Hoang Anh (Benjamin) Nguyen

Data Preprocessing - Data Transformation

1. Data Transformation: Steps and Rationale After cleaning the dataset, I applied a series of transformations to prepare the data for machine learning models. The dataset originally contained 1,140 rows and 40 columns. Following transformations, it expanded to 66 columns due to encoding and feature engineering.

2. Standardization of numerical columns All numeric features (such as attendance, shots, passes, tackles, etc.) were standardized using StandardScaler, which transforms values to have mean = 0 and standard deviation = 1. This was done because many machine learning algorithms (e.g., logistic regression, SVM, K-means) are sensitive to feature scale, and standardization ensures that variables measured on different scales contribute equally to the model.

3. Encoding categorical variables Categorical columns cannot be directly interpreted by numerical models. To address this: The stadium column was one-hot encoded, creating separate binary columns for each stadium. This avoids imposing an artificial order on stadium names and allows the model to treat each stadium as an independent category. The Home Team and Away Team columns were label encoded, converting team names into integer values. Label encoding was chosen here to keep the dataset manageable in size (since one-hot encoding teams could create a very large number of columns).

4. Feature engineering (new derived features) Two new columns were added to capture useful match-related insights: Goal Difference (goal_diff) = Home Goals – Away Goals. This feature measures the relative dominance of the home team over the away team. Total Goals (total_goals) = Home Goals + Away Goals. This feature captures the overall scoring level of a match, which can be useful in tasks like predicting match excitement or fan attendance.

5. Saving transformed dataset The transformed dataset was saved as mydata_transformed.csv, with the final shape being 1,140 rows and 66 columns. This dataset is now in a fully numerical, standardized,
The screenshot of the code in the next page

Generate + Code + Markdown

```python
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler, OneHotEncoder, LabelEncoder

# 1. Load dataset
df = pd.read_csv("mydata.csv")
print("Original shape:", df.shape)

# 2. Normalize / Standardize numerical columns
# Select numeric columns
num_cols = df.select_dtypes(include=np.number).columns.tolist()

# Standardize (mean=0, std=1)
scaler = StandardScaler()
df[num_cols] = scaler.fit_transform(df[num_cols])

print("\nStandardized numerical columns")

# 3. Encode categorical variables
cat_cols = df.select_dtypes(include="object").columns.tolist()

#One-hot encode "stadium" and label encode "Home Team" + "Away Team"
df_encoded = df.copy()

# One-hot encode "stadium"
if "stadium" in cat_cols:
    ohe = OneHotEncoder(sparse_output=False, drop="first")
    ohe_array = ohe.fit_transform(df[["stadium"]])
    ohe_df = pd.DataFrame(ohe_array, columns=ohe.get_feature_names_out(["stadium"]))
    df_encoded = pd.concat([df_encoded.drop(columns=["stadium"]), ohe_df], axis=1)
    print("One-hot encoded 'stadium'")

for col in ["Home Team", "Away Team"]:
    if col in df.columns:
        le = LabelEncoder()
        df_encoded[col] = le.fit_transform(df[col])
        print(f"Label encoded '{col}'")

# 4. Create at least two new features
# Goal Difference (Home - Away)
if "Goals Home" in df.columns and "Away Goals" in df.columns:
    df_encoded["goal_diff"] = df["Goals Home"] - df["Away Goals"]

# Total Goals
if "Goals Home" in df.columns and "Away Goals" in df.columns:
    df_encoded["total_goals"] = df["Goals Home"] + df["Away Goals"]

print("Created new features: 'goal_diff', 'total_goals'")

# 5. Save transformed dataset
df_encoded.to_csv("mydata_transformed.csv", index=False)
print("\nTransformed dataset saved as 'mydata_transformed.csv'")
print("Final shape:", df_encoded.shape)
```

✓ 0.1s Python

```
Original shape: (1140, 40)

Standardized numerical columns
One-hot encoded 'stadium'
Label encoded 'Home Team'
Label encoded 'Away Team'
Created new features: 'goal_diff', 'total_goals'

Transformed dataset saved as 'mydata_transformed.csv'
Final shape: (1140, 66)
```

Spaces: 4 LF {} Finish Setup Cell 1 of 1