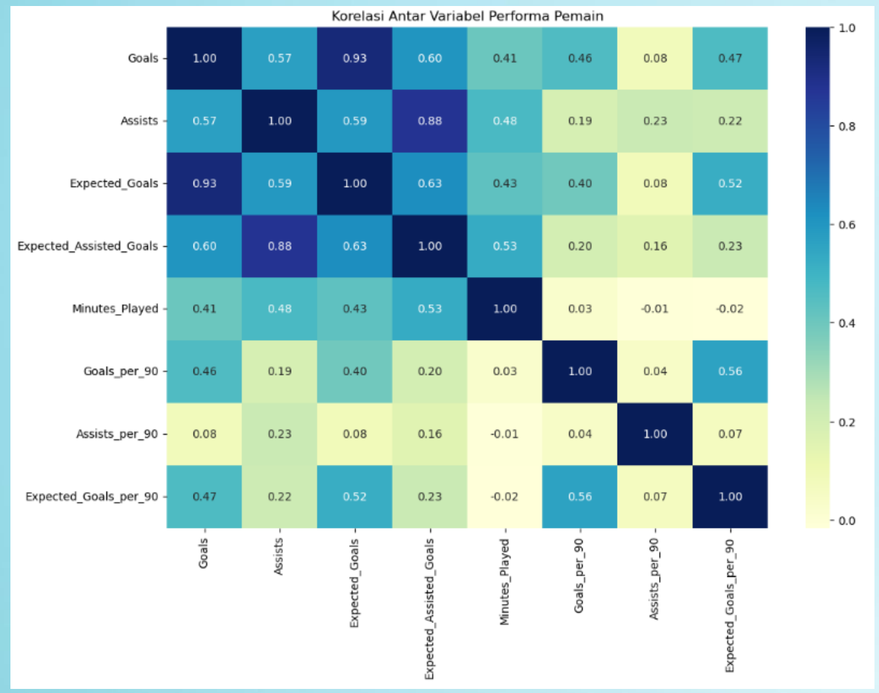



Insight:

Posisi Forward (FW) dan Forward/Midfielder (FW,MF) memiliki median Goals per 90 tertinggi, menunjukkan produktivitas tertinggi dalam mencetak gol, Sebaliknya, posisi Defender (DF) dan Goalkeeper (GK) menunjukkan nilai terendah — hasil ini sesuai dengan peran taktis masing-masing posisi.

# VISUALISASI

2. Heatmap Korelasi Variabel Performa Utama Pemain




## Insight:

Terdapat korelasi kuat antara Expected_Goals_per_90 dan Goals_per_90 ($r \approx 0.56$), yang menunjukkan bahwa peluang yang diciptakan berbanding lurus dengan hasil aktual. Hal ini juga menjadi dasar pemilihan xG sebagai variabel prediktor dalam model regresi.

# Analisis statistik

Analisis statistik dilakukan untuk menguji hubungan antar variabel serta perbedaan performa pemain berdasarkan posisi. Beberapa uji digunakan untuk memastikan hasil analisis valid baik pada data yang memenuhi maupun tidak memenuhi asumsi normalitas.

**1. Uji Korelasi Spearman**

Terdapat korelasi kuat antara Expected_Goals_per_90 dan Goals_per_90 (r ≈ 0.75), yang menunjukkan bahwa peluang yang diciptakan berbanding lurus dengan hasil aktual. Hal ini juga menjadi dasar pemilihan xG sebagai variabel prediktor dalam model regresi.

```
# Uji korelasi Spearman antara xG dan Goals
rho, pval = stats.spearmanr(df["Expected_Goals"], df["Goals"])
print(f"\nKorelasi Spearman antara Expected_Goals dan Goals: rho={rho:.3f}, p-value={pval:.4f}"
if pval < 0.05:
    print("→ Ada hubungan signifikan antara Expected Goals dan jumlah gol pemain.")
else:
    print("→ Tidak terdapat hubungan signifikan antara Expected Goals dan jumlah gol pemain.")
```

**Output**

```
Korelasi Spearman antara Expected_Goals dan Goals: rho=0.838, p-value=0.0000
→ Ada hubungan signifikan antara Expected Goals dan jumlah gol pemain.
```

**Penjelasan:**

Nilai korelasi Spearman sebesar ρ = 0.838 (p < 0.001) menunjukkan hubungan positif yang kuat.

Artinya, semakin tinggi peluang gol (xG), semakin tinggi pula gol aktual yang dicetak pemain.

# Analisis statistik

## 1. Uji Parametrik (ANOVA Satu Arah)

Uji One-Way ANOVA digunakan untuk melihat apakah terdapat perbedaan rata-rata Goals_per_90 antar kelompok posisi pemain (FW, MF, DF, GK).

```
# Uji beda performa antar posisi (ANOVA)
positions = df["Position"].unique()
groups = [df[df["Position"] == pos]["Goals_per_90"].dropna() for pos in positions]
f_stat, p_anova = stats.f_oneway(*groups)
print(f"\nHasil ANOVA untuk Goals_per_90 antar posisi: F={f_stat:.3f}, p-value={p_anova:.4f}")
if p_anova < 0.05:
    print("→ Terdapat perbedaan signifikan performa (Goals_per_90) antar posisi pemain.")
else:
    print("→ Tidak ada perbedaan signifikan performa antar posisi pemain.")
```

**Output**

```
Hasil ANOVA untuk Goals_per_90 antar posisi: F=41.160, p-value=0.0000
→ Terdapat perbedaan signifikan performa (Goals_per_90) antar posisi pemain.
```

**Penjelasan:**

NilaiNilai F = 41.160, p < 0.001, sehingga terdapat perbedaan signifikan rata-rata gol per posisi.

Hasil ini menunjukkan bahwa posisi pemain memang memengaruhi produktivitas mencetak gol.

# Analisis statistik

## 2. Uji Parametrik (ANOVA Satu Arah)

Uji One-Way ANOVA digunakan untuk melihat apakah terdapat perbedaan rata-rata Goals_per_90 antar kelompok posisi pemain (FW, MF, DF, GK).

```
# Uji beda performa antar posisi (ANOVA)
positions = df["Position"].unique()
groups = [df[df["Position"] == pos]["Goals_per_90"].dropna() for pos in positions]
f_stat, p_anova = stats.f_oneway(*groups)
print(f"\nHasil ANOVA untuk Goals_per_90 antar posisi: F={f_stat:.3f}, p-value={p_anova:.4f}")
if p_anova < 0.05:
    print("→ Terdapat perbedaan signifikan performa (Goals_per_90) antar posisi pemain.")
else:
    print("→ Tidak ada perbedaan signifikan performa antar posisi pemain.")
```

**Output**

```
Hasil ANOVA untuk Goals_per_90 antar posisi: F=41.160, p-value=0.0000
→ Terdapat perbedaan signifikan performa (Goals_per_90) antar posisi pemain.
```

**Penjelasan:**

NilaiNilai F = 41.160, p < 0.001, sehingga terdapat perbedaan signifikan rata-rata gol per posisi.

Hasil ini menunjukkan bahwa posisi pemain memang memengaruhi produktivitas mencetak gol.

# Analisis statistik

**3. Uji Non-Parametrik (Kruskal–Wallis dan Mann–Whitney U)**

Sebagai alternatif ANOVA yang tidak mengasumsikan normalitas, dilakukan uji Kruskal–Wallis terhadap Goals_per_90 antar posisi.

```
Position groups: ['DF' 'Other' 'MF' 'FW' 'FW,MF']
Kruskal-Wallis: H=381.6653, p=0.000000
→ Terdapat perbedaan signifikan antar grup posisi (non-parametrik).
```

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i (R_i - \bar{R})^2$$

**Penjelasan:**

Nilai H = 381.6653, p < 0.001, menandakan terdapat perbedaan signifikan antar posisi.

Uji lanjutan Mann–Whitney U dengan koreksi FDR (False Discovery Rate) menunjukkan perbedaan paling signifikan antara Forward (FW) dengan Defender (DF) dan Goalkeeper (GK).

**Penjelasan:**

Ndengan:

N = total seluruh observasi (pemain),

$kk$ = jumlah grup (posisi: DF, MF, FW, FW,MF, Other),

$ni$= jumlah pemain dalam grup ke-i,

$Ri$= rata-rata ranking dalam grup ke-I

,$R^-$ = rata-rata ranking keseluruhan.

# Demo

# Thank You