Mejora de rendimiento: Análisis para transformar el juego

PROYECTO FINAL – DATA ANALYTICS 2.0

Sofía Echeverría - María Paz Camino – Leonel Fuhrmann
Agustín Brandt

15-08-2025

1. Introducción

El presente informe tiene como objetivo presentar el desarrollo y los principales hallazgos de un proyecto de análisis de datos aplicado al ámbito deportivo profesional, centrado en la liga NBA.

El análisis se desarrolló en múltiples etapas, que incluyeron la limpieza y transformación de grandes volúmenes de datos, la construcción de un modelo analítico, el diseño de visualizaciones interactivas y la generación de métricas clave orientadas a la toma de decisiones estratégicas.

Finalmente, el informe concluye con un conjunto de conclusiones, reflexiones metodológicas y recomendaciones basadas en evidencia, que permiten evaluar el potencial de esta propuesta para mejorar el rendimiento competitivo del equipo involucrado.

2. Propuesta de Proyecto

El presente proyecto se enmarca en el análisis de datos del deporte profesional, específicamente de la NBA, con el objetivo de brindar recomendaciones estratégicas para posibles fichajes, en función del desempeño de los jugadores y las capacidades presupuestarias del equipo. Para ello, se analizarán variables como edad, peso, características físicas, rendimiento deportivo, historial de victorias y derrotas, y salarios. Si bien inicialmente se contempló una comparación entre el equipo con más derrotas y el más exitoso, la falta de información cualitativa sobre estos últimos llevó a reformular el enfoque. En su lugar, se trabajará sobre la identificación de perfiles de jugadores con buen desempeño que resulten viables para el equipo que solicita el análisis, integrando criterios deportivos y financieros. Este trabajo cobra especial relevancia ante la necesidad de preservar el apoyo de los patrocinadores, quienes han manifestado preocupación por el rendimiento reciente del equipo y podrían reconsiderar su continuidad si no se implementan estrategias de mejora basadas en evidencia.

El desarrollo del proyecto se llevará a cabo aplicando la metodología ágil Scrum, lo cual permite organizar el trabajo en iteraciones breves y enfocadas. El proceso constará de dos Sprints, donde se desarrollarán las principales etapas técnicas del proyecto: limpieza y estructuración de datos, análisis exploratorio, definición de métricas clave, visualización de resultados y automatización del flujo de trabajo. El proyecto culminará con una demo final, donde se presentarán los entregables alcanzados.

Durante toda la ejecución, se realizarán Daily Meetings para el seguimiento de avances, la coordinación de tareas y la resolución de bloqueos. El Scrum Master será un miembro del equipo docente de Henry, quien facilitará el proceso y asegurará el cumplimiento de la metodología.

Los estudiantes asumirán el rol de Developers, siendo responsables de la implementación técnica del proyecto, mientras que el rol de Product Owners será ocupado por los evaluadores, quienes revisarán el cumplimiento de los objetivos en cada entrega y brindarán feedback sobre la evolución del trabajo.

Las tareas técnicas incluyen la creación de la base de datos en MySQL, procesamiento y transformación del dataset con Python (usando bibliotecas como Pandas y Numpy), desarrollo de visualizaciones exploratorias y reportes ejecutivos en Power BI, así como la implementación de medidas DAX y automatización del flujo de datos.

También se diseñará un modelo entidad-relación para estructurar correctamente la información y se aplicarán buenas prácticas de limpieza, análisis y visualización para garantizar la calidad de los resultados.

Este enfoque ágil busca fomentar la colaboración, la mejora continua y la entrega incremental de valor, alineando las mejores prácticas del desarrollo profesional con los desafíos del análisis de datos en contextos reales.

3. Organización y Documentación Inicial del Proyecto

Para dar inicio al desarrollo del proyecto, se llevaron a cabo una serie de tareas fundamentales orientadas a la organización del equipo, la gestión del entorno colaborativo y la documentación técnica preliminar. Estas acciones permitieron establecer una base clara y ordenada para avanzar de forma eficiente en las etapas analíticas posteriores.

Las actividades desarrolladas en esta etapa fueron:

• Creación del canal de trabajo en **Discord**, utilizado como espacio central de comunicación y coordinación entre los integrantes del equipo. Allí se comparten documentos, avances, materiales de consulta y se realizan reuniones virtuales periódicas.

• Descarga y revisión del dataset principal desde **Kaggle,** compuesto por 16 archivos vinculados a estadísticas de partidos, jugadores, equipos y otros aspectos relevantes del universo NBA. Adicionalmente, se descargó un dataset complementario denominado Salaries, con información sobre los sueldos de los jugadores, que permitirá enriquecer el análisis incorporando variables económicas al estudio de rendimiento y fichajes. Posteriormente, se almacenaron los archivos CSV crudos en una carpeta compartida de Google Drive, que centraliza el acceso a los datos originales y garantiza su disponibilidad para todos los miembros del equipo.

• Elaboración de un **Diccionario de Datos**, en el que se detalla la estructura de cada una de las tablas del dataset. Incluye los nombres de las columnas, el tipo de dato correspondiente, ejemplos concretos de los valores encontrados, traducción al español de los nombres de cada campo y porcentaje de valores nulos por columna (Anexo 1).

• Creación de un repositorio en **GitHub** para alojar el proyecto de forma colaborativa. El repositorio PF_NBA_EQUIPO1 permite el seguimiento de versiones, la carga de documentación y el control de cambios por parte de los cuatro integrantes del equipo.

• Apertura de una cuenta en **Google Cloud Platform** con el objetivo de trabajar en la nube de manera indistinta y acceder a un entorno de procesamiento centralizado. Previamente se realizó el análisis exploratorio y el proceso de ETL utilizando Python. Luego, los archivos CSV procesados fueron cargados directamente en **BigQuery,** donde se estructuraron como base de datos para facilitar su consulta y análisis posterior.

• Inicio del archivo de documentación del proyecto, donde se registran cronológicamente las actividades realizadas, decisiones tomadas, herramientas utilizadas y aprendizajes del proceso. Este archivo se actualizará de forma continua a lo largo del desarrollo del trabajo.

• Creación de un tablero en **Trello,** destinado a la planificación y seguimiento de tareas del proyecto. Allí se definieron columnas por etapa de trabajo, se asignaron responsables y se registraron avances, lo que permite una gestión ágil y visual del proceso.

Estas tareas nos permitieron no solo ordenar el trabajo en equipo y facilitar la colaboración, sino también establecer una base común de comprensión sobre el dataset.

4. Definición del universo de análisis: selección de equipos

Con el objetivo de delimitar el universo de análisis y enfocar la propuesta en un caso concreto, se realizó un análisis exploratorio preliminar utilizando la tabla Games, que contiene información detallada de cada partido, incluyendo fecha, nombre del equipo, sigla abreviada y resultado del encuentro, tanto en condición de local como de visitante. Las columnas que indican el resultado de cada equipo están representadas por las letras W (win) y L (loss), es decir, victoria o derrota. El procedimiento consistió en:

• Filtrar los datos correspondientes a los últimos cinco años, entre 2018 y 2023, para enfocar el análisis en la performance reciente de los equipos.

• Agrupar los datos por equipo y contabilizar la cantidad total de partidos ganados (W) y perdidos (L), diferenciando entre partidos jugados como local y visitante.

• Calcular, para cada equipo, la diferencia entre partidos ganados y perdidos en el período analizado. Ordenar los resultados de mayor a menor según dicha diferencia, con el fin de identificar a los equipos con mejor performance global.

• De este análisis surgieron los 10 equipos con mejor diferencial de victorias, entendidos como aquellos con un rendimiento destacado en las últimas temporadas. El equipo que ocupó el décimo lugar en el ranking fue *Los Angeles Lakers.*

•

En función de estos resultados, se definió continuar el análisis centrado en dicho equipo, asumiendo que son ellos quienes han solicitado este trabajo.

A continuación, se presenta el desarrollo técnico de este análisis, disponible en el notebook denominado *Game_CSV_filtrada.ipynb*

```
import pandas as pd

## se calcularon la cantidad de partidos ganados y periddos de local y de visitante.
df = pd.read_csv(r"C:\Users\PATRICIA\Desktop\proyecto_final\Nueva carpeta\game.csv",
sep=';')
print(df.columns)
df.columns = df.columns.str.strip()

home = df[['team_name_home','wl_home']].copy()
home['win'] = (home['wl_home'] == 'W').astype(int)
home['loss'] = (home['wl_home'] == 'L').astype(int)
home.rename(columns={'team_name_home': 'team'}, inplace=True)


away = df[['team_name_away', 'wl_away']].copy()
away['win'] = (away['wl_away'] == 'W').astype(int)
away['loss'] = (away['wl_away'] == 'L').astype(int)
away.rename(columns={'team_name_away': 'team'}, inplace=True)

all_games = pd.concat([home[['team', 'win', 'loss']], away[['team', 'win',
'loss']]])

result = all_games.groupby('team').sum().reset_index()

result['total_games'] = result['win'] + result['loss']
result['win_pct'] = result['win'] / result['total_games']

print(result)

# limpiar nombres de columnas
df.columns = df.columns.str.strip()

df['game_date'] = pd.to_datetime(df['game_date'], dayfirst=True)
# convertir game_date a datetime
df['game_date'] = pd.to_datetime(df['game_date'])

# filtrar fechas > 2018
df = df[df['game_date'] >= pd.Timestamp('2018-01-01')
```

```python
# armar datos para equipo local
home = df[['team_name_home', 'wl_home']].copy()
home['win'] = (home['wl_home'] == 'W').astype(int)
home['loss'] = (home['wl_home'] == 'L').astype(int)
home.rename(columns={'team_name_home': 'team'}, inplace=True)

# armar datos para equipo visitante
away = df[['team_name_away', 'wl_away']].copy()
away['win'] = (away['wl_away'] == 'W').astype(int)
away['loss'] = (away['wl_away'] == 'L').astype(int)
away.rename(columns={'team_name_away': 'team'}, inplace=True)

# unir locales y visitantes
all_games = pd.concat([home[['team', 'win', 'loss']], away[['team', 'win',
'loss']]])

# agrupar por equipo
result = all_games.groupby('team').sum().reset_index()
result['total_games'] = result['win'] + result['loss']
result['win_pct'] = result['win'] / result['total_games']

# mostrar resultado
print(result)

# guardar en CSV
result.to_csv("resumen_victorias_derrotas_2019_en_adelante.csv", index=False)

## Se realizó la diferencia entre ganados y perdido por equipo en los últimos cinco
años y se pidió que muestre los mejores 10
result['win_loss_diff'] = result['win'] - result['loss']


result = result.sort_values(by='win_loss_diff', ascending=False).head(10)


print(result[['team', 'win', 'loss', 'win_loss_diff']])


result[['team', 'win', 'loss',
'win_loss_diff']].to_csv("equipos_diferencia_win_loss.csv", index=False)
```

```
## Se muestra nuevo archivo
result['win_loss_diff'] = result['win'] - result['loss']


top10_diff = result.sort_values(by='win_loss_diff', ascending=False).head(10)


top_teams = top10_diff['team'].tolist()
print("Top 10 equipos:", top_teams)


df_top = df[
    (df['team_name_home'].isin(top_teams)) |
    (df['team_name_away'].isin(top_teams))
]


print(df_top.head())


df_top.to_csv("partidos_top10_equipos.csv", index=False)
```

El nuevo dataset contiene 4238 registros.

```
## Cantidad de filas nuevo CSV
num_filas = len(df_top)
print(f"Número de filas: {num_filas}")


tamano_bytes = df_top.memory_usage(deep=True).sum()
tamano_mb = tamano_bytes / (1024 ** 2)
print(f"Tamaño en memoria: {tamano_mb:.2f} MB")
```

5. Importación y modelado de datos en BigQuery

Se integraron y organizaron en BigQuery (Google Cloud) las bases de datos procesadas previamente, dejándolas listas para su posterior carga en Power BI. En esta etapa, se realizó la selección de las tablas relevantes para el análisis, que incluyen: *play_by_play, game, team_details, draft_combine_stats, draft_history, other_stats, common_player_info, player, salary, inactive_player y lines_score.*

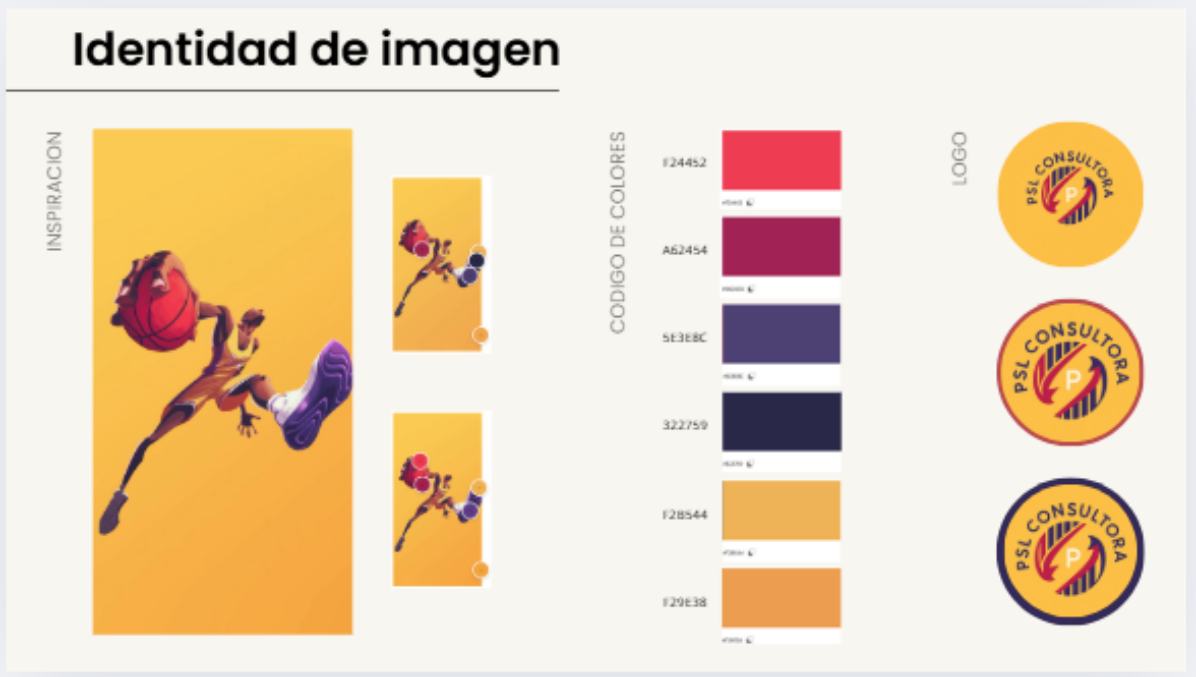
A partir de estas tablas, se identificaron las claves primarias (PK) y claves foráneas (FK) necesarias para construir el modelo de entidad-relación. Para el diseño visual del modelo se utilizó la herramienta Lucidchart. A continuación, se adjunta la imagen correspondiente.

6. Automatización de la ingesta de datos

Se trabajó en la automatización del proceso de ingesta de datos hacia BigQuery, con el objetivo de facilitar futuras actualizaciones y asegurar la escalabilidad del proyecto. Esta automatización permite cargar nuevos archivos CSV o actualizar los existentes de forma eficiente, reduciendo la intervención manual y minimizando errores. Se dejó documentado el procedimiento para replicar esta tarea en próximos ciclos de análisis.

7. Definición de la Identidad de imagen y logo

Se definió la identidad visual del proyecto, incluyendo la creación de un logo representativo y la elección de una paleta de colores, tipografía y estilo gráfico coherente. Esta identidad se aplicó en los distintos entregables, como presentaciones, dashboards y documentación, buscando reforzar la unidad estética y la profesionalidad del equipo. El logo fue diseñado considerando los valores del equipo y el propósito del análisis.




Identidad de imagen
TIPOGRAFIA
TITULOS Y SUBTITULOS
>>
Poppins
Poppins AaBbCc
Thin
ExtraLight
Light
Regular
Medium
SemiBold
Bold
ExtraBold
Black
EJEMPLO
Objetivo
TEXTO GENERAL
>>
Open Sans
Open Sans AaBbCc
Light
Regular
Bold
ExtraBold
EJEMPLO
Base de datos