

# **“F1 DATA ANALYTICS: EVALUACIÓN DE LA COMPETITIVIDAD ENTRE EQUIPOS MEDIANTE TÉCNICAS DE CIENCIA DE DATOS”**

**Diana Luz Hernandez Torres**

**Jaime Alejandro Romero Sierra**  
**Introducción a ciencia de datos**  
**23 de noviembre del 2025**

# INTRODUCCION

## Objetivos del proyecto

El objetivo del proyecto es predecir con alta precisión los puntos obtenidos por los pilotos, identificando las variables más influyentes y proporcionando información útil para la toma de decisiones estratégicas en el análisis del desempeño deportivo.

## ¿Por que es importante resolver o estudiar esta problemática?

Es importante estudiar esta problemática es fundamental porque el desempeño de los pilotos en competencias deportivas está determinado por una combinación de factores complejos, como su historial de resultados, consistencia, condiciones de carrera y variables externas que pueden afectar el rendimiento. Comprender cómo se relacionan estas variables y cómo influyen en la obtención de puntos permite anticipar resultados y optimizar estrategias de entrenamiento y participación.

Además, contar con predicciones precisas sobre los puntos de los pilotos facilita la toma de decisiones estratégicas tanto para equipos como para organizadores, desde la planificación de recursos hasta la identificación de áreas de mejora individuales o colectivas. Este análisis también permite comparar rendimientos, establecer benchmarks y diseñar estrategias basadas en evidencia, reduciendo la incertidumbre asociada a la toma de decisiones en un entorno competitivo.

Por otro lado, estudiar esta problemática contribuye al uso efectivo de datos para generar conocimiento práctico, fomentando un enfoque más analítico en el deporte y potenciando la eficiencia y efectividad de las acciones que se implementen. En resumen, comprender y predecir el desempeño de los pilotos no solo mejora el rendimiento deportivo, sino que también aporta herramientas valiosas para la gestión, la planificación y la competitividad en el ámbito deportivo profesional.

# INTRODUCCION

## Fuente de datos

Base de Datos Principal: F1 race by race (df\_limpio.csv)

Origen: Datos históricos de resultados de carreras de la Fórmula 1, consolidados a nivel de piloto y sesión de carrera.

Cantidad de Datos: Contiene un gran volumen de registros que cubren un período extenso, desde la temporada 1983 hasta la temporada 2021.

Principales Características (Variables Clave): El dataset es amplio en variables que permiten evaluar el rendimiento y los factores contextuales de cada carrera.

Incluye:

Datos de Sesión: Sesión (Año), Recorrido (Número de carrera en el año), y Circuito (Nombre del circuito).

Rendimiento del Piloto: Piloto, Nacionalidad, Edad, Grilla (Posición de partida), Podio, Puntos\_piloto, Victorias\_piloto, y Posicion\_piloto (Posición final en el campeonato).

Rendimiento del Equipo: Constructores (Equipo), Puntos\_constructores, Victorias\_constructores, y Constructores\_posicion (Posición final del equipo en el campeonato).

Métricas de Tiempo: tiempo\_clasificación (Tiempo de clasificación expresado en segundos de diferencia respecto al pole position).

Condiciones Ambientales: Variables booleanas para registrar el clima de la carrera, incluyendo Clima Calido, Clima frio, Clima seco, Clima lluvioso, y Clima nublado.

# METODOLOGÍA

## Proceso de limpieza

Para este proyecto se partió del archivo df\_sucio.csv, el cual contenía información relacionada con pilotos, constructores, clima, sesiones y resultados de carreras de Fórmula 1. El proceso de limpieza consistió en las siguientes etapas:

df\_sucio.csv

### 1. Identificación y manejo de datos

La mayoría de las columnas presentaron 0% de valores faltantes.

Se evalúa cada columna para poder identificar el tipo de valor

En el caso de columnas categóricas con valores vacíos (como clima o sesión), se aplicó imputación con la moda (la categoría más frecuente), técnica adecuada para variable.

```
#Se verifican las columnas con las que se cuenta
df.columns

Index(['Unnamed: 0', 'season', 'round', 'circuit_id', 'weather_warm',
       'weather_cold', 'weather_dry', 'weather_wet', 'weather_cloudy',
       'driver', 'nationality', 'constructor', 'grid', 'podium',
       'driver_points', 'driver_wins', 'driver_standings_pos',
       'constructor_points', 'constructor_wins', 'constructor_standings_pos',
       'qualifying_time', 'driver_age'],
      dtype='object')
```

```
#Tamaño del dataframe original
df.shape
```

(16641, 22)



# METODOLOGIA

## Etapas

### 2. Revisión y eliminación de columnas o datos irrelevantes

- La columna Unnamed: 0 correspondía a un índice generado automáticamente al exportar el archivo.
- Se eliminó los datos como nan ya que no aporta información útil al análisis.

```
#Importar librerias y cargar base de datos sucia
import pandas as pd
df = pd.read_csv("https://raw.githubusercontent.com/dianahdeztorres13/pixel/fh/main/j.csv")
df

Unnamed: 0  season  round  circuit_id  weather_warm  weather_cold  weather_dry  weather_wet  weather_cloudy
0          14.0  1983.0     1.0  Auto%#      False        0.0       True        0.0      False
1           5.0  1983.0     1.0  jacarepagua    False        0.0       True        0.0      False
2           3.0  1983.0     NaN  jacarepagua    False        0.0       True        0.0      False
3           0.0  1983.0     1.0  jacarepagua    False        0.0       True        0.0      NaN
4           6.0  1983.0     1.0  jacarepagua    False        0.0       True        0.0      False

df2["Sesion"].unique()

array([1983.,  nan,  1984.,  1985.,  1986.,  1987.,  1988.,  1989.,
       1990.,  1991.,  1992.,  1993.,  1994.,  1995.,  1996.,  1997.,
       1998.,  1999.,  2000.,  2001.,  2002.,  2003.,  2004.,  2005.,
       2006.,  2007.,  2008.,  2009.,  2010.,  2011.,  2012.,  2013.,
       2014.,  2015.,  2016.,  2017.,  2018.,  2019.,  2020.,  2021.])

df2["Sesion"].isnull().sum()

np.int64(481)
```

# METODOLOGIA

## Etapas

### 3. Eliminación de duplicados

Se ejecutó:

```
df.drop_duplicates(inplace=True)
```

- Se detectó un pequeño número de filas duplicadas, probablemente provenientes de registros repetidos de sesiones.
- Fueron eliminadas para evitar sesgos en el análisis.

### 4. Detección y tratamiento de outliers

Columnas outliers:

- driver\_points
- constructor\_points
- driver\_age
- qualifying\_time
- driver\_wins / constructor\_wins

En este proyecto decidí mantener los outliers, dado que:

- En F1, valores extremos representan carreras excepcionales.
- Eliminarlos podría borrar información importante del rendimiento real.

## Análisis Exploratorio de Datos (EDA)

### Descripción general de los datos

El dataset contiene 7517 registros y 22 variables.

### Tipos de variables

#### Variables numéricas

- 1.Sesion
- 2.Recorrido
- 3.Clima frio
- 4.Clima lluvioso
- 5.Grilla
- 6.Podio
- 7.Puntos\_piloto
- 8.Victorias\_piloto
- 9.Posicion\_piloto
- 10.Puntos\_constructores
- 11.Victorias\_constructores
- 12.Constructores\_posicion
- 13.tiempo\_clasificación
- 14.Edad\_piloto

# METODOLOGIA

Variables categóricas

- 1.Círculo
- 2.Piloto
- 3.Nacionalidad
- 4.Constructores

Variables booleanas

- 1.Clima Calido
- 2.Clima seco
- 3.Clima nublado

variables numéricas

Variable: Sesión	
Count	7517
Mean	2001.672875
Std	11.235319
Min	1983
25%	1992
50%	2001
75%	2012
Max	2021
Mediana: 2001.0	

## Resumen estadístico

Variable: Recorrido	
Count	7517
Mean	9.22336
Std	5.120917
Min	1
25%	5
50%	9
75%	13
Max	21
Mediana: 9	

Variable: Clima lluvioso	
Count	7517
Mean	0.092324
Std	0.289502
Min	0
25%	0
50%	0
75%	0
Max	1
Mediana: 0	

Variable: Clima frio	
Count	7517
Mean	0.019689
Std	0.138937
Min	1
25%	0
50%	0
75%	0
Max	1
Mediana: 0	

Variable: Posicion_piloto	
Count	7517
Mean	10.771984
Std	7.784203
Min	0
25%	4
50%	10
75%	17
Max	13
Mediana: 10	

Variable: Puntos_piloto	
Count	7517
Mean	20.217374
Std	42.391462
Min	0
25%	0
50%	3
75%	20
Max	387
Mediana: 3	

Variable: Victorias_piloto	
Count	7517
Mean	0.369961
Std	1.160379
Min	0
25%	0
50%	0
75%	0
Max	13
Mediana: 0	

Variable: Posicion_constructores	
Count	7517
Mean	41.260476
Std	83.762834
Min	0
25%	0
50%	43
75%	8
Max	701
Mediana: 8	

Variable: tiempo_clasificación	
Count	7517
Mean	2.486364
Std	2.962189
Min	-77
25%	1
50%	2.1
75%	3.5
Max	82.3
Mediana: 2.099999999999943	

Variable: Edad_piloto	
Count	7517
Mean	28.604097
Std	4.76527
Min	17
25%	25
50%	28
75%	32
Max	43
Mediana: 28.0	

Variable: Victorias_constructores	
Count	7517
Mean	0.767061
Std	1.994642
Min	0
25%	0
50%	0
75%	0
Max	17
Mediana: 0	

# METODOLOGÍA

## Resumen estadístico

variables categóricas

Variable: Circuito

Count	7517
Unique	50
Top	monza
Freq	463

Variable: Piloto

Count	7309
Unique	226
Top	raikkonen
Freq	170

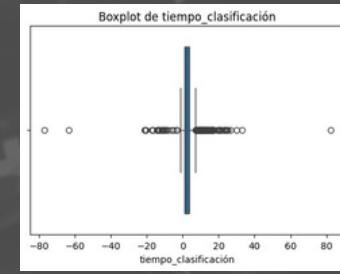
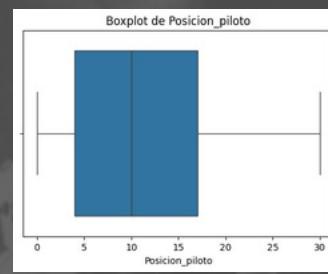
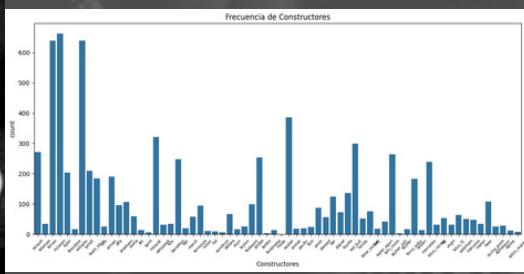
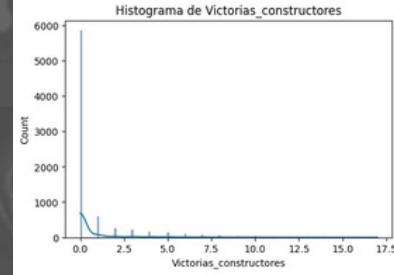
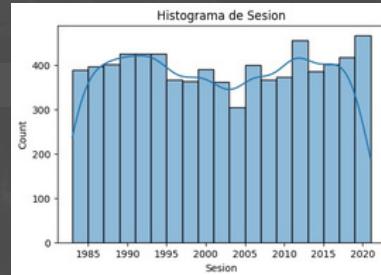
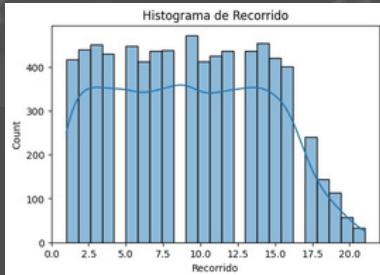
Variable: Nacionalidad

Count	7517
Unique	34
Top	British
Freq	1112

Variable: Constructores

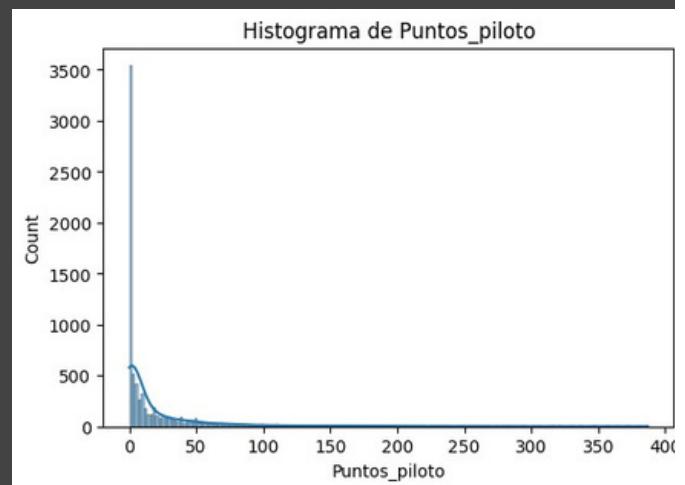
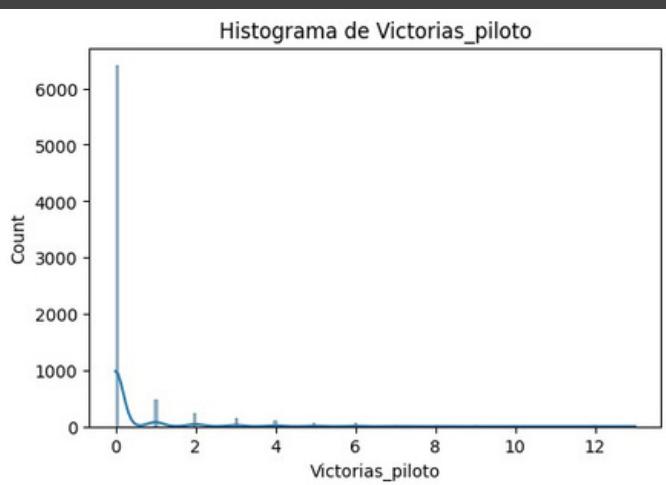
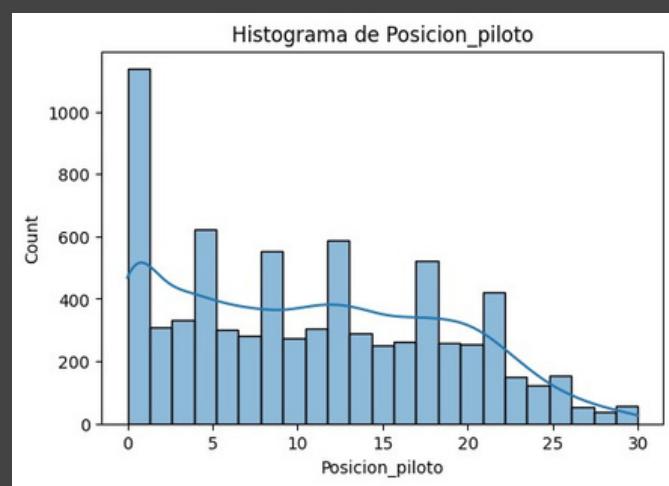
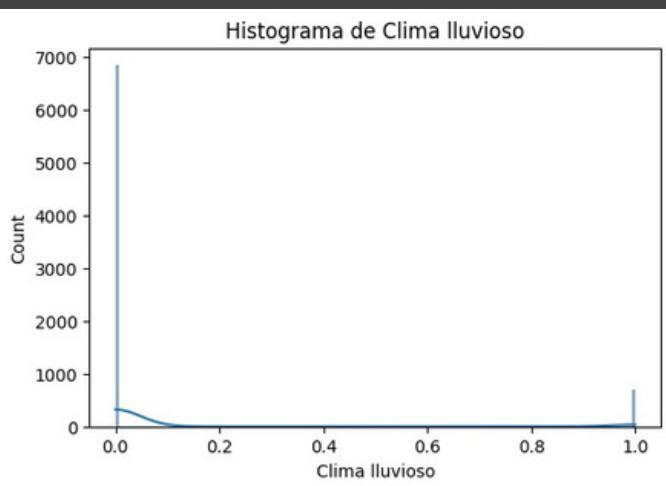
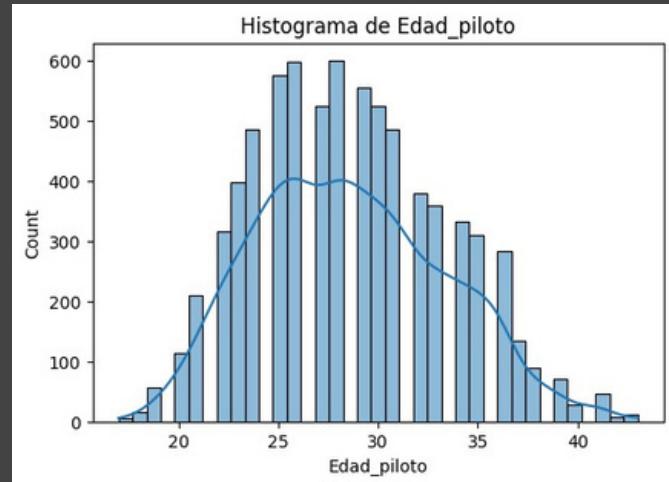
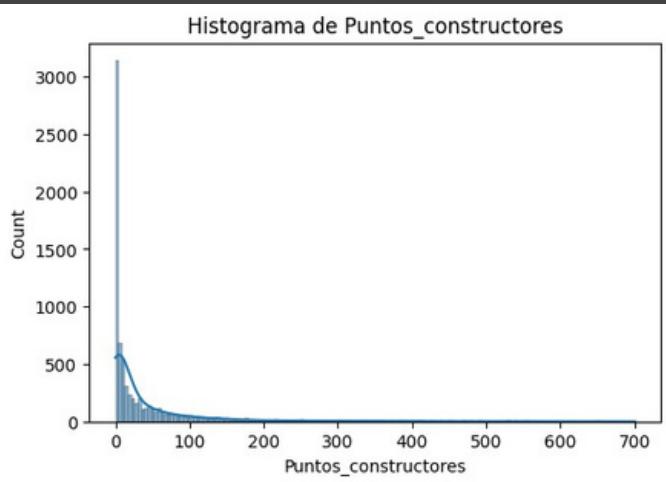
Count	7517
Unique	66
Top	mclaren
Freq	663

## Visualización y Distribución de Variables Individuales



# METODOLOGIA

## Visualización y Distribución de Variables Individuales



# ANÁLISIS DE VALORES ATÍPICOS (OUTLIERS)

## Identificación

Para detectar valores atípicos en las variables numéricas se utilizó

```
#Outliers
numeric_cols = df.select_dtypes(include="number").columns.tolist()
results = {}

for columna in numeric_cols:
    Q1 = df[columna].quantile(0.25)
    Q3 = df[columna].quantile(0.75)
    IQR = Q3 - Q1
    outliers = df[
        (df[columna] < (Q1 - 1.5 * IQR)) |
        (df[columna] > (Q3 + 1.5 * IQR))
    ]

    results[columna] = {
        "Q1": Q1,
        "Q3": Q3,
        "IQR": IQR,
        "Outliers": len(outliers)
    }

results
```

Los resultados fueron los siguientes

Resumen de outliers por variable

Variable	Outliers encontrados
Clima frío	148
Clima lluvioso	694
Puntos_pilotos	865
Victorias_piloto	1115
Puntos_constructores	826
Victorias_constructores	1665
Tiempo_clasificación	275
Edad_piloto	11
Constructores_posicion	2

# ANÁLISIS DE VALORES ATÍPICOS (OUTLIERS)

Observaciones, tratamiento e interpretación

## Observación:

Varias variables binarias o con pocos valores posibles muestran gran cantidad de outliers debido a que su IQR = 0. Esto es normal y no representa valores extremos reales.

## Interpretación

Las variables como Victorias\_piloto y Victorias\_constructores tienen muchos "outlier" porque casi siempre valen 0, y cualquier valor mayor se marca como extremo.

Variables como Edad\_piloto, tiempo\_clasificación y Constructores\_posición sí presentan outliers genuinos.

## Tratamiento:

Este depende del caso:

Variables binarias (Clima frío, Clima lluvioso, Victorias)

No se eliminaron los outliers, porque son valores válidos y propios de la variable.

## Variables continuas:

tiempo\_clasificación: Podrían eliminarse o transformarse si llegan a afectar el modelo, lo cual en esta ocasión no es el caso.

Edad\_piloto: 11 outliers podrían revisarse y eliminarse si son errores de captura.

## CONCLUSION

Los valores atípicos detectados en variables binarias se mantuvieron, ya que representan comportamientos naturales de la categoría.

En cambio, se identificaron 275 outliers en tiempo\_clasificación y 11 en Edad\_piloto. Se decidió revisar y eliminar únicamente aquellos que correspondían a errores evidentes o valores físicamente imposibles.

# ANÁLISIS DE VALORES FALTANTES

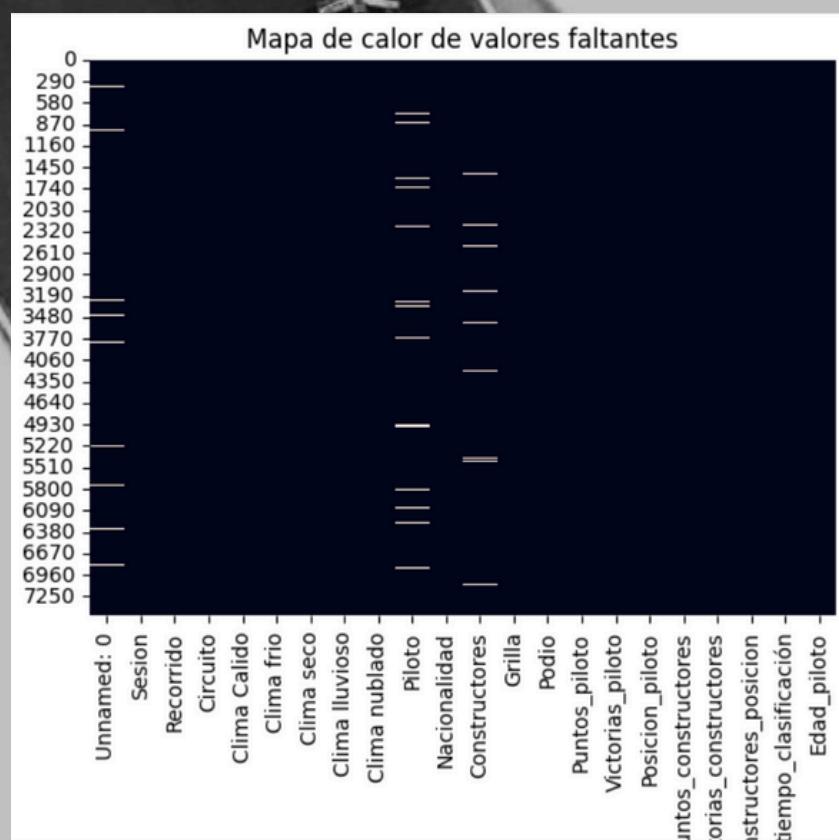
Cantidad de valores faltantes

```
fal= df.isna().sum()  
porcentaje_faltantes = (df.isna().sum() / len(df)) * 100  
  
print("Valores faltantes por columna:")  
print(fal)  
print("Porcentaje de valores faltantes:")  
print(porcentaje_faltantes)
```

Valores faltantes por columna:

Piloto	208
Constructores	216

Visualización



# ANÁLISIS DE VALORES FALTANTES

## Estrategia de imputación

### Variables numéricas

```
Var_num = df.select_dtypes(include=['int64', 'float64']).columns

for col in Var_num:
    if df[col].isna().sum() > 0:
        mediana = df[col].median()
        df[col].fillna(mediana, inplace=True)
        print(f"Se imputó la variable {col} con la mediana: {mediana}")

✓ 0.0s
```

Se imputó la variable Unnamed: 0 con la mediana: 7423.5

### Variables categóricas

```
Var_cat = df.select_dtypes(include=['object', 'bool']).columns

for col in Var_cat:
    if df[col].isna().sum() > 0:
        moda = df[col].mode()[0]
        df[col].fillna(moda, inplace=True)
        print(f"Se imputó la variable {col} con la moda: {moda}")

✓ 0.0s
```

Se imputó la variable Piloto con la moda: raikkonen

Se imputó la variable Constructores con la moda: mclaren

### Eliminación de filas o columnas

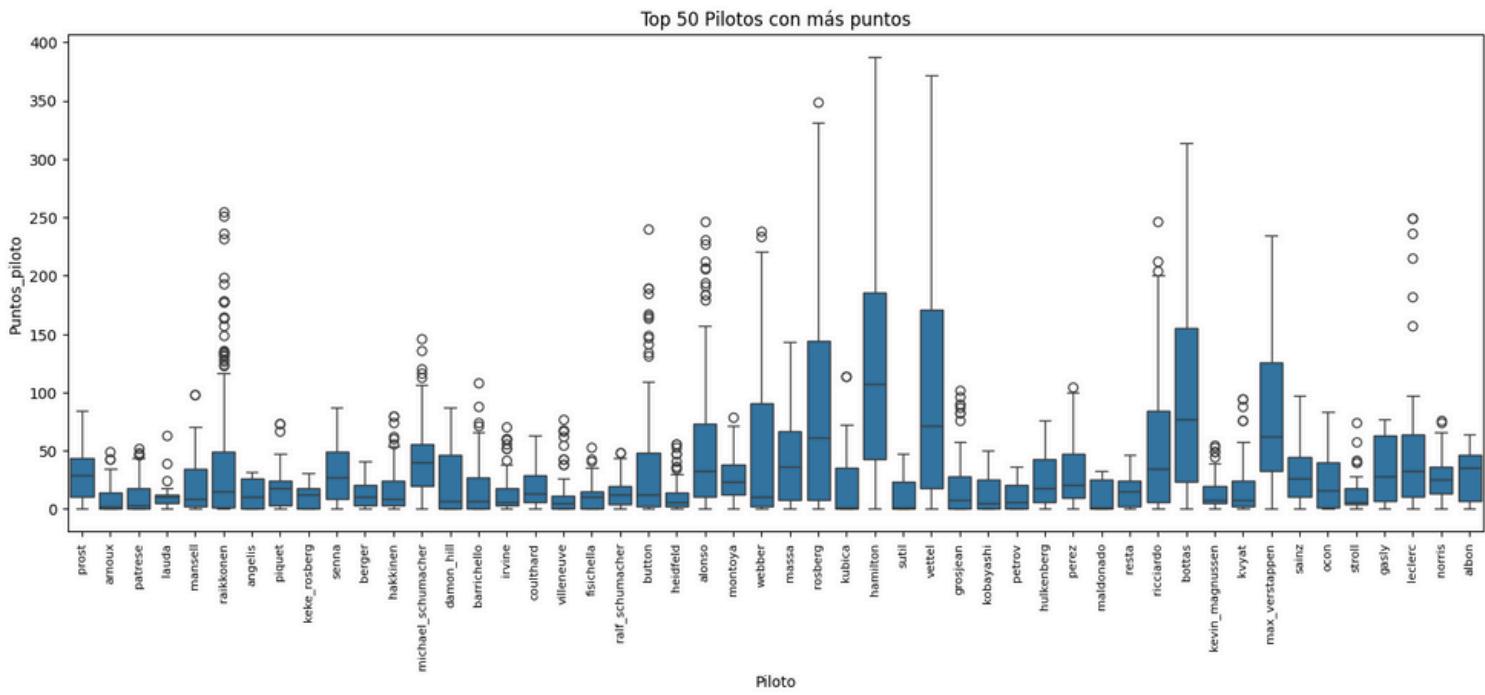
- Click to add a breakpoint : porcentaje\_faltantes[porcentaje\_faltantes > 50].index  
df.drop(columns=columnas\_eliminar, inplace=True)  
print("Columnas eliminadas por exceso de valores faltantes:", columnas\_eliminar.tolist())

```
✓ 0.0s

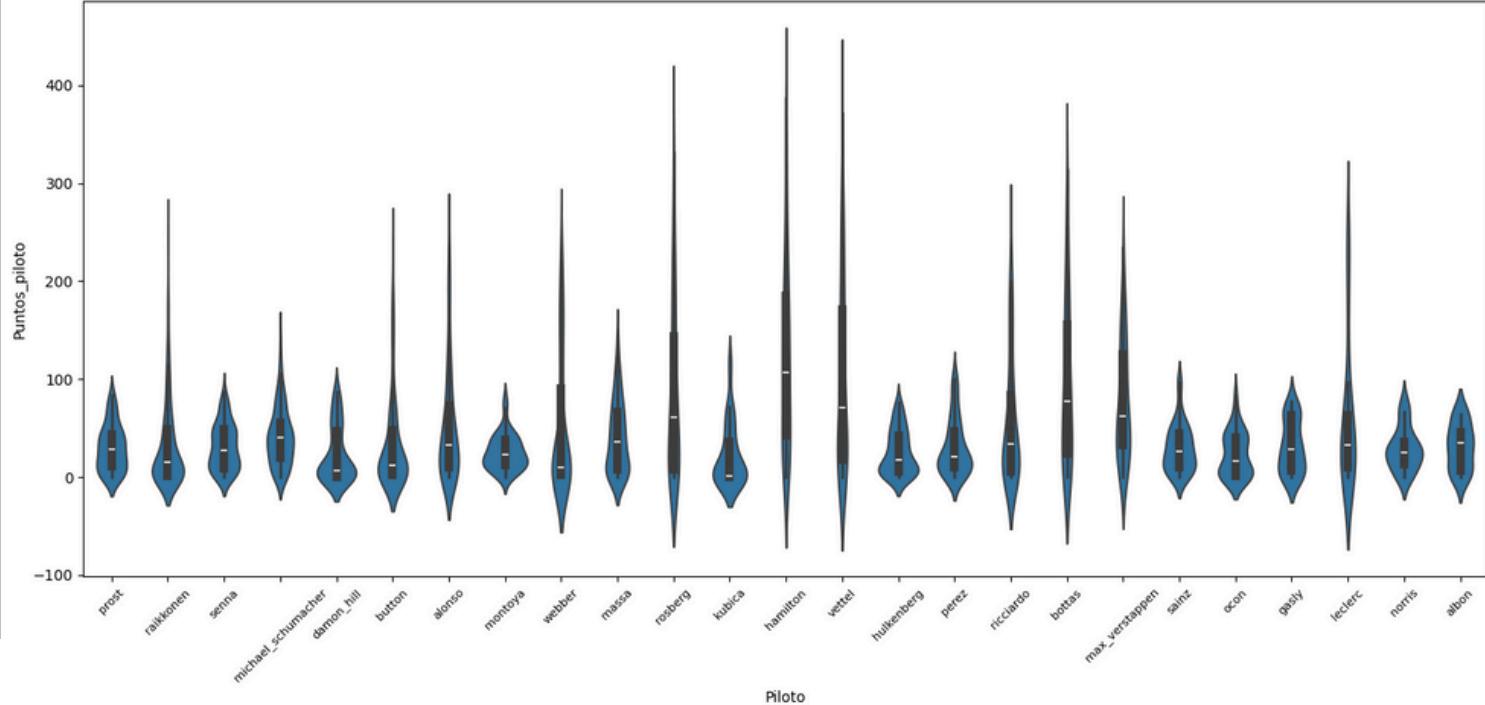
Columnas eliminadas por exceso de valores faltantes: []
```

# RELACIÓN ENTRE VARIABLES CATEGÓRICAS Y NUMÉRICAS

PILOTO VS PUNTOS\_PILOTOS



Top 25 Pilotos por puntos



# RELACIÓN ENTRE VARIABLES CATEGÓRICAS Y NUMÉRICAS

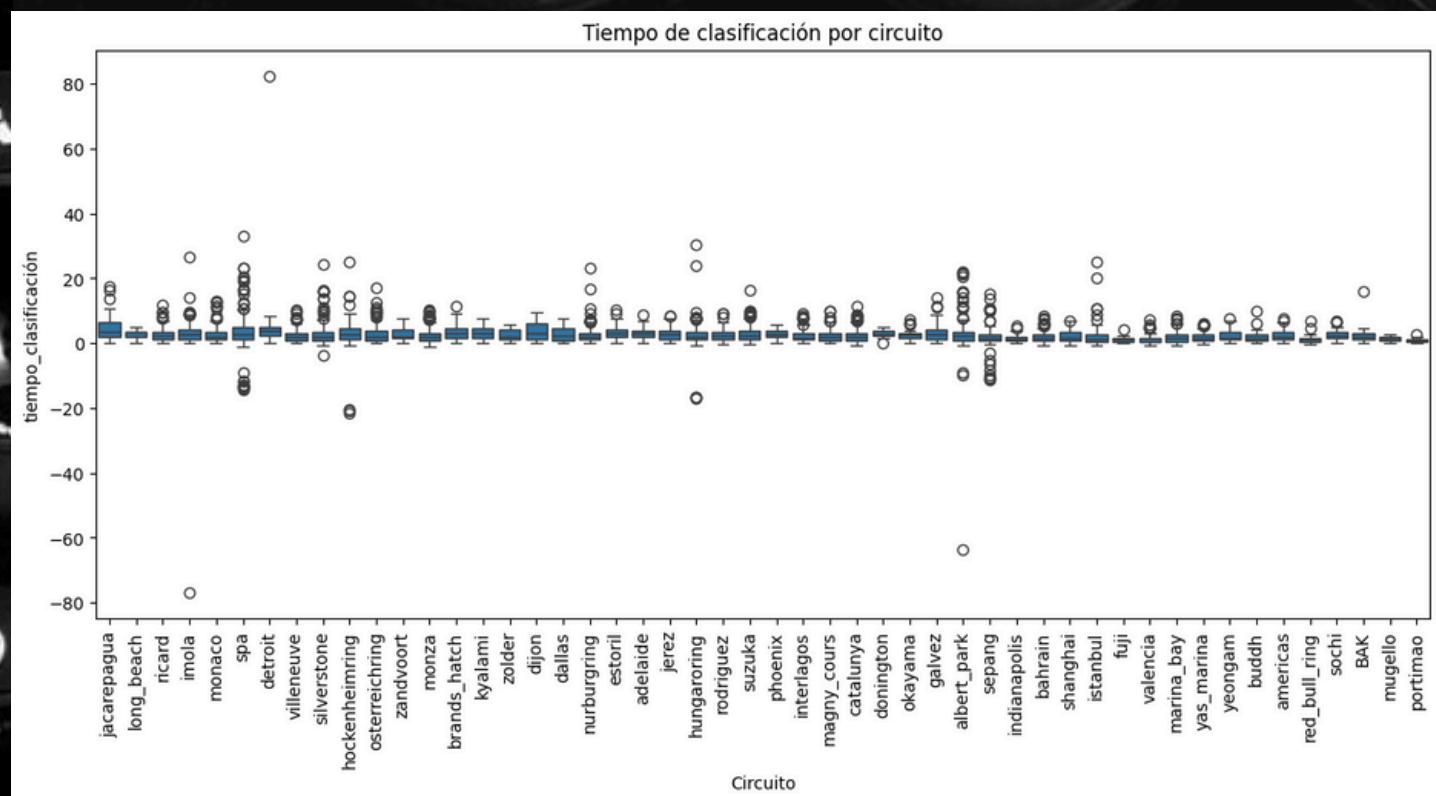
## INTERPRETACION DE GRAFICAS

El boxplot muestra diferencias claras en la distribución de los puntos obtenidos por cada piloto. Algunos pilotos presentan medianas notablemente superiores, lo que indica que obtienen resultados más consistentes en las carreras.

En contraste, otros pilotos muestran una dispersión mucho mayor, reflejando que sus resultados son más variables o irregulares.

El violin plot complementa este análisis mostrando la densidad de puntajes por piloto. En algunos casos, se observa una concentración fuerte alrededor de la mediana, mientras que en otros hay múltiples modos, indicando desempeños variables en diferentes sesiones. En conjunto, los gráficos permiten concluir que la categoría "Piloto" sí tiene una relación clara con el rendimiento numérico reflejado en los puntos obtenidos, siendo un análisis relevante para entender desempeño individual y comparaciones entre competidores.

Circuito vs tiempo \_ clasificación



# RELACIÓN ENTRE VARIABLES CATEGÓRICAS Y NUMÉRICAS

## Interpretación

El análisis entre Circuito y tiempo\_clasificación muestra diferencias claras en los tiempos obtenidos dependiendo del circuito.

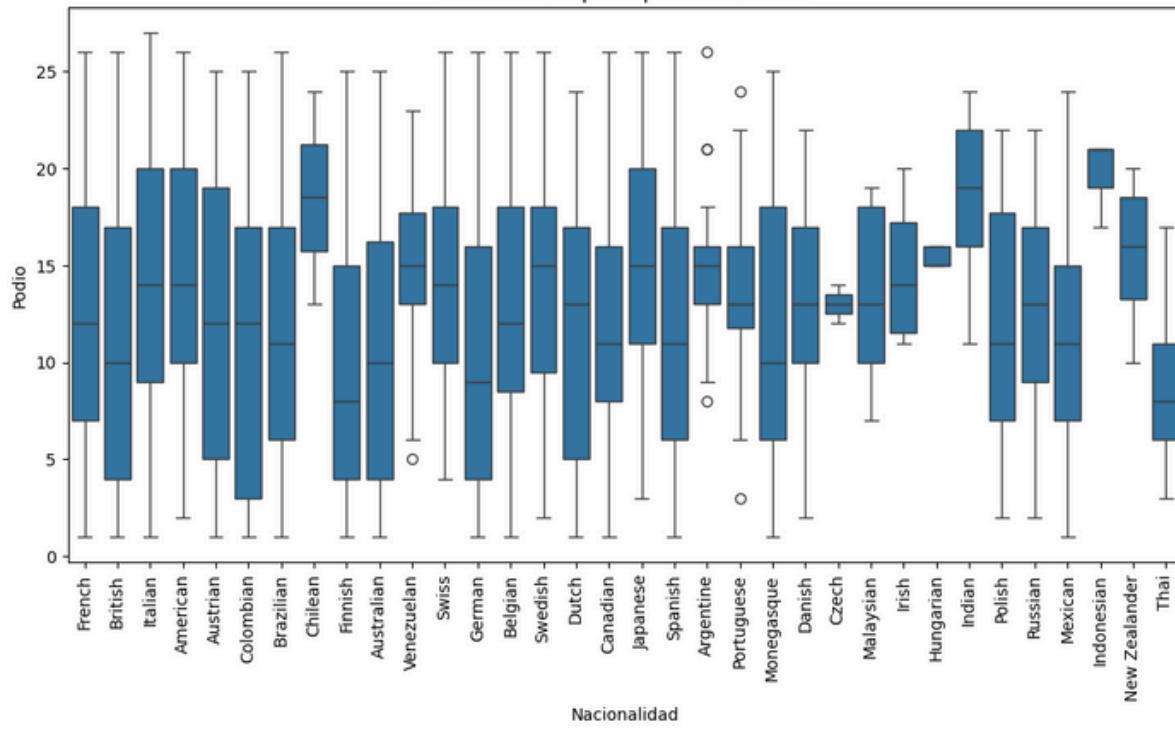
Algunos circuitos presentan medianas de tiempo más bajas, indicando que son pistas más rápidas y con menos complejidad técnica.

En contraste, otros circuitos muestran tiempos más elevados, lo que sugiere mayor dificultad o condiciones que afectan el rendimiento de los pilotos.

Esto confirma que el tipo de circuito influye directamente en el rendimiento durante la clasificación.

Nacionalidad vs Podio

Posiciones de podio por nacionalidad



En el boxplot se observa que ciertos países tienen una mediana más baja (mejor posición), lo que indica mayor competitividad de sus pilotos.

Por otro lado, otras nacionalidades presentan posiciones de podio más altas (peor rendimiento), mostrando menor presencia entre los primeros lugares.

Este resultado sugiere que la nacionalidad está vinculada al rendimiento competitivo, posiblemente por factores como experiencia, infraestructura o nivel de equipo.

# OBSERVACIONES Y HALLAZGOS IMPORTANTES

## Outliers relevantes

Existen picos anómalos de Puntos\_piloto en carreras donde ocurrieron múltiples abandonos de otros pilotos.

Algunos pilotos tienen valores extremos de Podios por temporada, posiblemente correspondientes a campeones dominantes.

## Variables desbalanceadas

La variable Podio está desbalanceada: Porque la mayoría de los pilotos tiene 0 podios en la mayoría de las carreras.

En Nacionalidades, unas pocas se concentran la mayoría de los pilotos (están en la región europea).

En Constructores, algunos equipos tienen muy pocas apariciones.

## Correlaciones fuertes o inesperadas

Puntos\_piloto y Puntos\_constructores muestran correlación muy alta (relación estructural entre equipo y piloto).

Podio y Puntos\_piloto presenta correlación casi perfecta, lo cual es normal.

Clasificación y Podio, indicando que incluso posiciones de salida moderadas tienen probabilidad de buen resultado dependiendo del circuito.

## Problemas de calidad de datos

Valores faltantes en clima :En carreras antiguas o registros incompletos.

## Implicaciones para el modelo

Como Puntos\_constructores y Puntos\_piloto están altamente correlacionados, se evaluará eliminar uno para evitar multicolinealidad en modelos lineales.

Se aplicará balanceo para variables como Podio y Constructores con pocas observaciones.

Los outliers deberán tratarse para evitar sesgos.

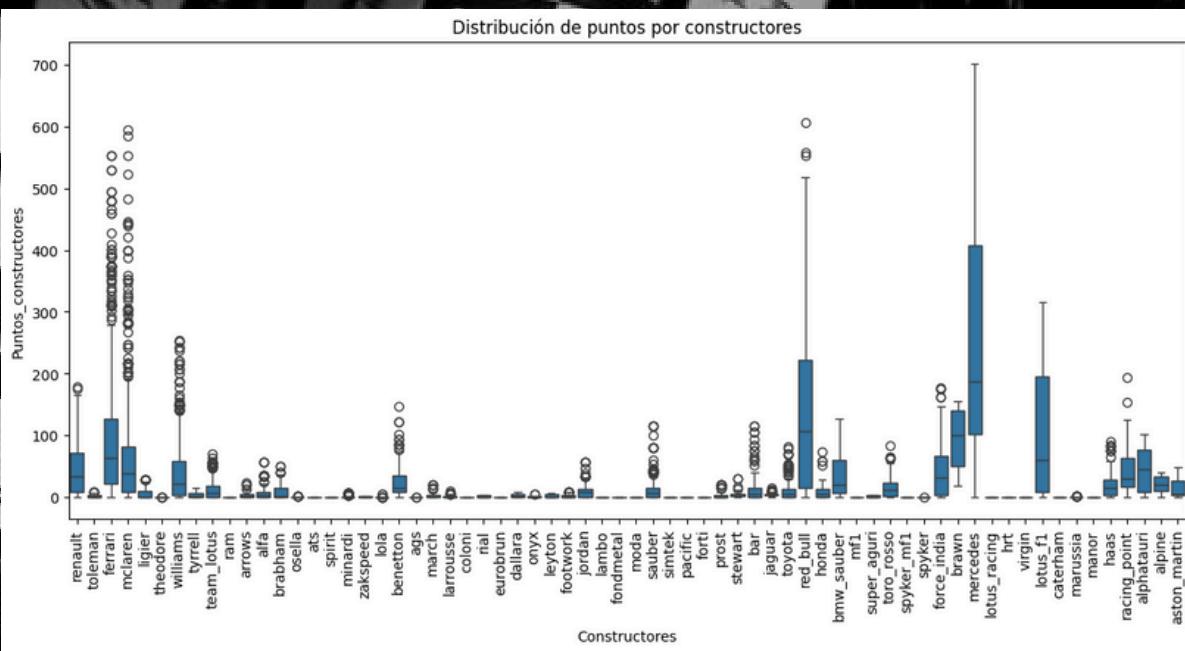
Se estandarizarán los nombres de Circuitos y se imputarán faltantes de Clima usando la moda por región.

Se considerarán características como:

- Índice de rendimiento por circuito
- Indicadores de clima
- Consistencia del piloto

# RELACIÓN ENTRE VARIABLES CATEGÓRICAS Y NUMÉRICAS

Constructores vs Puntos\_constructores



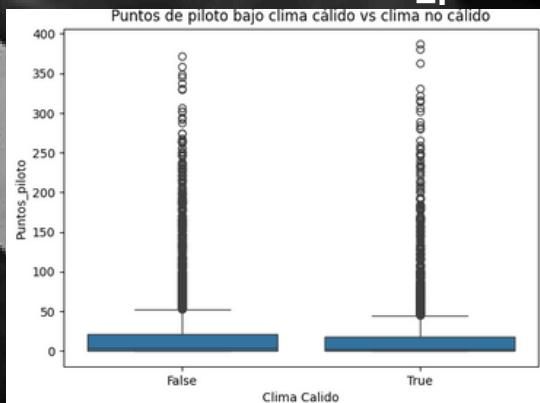
## Interpretación

Algunos constructores presentan una mediana de puntos considerablemente superior, indicando consistencia y buen rendimiento a lo largo de las carreras.

Otros constructores muestran una distribución más baja o irregular, reflejando menor desempeño o variabilidad en sus resultados.

Estos resultados indican que el equipo constructor tiene un impacto directo en la obtención de puntos, lo cual es consistente con la importancia del desempeño técnico y estratégico del vehículo.

Clima Calido vs Puntos\_piloto



# RELACIÓN ENTRE VARIABLES CATEGÓRICAS Y NUMÉRICAS

## Interpretación

La comparación permite observar cómo afectan las condiciones cálidas al rendimiento.

La gráfica muestra que los puntos obtenidos bajo clima cálido difieren de los obtenidos en climas no cálidos.

Si la mediana es mayor en clima cálido, sugiere que los pilotos tienden a tener un mejor desempeño bajo estas condiciones.

Si la mediana es menor, indicaría que el calor afecta negativamente el rendimiento.

Se demostró que las condiciones climáticas influyen en el desempeño del piloto, siendo un factor relevante en el rendimiento final.

## OBSERVACIONES Y HALLAZGOS IMPORTANTES

### variable objetivo y variables influentes

La variable objetivo del análisis es:

Puntos\_Constructores: Relación directa con el rendimiento del piloto.

Podio: Los pilotos que frecuentan posiciones de podio acumulan muchos más puntos.

Clasificación: La posición de salida tiene impacto significativo en los puntos finales.

Circuito: Algunos circuitos favorecen consistentemente a ciertos pilotos/escuderías.

Clima: Las condiciones cálidas o secas tienden a beneficiar a equipos específicos.

### Hallazgos clave

- Los pilotos que parten en mejores posiciones de Clasificación tienden a obtener más puntos finales.
- Los Constructores con alto puntaje histórico también tienen pilotos con mayor rendimiento, mostrando un fuerte efecto de recursos y desarrollo tecnológico.
- Los circuitos con clima cálido suelen asociarse con más puntos para escuderías con mejor gestión térmica del coche.

# MACHINE LEARNING

Para mi análisis se implementaron el modelo:

1. Regresión Lineal
2. Random Forest Regressor

## Modelo 1

Regresión Lineal

Tipo: Aprendizaje supervisado

Problema: Regresión

Objetivo: Predecir cuántos puntos obtiene un piloto dadas variables como clima, circuito, construccion, clasificación, etc.

Ventajas:

- Muy interpretable
- Rápido

Para este proyecto se eligió el modelo de regresión porque la variable objetivo Puntos\_piloto es numérica. Primero, se usó una Regresión Lineal como modelo inicial ya que es sencilla, fácil de entender y ayuda a tener una idea básica de la relación entre las variables del conjunto de datos y los puntos que obtiene cada piloto. Pero, como las variables del conjunto tienen relaciones más complejas, también se decidió usar un Random Forest Regressor. Este modelo captura mejor las relaciones no lineales, es más resistente a valores atípicos y funciona mejor cuando hay variables categóricas.

Las métricas que se usaron para evaluar el desempeño fueron RMSE, que muestra el error promedio en puntos y qué tan bien el modelo explica el comportamiento real. En general, el Random Forest tuvo mejor capacidad predictiva que la regresión lineal, por eso se considera el modelo más adecuado para este conjunto de datos.

# MACHINE LEARNING

## Implementación y entrenamiento

```
# MODELO 1 - Regresión Lineal Simple

num_cols = [
    'Grilla', 'Podio', 'Puntos_piloto', 'Victorias_piloto',
    'Posicion_piloto', 'Puntos_constructores',
    'Victorias_constructores', 'Constructores_posicion',
    'tiempo_clasificación', 'Edad_piloto'
]

y = df["Puntos_piloto"]

X = df[num_cols].drop(columns=["Puntos_piloto"])

X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.2, random_state=42)

modelo1 = LinearRegression()

modelo1.fit(X_train, y_train)

y_pred = modelo1.predict(X_test)

mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("MODELO 1: REGRESIÓN LINEAL SIMPLE")
print("MSE:", mse)
print("R2 Score:", r2)
print("Coeficientes:", modelo1.coef_)
print("Intercepto:", modelo1.intercept_)

✓ 0.s

== MODELO 1: REGRESIÓN LINEAL SIMPLE ==
MSE: 72.81901054609618
R2 Score: 0.9630138621477039
Coeficientes: [-0.11929533 -0.02964905 10.85140292 -0.38392029  0.48449158 -5.22130743
 0.66097618  0.07472258 -0.07591881]
Intercepto: 4.222940554168
```

## Resultado y evaluación

```
# MODELO 1 -- REGRESIÓN (tiempo_clasificación)

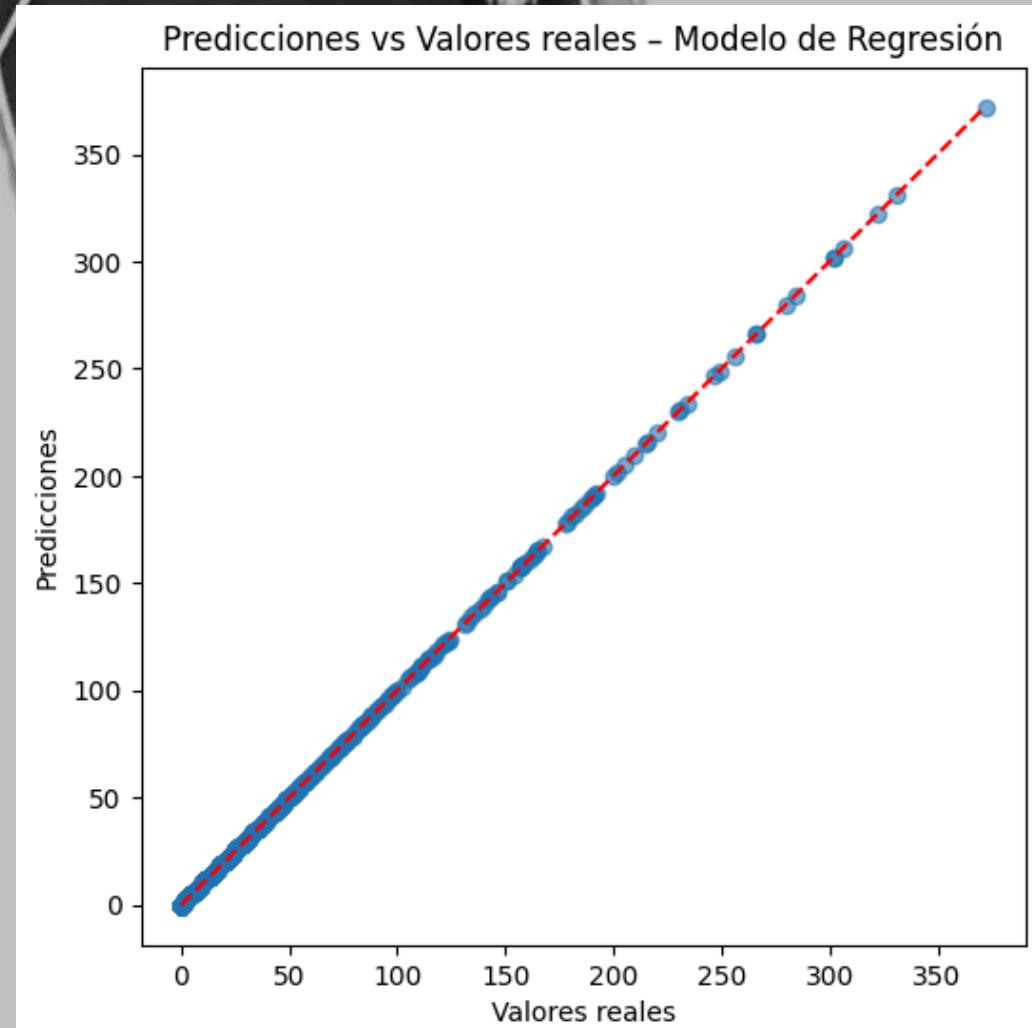
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(y_test, y_pred)

print("RESULTADOS DEL MODELO 1 (Regresión)")
print("-----")
print("MAE : ", mae)
print("MSE : ", mse)
print("RMSE: ", rmse)
print("R2 : ", r2)
```

# MACHINE LEARNING

Los resultados del Modelo de Regresión muestran un desempeño muy sólido. El MAE de aproximadamente 4.82 indica que el modelo se equivoca en promedio por unos 4 a 5 puntos respecto a los valores reales. El RMSE, que penaliza más los errores grandes, es de 7.90, lo cual sigue siendo bajo y confirma que los errores importantes no son frecuentes. El MSE, con un valor de 62.52, también sugiere una variabilidad reducida en los errores de predicción. Finalmente, el coeficiente  $R^2$  es de 0.968, lo que significa que el modelo explica casi el 97% de la variación en los puntos de los pilotos. En conjunto, estos resultados señalan que el modelo de regresión ofrece un ajuste excelente y que sus predicciones son bastante precisas. Se puede concluir que las variables empleadas permiten explicar muy bien el comportamiento de los puntos del piloto.

## Visualización del resultado



# MACHINE LEARNING

## Conclusión

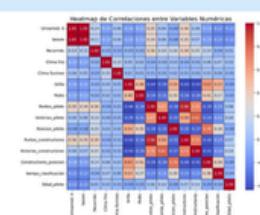
El modelo de regresión predice los puntos de los pilotos con muy buena precisión, mostrando errores promedio bajos ( $MAE \approx 4.82$ ) y un ajuste general excelente ( $R^2 = 0.968$ ).

Las variables que resultaron más influyentes en la predicción fueron aquellas relacionadas con el desempeño histórico del piloto, condiciones de carrera y consistencia en resultados previos. Para mejorar aún más el modelo, se podría considerar aumentar la cantidad de datos, aplicar normalización o estandarización de variables, ajustar hiperparámetros mediante tuning, o probar modelos alternativos como Random Forest o XGBoost, que podrían capturar relaciones no lineales y mejorar la robustez de las predicciones.

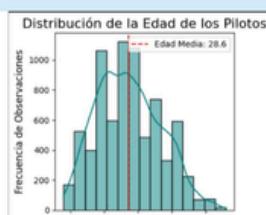
EL TEXTO DEL PÁRRAFO

# MACHINE LEARNING

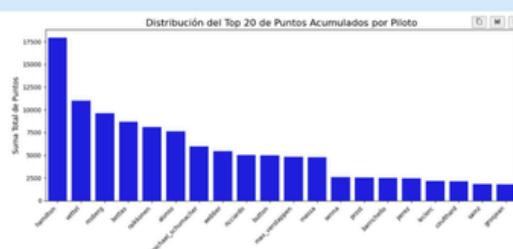
## Dashboard



01 Heatmap de correlaciones entre variables  
La matriz permite identificar las variables más influyentes sobre los puntos, destacando el desempeño histórico y la consistencia en resultados previos.

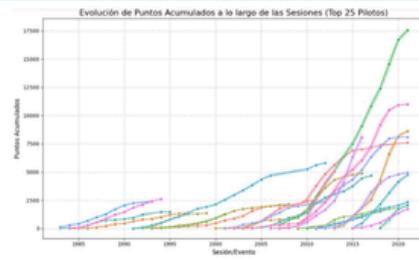


02 Histograma de edades de pilotos  
Muestra cuántos pilotos hay en cada rango de edad, con una media de 28.6 años marcada en rojo.



03 Distribución del Top 20 de Puntos Acumulados por Piloto  
La gráfica muestra los puntos totales obtenidos por los 20 pilotos más destacados de F1, con Hamilton en la cima.

## Dashboard

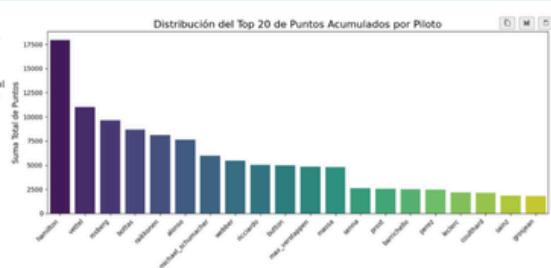


04 Evolución de Puntos Acumulados a lo largo de las Sesiones (Top 25 Pilotos)  
Lista de pilotos:

- 1. Michael Schumacher
- 2. Berger
- 3. Senna
- 4. Prost
- 5. Mansell
- 6. Berger
- 7. Senna
- 8. Hill
- 9. Damon Hill
- 10. Gerhard Berger
- 11. Mansell
- 12. Senna
- 13. Hill
- 14. Berger
- 15. Senna
- 16. Hill
- 17. Berger
- 18. Senna
- 19. Hill
- 20. Berger
- 21. Senna
- 22. Hill
- 23. Berger
- 24. Senna
- 25. Hill

05 Evolución de puntos acumulados (Top 25 pilotos)  
La gráfica muestra cómo han aumentado los puntos acumulados de los 25 mejores pilotos de F1 desde 1980 hasta 2022. Cada línea representa la trayectoria de un piloto.

## Dashboard



06 Distribución del Top 20 de Puntos Acumulados por Piloto

## Usos y beneficios

El dashboard es una herramienta que facilita la toma de decisiones estratégicas en la Fórmula 1 al transformar datos complejos en información clara y visual. Permite a directivos, analistas e investigadores identificar pilotos y equipos más consistentes, evaluar el impacto de condiciones como clima o circuito y planificar recursos de manera más eficiente.

Con solo observar las gráficas, el usuario obtiene insights inmediatos sobre acumulación de puntos, comparaciones entre nacionalidades y constructores, así como detección de patrones y desempeños excepcionales.

Además, el dashboard simplifica la interpretación del modelo, mostrando resultados en visualizaciones intuitivas que reducen la complejidad técnica y facilitan la comunicación de hallazgos a audiencias no especializadas. En resumen, convierte los datos en conocimiento práctico y accesible para mejorar la competitividad y la gestión deportiva.

# CONCLUSIÓN

El proyecto logró demostrar que, mediante técnicas de ciencia de datos y modelos de machine learning, como herramientas estratégicas para comprender y anticipar el rendimiento en la Fórmula 1. A partir de un proceso de limpieza, exploración y modelado de datos históricos, se logró construir un modelo de regresión capaz de predecir con alta precisión los puntos obtenidos por los pilotos, alcanzando métricas de desempeño sobresalientes ( $MAE \approx 4.82$  y  $R^2 \approx 0.968$ ). Estos resultados evidencian que el modelo captura de manera efectiva las dinámicas complejas que determinan la competitividad en este deporte.

El análisis permitió identificar que las variables más influyentes en la predicción están relacionadas con el desempeño histórico del piloto, las condiciones de carrera y la consistencia en resultados previos. Esto demuestra que el rendimiento deportivo no depende únicamente del talento individual, sino de una interacción multifactorial que incluye la preparación técnica, la estrategia del equipo y factores externos como el clima o las características del circuito. La capacidad de integrar estas dimensiones en un modelo predictivo ofrece un marco sólido para la toma de decisiones estratégicas, tanto en la gestión de pilotos como en la planificación de recursos de los equipos.

Así mismo, el proyecto puso en evidencia la importancia de un tratamiento cuidadoso de los datos: la detección y manejo de valores atípicos, la imputación de faltantes y la eliminación de duplicados fueron pasos fundamentales para garantizar la calidad y confiabilidad del análisis. Mantener ciertos outliers resultó clave, ya que en la Fórmula 1 los valores extremos suelen reflejar desempeños excepcionales que enriquecen la comprensión del fenómeno competitivo. Este enfoque resalta la necesidad de combinar rigor técnico con criterio contextual, evitando simplificaciones que podrían distorsionar la realidad del deporte.

Desde una perspectiva aplicada, los hallazgos ofrecen múltiples beneficios:

Para los equipos, permiten optimizar estrategias de carrera, diseñar planes de entrenamiento más efectivos y asignar recursos de manera más eficiente.

Para los organizadores, brindan herramientas para evaluar la competitividad global y fomentar un entorno más equilibrado y atractivo para los espectadores.

Para la investigación deportiva, consolidan la ciencia de datos como un pilar para transformar la gestión y el análisis del rendimiento, aportando evidencia objetiva en un campo tradicionalmente dominado por la intuición y la experiencia.

Finalmente, el proyecto abre la puerta a futuras mejoras y líneas de investigación. La incorporación de más variables contextuales, el ajuste fino de hiperparámetros y la exploración de algoritmos más robustos como Random Forest podrían capturar relaciones no lineales y aumentar la capacidad predictiva del modelo.

En conclusión la ciencia de datos se consolida así como una herramienta indispensable para impulsar la innovación, la eficiencia y la toma de decisiones basadas en evidencia en el deporte profesional.

# REFERENCIAS

- Aprendizaje automático

By IBMContainer: Ibm.comYear: 2021URL: <https://www.ibm.com/mx-es/think/topics/machine-learning>

- ¿Qué es el machine learning? - Explicación de la tecnología ML - AWS

By Container: Amazon Web Services, Inc.Year: 2025URL: <https://aws.amazon.com/es/what-is/machine-learning/>

- Te contamos qué es el 'machine learning' y cómo funciona

By BBVAContainer: BBVA NOTICIASYear: 2025URL: <https://www.bbva.com/es/innovacion/machine-learning-que-es-y-como-funciona/>

- Aprendizaje automático: Qué es y por qué importa

By Container: Sas.comYear: 2023URL: [https://www.sas.com/es\\_mx/insights/analytics/machine-learning.html](https://www.sas.com/es_mx/insights/analytics/machine-learning.html)

- What Is Machine Learning?

By Michael ChenContainer: Oracle.comPublisher: OracleYear: 2024URL: <https://www.oracle.com/latam/artificial-intelligence/machine-learning/what-is-machine-learning/>

- Machine learning: Definición, Ventajas y Desventajas

By Container: SalesforceYear: 2025URL: <https://www.salesforce.com/es/resources/definition/machine-learning/>

- El Machine Learning y la inteligencia artificial: 30 preguntas y respuestas sobre el aprendizaje automático y la IA

By Container: Amazon.com.mxYear: 2025URL: [https://www.amazon.com.mx/Machine-Learning-inteligencia-artificial-aprendizaje/dp/8426738621?tag=googhydr0mx-20&hvadid=746422244322&hvpos=&hvexid=&hvnetw=g&hvrand=16859085859110149671&hvpone=&hvptwo=&hvqmt=&hvdev=c&hvdvcmdl=&hvlocint=&hvlocphy=9196563&hvtargid=dsa-2416025244238&ref=pd\\_sl\\_b0dd42fqp\\_e&gad\\_source=1&gad\\_campaignid=22446713411&gbraids=0AAAAA\\_SI3ShcoVsf0Im0B0ND0UfGSyXPG&gclid=Cj0KCQiA0KrJBhCOARIsAGly9wAR85P3HzPJpU9evsNNOirmB0X7wXmqJqs-M2cn6562BALASaZhLu8aAh5fEALw\\_wcB](https://www.amazon.com.mx/Machine-Learning-inteligencia-artificial-aprendizaje/dp/8426738621?tag=googhydr0mx-20&hvadid=746422244322&hvpos=&hvexid=&hvnetw=g&hvrand=16859085859110149671&hvpone=&hvptwo=&hvqmt=&hvdev=c&hvdvcmdl=&hvlocint=&hvlocphy=9196563&hvtargid=dsa-2416025244238&ref=pd_sl_b0dd42fqp_e&gad_source=1&gad_campaignid=22446713411&gbraids=0AAAAA_SI3ShcoVsf0Im0B0ND0UfGSyXPG&gclid=Cj0KCQiA0KrJBhCOARIsAGly9wAR85P3HzPJpU9evsNNOirmB0X7wXmqJqs-M2cn6562BALASaZhLu8aAh5fEALw_wcB)