

1. Objetivo General

Realizar un análisis de los partidos de futbol entre dos países para poder predecir quién podría ser el ganador en su próximo encuentro.

2. Propuesta y Organización CRISP-DM

2.1. Business Understanding (Entendimiento del negocio/problema)

- Define claramente cuál es el problema de negocio o de investigación.

Se necesita saber el posible ganador del próximo partido de Ecuador vs Brasil.

- Describe el contexto: ¿cuál es la hipótesis principal?, ¿qué necesitas comprobar o resolver?, ¿cuál es tu pregunta de investigación?

Tengo una base con los resultados de todos los partidos de futbol desde 1872 hasta 2025

En base a esta información se requiere predecir el próximo ganador en el partido que se va a llevar a cabo entre Ecuador y Brasil en junio de 2025

Mi pregunta sería quién será el ganador del siguiente partido entre Ecuador y Brasil

- Explica el valor que proporcionará tu producto de datos MVP (¿por qué es relevante para tu institución, empresa o investigación?).

Es importante para saber si vamos a poder ir al mundial 2026

2.2. Data Understanding (Entendimiento de los datos)

- Explica de dónde provienen tus datos (fuentes, tipo de datos, frecuencia de actualización, etc.).

La información proviene de la plataforma Kaggle www.kaggle.com

Esta información se actualiza semanal o diariamente depende de los partidos llevados a cabo.

La base utilizada se la bajó un día después del partido Ecuador vs Brasil, el 26 de marzo de 2025, y ya tenía la información actualizada del 25 de marzo de 2025.

El archivo es .csv

Los tipos de datos son:

```
dtype: object
date          datetime64[ns]
home_team      object
away_team      object
home_score     int64
away_score     int64
tournament     object
city           object
country        object
neutral        bool
```

- Describe de manera general las variables y su posible relevancia o relación con el problema.

date: es la fecha en la que se llevó a cabo el partido

home_team: es el equipo que juega en casa

away_team: es el equipo que juega de visitante

home_score: es el puntaje del equipo que juega en casa

away_score: es el puntaje del equipo que juega de visitante

tournament: el torneo en el que participaron

city: la ciudad donde se llevó a cabo el encuentro

country: el país donde se llevó a cabo el encuentro

neutral: indica si el partido se llevó a cabo en un lugar neutral

De los campos indicados `home_team`, `away_team`, `home_score` y `away_score` son claves para el análisis

- Identifica los posibles desafíos: datos faltantes, duplicados, inconsistencias, calidad y confiabilidad.

Los datos son confiables y sin inconsistencias, se validaron los datos de los partidos para verificar su veracidad. Tampoco tiene duplicidad

2.3. Data Preparation (Preparación de los datos)

- Lista las tareas de limpieza y transformación necesarias (Data Wrangling).

Se debe revisar el tipo de dato de los campos

Se debe verificar que no existan duplicados con la función `duplicated`

Se debe contar las filas y columnas

Se debe validar valores faltantes con la función `isnull`

Se debe asegurar de que la fecha sea tipo fecha con la función `to_datetime`

Se debe ordenar por fecha con la función `sort_values`

Se debe crear promedios de puntajes con la función `mean`

Se crea campos, año y mes

- Documenta la extracción de datos y la manipulación para llegar a la forma deseada.

Se crea las variables **resultado**, **oponente**, **partidos_previos** y **porcentaje_ganado** para usarlos como features.

Se crea las columnas dummies en base a **oponente** para que se creen columnas por cada país con el que Ecuador ha jugado.

Se transforma el campo resultado que tiene valores categóricos en valores numéricos: columna 'resultado' son 'gana', 'pierde' y 'empata',

2.4. Modeling (Modelado)

Para el modelo se usa Clasificación

Regresión Logística : LogisticRegression

```
from sklearn.linear_model import LogisticRegression

# Inicializa y entrena el modelo de Regresión Logística
logreg = LogisticRegression(solver='liblinear', max_iter=1000, random_state=42)
logreg.fit(X_train, y_train)

# Hacer predicciones con el set de pruebas
y_pred = logreg.predict(X_test)
y_pred_proba = logreg.predict_proba(X_test)[:, 1] # Probabilidad de ganar
```

2.5. Evaluation (Evaluación)

- Particionamiento de Datos

```

import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder

# Veamos las features y realizamos one-hot encoding en 'oponente'
features = ['partidos_previos', 'porcentaje_ganado']
X = pd.get_dummies(df_ecuador, columns=['oponente'], drop_first=True)
X = X[features + [col for col in X.columns if 'oponente_' in col]]

# Convertir 'resultado' a numero
le = LabelEncoder()
y = le.fit_transform(df_ecuador['resultado'])

# Separar datos en entrenamiento y prueba
#test_size=0.2: Indica que el 20% de los datos se asignarán al conjunto de prueba (X_test, y_test) y el 80% restante al conjunto de entrenamiento (X_train, y_train)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)

print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)
display(X_train.head())
display(X_test.head())

```

(463, 65) (116, 65) (463,) (116,)

La división de datos se realiza una sola vez, utilizando `train_test_split`, para crear los conjuntos de entrenamiento y prueba.

- Métricas de desempeño

Análisis: Accuracy: El modelo tiene una exactitud del 51.72%, lo que significa que predice correctamente un poco más de la mitad de las veces. Esto puede indicar que el modelo no es muy preciso y hay margen de mejora. Precision: Un valor de 0.3338 sugiere que cuando el modelo predice gana, solo acierta alrededor del 33.38% de las veces. Esto indica una alta tasa de falsos positivos. Recall: Con un valor de 0.4415, el modelo está capturando alrededor del 44.15% de los casos positivos reales. Esto implica que hay una cantidad significativa de falsos negativos. F1-score: Un valor de 0.3755 indica un rendimiento moderado, pero con margen de mejora en la precisión y la recuperación. ROC AUC Score: Un valor de 0.6136 sugiere un rendimiento aceptable, mejor que una clasificación aleatoria (0.5), pero aún con espacio para mejorar. Un valor más cercano a 1 indicaría un mejor rendimiento.

```

from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, roc_auc_score
from sklearn.metrics import classification_report

# Evaluación del módulo
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred, average='macro', zero_division=0)
recall = recall_score(y_test, y_pred, average='macro', zero_division=0)
f1 = f1_score(y_test, y_pred, average='macro', zero_division=0)

# Recalcula y_pred_proba para obtener probabilidades para todas las clases
y_pred_proba = logreg.predict_proba(X_test)
roc_auc = roc_auc_score(y_test, y_pred_proba, multi_class='ovr')

|
print(f"Accuracy: {accuracy:.4f}")
print(f"Precision: {precision:.4f}")
print(f"Recall: {recall:.4f}")
print(f"F1-score: {f1:.4f}")
print(f"ROC AUC Score: {roc_auc:.4f}")

# Reporte
print("\nReporte:\n", classification_report(y_test, y_pred))

```






Accuracy: 0.5172
 Precision: 0.3338
 Recall: 0.4415
 F1-score: 0.3755
 ROC AUC Score: 0.6136

2.6. Organización de trabajo


📁 / Maestría / Fundamentos-DS / partidos-de-futbol /

<input type="checkbox"/> Name	Modified	File Size
📁 DATASET	10 days ago	
📁 NOTEBOOKS	19 minutes ago	
📁 REPORTS	10 days ago	
📄 LICENSE	21 days ago	1.1 KB
📄 README.md	10 days ago	35 B

📁 / Maestría / Fundamentos-DS / partidos-de-futbol / NOTEBOOKS /

 Name	Modified	File Size
•  DATAWRANGLING.ipynb	2 hours ago	4.6 MB
•  EDA.ipynb	2 hours ago	21 KB
•  Ingeniería_de_caracteristicas_parti...	1 hour ago	35.1 KB
✓ •  Modelo.ipynb	18 minutes ago	70.6 KB




📁 / Maestría / Fundamentos-DS / partidos-de-futbol / DATASET /

 Name	Modified	File Size
📁 CLEAN	23 hours ago	
📁 ROW	10 days ago	

📁 / Maestría / Fundamentos-DS / partidos-de-futbol / DATASET / CLEAN /

 Name	Modified	File Size
 df_ecuador.csv	23 hours ago	61 KB

📁 / Maestría / Fundamentos-DS / partidos-de-futbol / REPORTS /

 Name	Modified	File Size
✓  CRISP-DM-Completa.pdf	23 seconds ago	65.4 KB
 Propuesta CRISP-DM.pdf	10 days ago	65.4 KB