Implementación de Machine Learning para el pronóstico de resultados en los torneos de Basketball de la división 1 de la NCAA

Felipe Ramírez Vargas

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor
Carmen Elena Patiño Rodriguez, Phd

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2023

Cita	(Ramírez Vargas, F. 2023)
Referencia **Estilo APA 7 (2020)**	Ramírez Vargas, F. (2023). *Implementación de Machine Learning para el pronóstico de resultados en los torneos de Basketball de la división 1 de la NCAA* Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte V.

Centro de Investigación Ambientales y de Ingeniería (CIA).

Centro de Documentación Ingeniería (CENDOI)

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.
Decano: Julio Cesar Saldarriaga Molina
Jefe departamento: Diego José Luis Botia Valderrama

Dedicatoria

Este trabajo se lo dedico a Dios, por permitirme cada día ser un mejor ser humano y obtener las capacidades necesarias para el desarrollo de un ejercicio profesional que trascienda e impacte en la sociedad. A mis padres Adriana Vargas Rivera e Iván Dario Ramirez Giraldo quienes siempre han estado en mi vida para apoyarme y hacer de mi un mejor ser humano y profesional y por último, a mi mascota Lucas, que desde el cielo me sigue acompañando en todo momento, en especial en mis momentos de concentración y estudio como lo hacía habitualmente cuando estuvo a mi lado.

Agradecimientos

Agradezco especialmente a la profesora Carmen Elena Patiño por guiarme en el desarrollo de la monografía, siempre brindarme los espacios necesarios para la solución de inquietudes, y permitiéndome desarrollar este ejercicio académico de la mejor manera. Agradezco a la Universidad de Antioquia por permitirme realizar la especialización en analítica y ciencia de datos, en donde adquirí conocimientos que me permitirán tener un mejor desarrollo de mi vida profesional, y buscar crecer desde el ámbito empresarial en este campo.

Tabla de contenido

Lista de tablas

Lista de Ilustraciones

Siglas, acrónimos y abreviaturas

PhD	Philosophiae Doctor
UdeA	Universidad de Antioquia
NCAA	National Collegiate Athletic

Resumen

La presente monografía es el resultado de la materialización del trabajo necesario para la especialización analítica y ciencia de datos cursada en la Universidad de Antioquia -UdeA. En este documento se da recuento de las metodologías y estrategias implementadas para generar el análisis necesario para la predicción de resultados en los torneos de basketball de la competencia National Collegiate Athletic - NCAA. Dicho trabajo y análisis se hace en el marco de la competencia de la plataforma kaggle.com

Como ejercicio académico se analizó la data suministrada y se generó una limpieza de la misma, posteriormente se plantean metodologías en pro de predecir resultados de los diversos encuentros deportivos, y se genera una evaluación de los modelos implementados. A su vez, se compara y alimenta el modelo con los ejercicios realizados por parte de los concursantes ganadores de la competencia en la plataforma Kaggle. De manera paralela, se investiga cómo la ciencia de datos, y las metodologías de machine learning benefician los negocios de casas de apuestas, en función de proyecciones de resultados que se pueden generar.

De esta manera, se busca generar comparaciones y relaciones en el comportamiento del modelo generado según variaciones que se hagan en el tratamiento de la data y variables estadísticas, con el fin de poder encontrar un modelo predictivo que se encuentre más ajustado a la realidad y a su vez, compararlo con el ejercicio implementado por los ganadores de la competencia y nutrirlo con alternativas que hayan propuesto estos competidores.

Las predicciones deportivas no son únicamente un modelo de negocio que se implementa por casas de apuestas, sino que, a su vez permiten a los equipos de diferentes deportes tomar decisiones frente a las estrategias a implementar a nivel deportivo y gerencial, las cuales repercuten directamente en las finanzas de esta industria.

Palabras clave: Machine Learning, torneo, NCAA, predicción.

Github: https://github.com/FelipeRam22/Monografiafinal

Abstract

This monograph is the result of the materialization of the work necessary for the analytical and data science specialization taken at the University of Antioquia -UdeA. This document provides an account of the methodologies and strategies implemented to generate the analysis necessary to predict results in the basketball tournaments of the National Collegiate Athletic - NCAA competition.

Said work and analysis is done within the framework of the competence of the kaggle.com platform as an academic exercise, the data provided was analyzed and a cleaning of it was generated. Subsequently, methodologies were proposed to predict the results of the various sporting events, and an evaluation of the implemented models was generated. In turn, the model is compared and fed with the exercises performed by the winning contestants of the competition on the kaggle platform. In parallel, we investigate how data science and machine learning methodologies benefit betting house businesses, based on projections of results that can be generated.

In this way, it seeks to generate comparisons and relationships in the behavior of the generated model according to variations that are made in the treatment of data and statistical variables, in order to be able to find a predictive model that is more adjusted to reality and its time, compare it with the exercise implemented by the winners of the competition and nourish it with alternatives that these competitors have proposed. Sports predictions are not only a business model that is implemented by betting houses, but they also allow teams from different sports to make decisions regarding the strategies to be implemented at a sporting and managerial level, which have a direct impact on the finances of this industry.

*Keywords***:** Machine Learning, tourney, NCAA, prediction

1. Descripción del problema

El proyecto a ejecutar es la solución a una de las competencias con incentivo económico que ofrece la plataforma Kaggle.com denominado “March Machine Learning Mania 2023” (March Machine Learning Mania 2023, 2023) Por medio de los dataset suministrados se buscará generar una predicción utilizando metodologías de Machine learning para determinar la probabilidad de ganar un equipo a otro en la competencia National Collegiate Athletic - NCAA, en el marco de un ejercicio académico.

La solución al reto de la competencia de la plataforma kaggle.com, se hace en pro de atender a la monografía necesaria para optar por el título de especialista en analítica y ciencia de datos de la Universidad de Antioquia. De esta manera, se buscará implementar los conocimientos adquiridos a lo largo del programa académico, implementando un modelo de machine learning que permita atender la problemática planteada a través de la plataforma Kaggle. Adicional a ello, la competencia tuvo unos equipos ganadores, por lo que, se comparará con los modelos implementados por dichos equipos y se alimentará el modelo con los ejercicios que se considere beneficien el modelo a ejecutar. De esta manera, desde la óptica de un ejercicio académico se puede vislumbrar los beneficios de diferentes metodologías de aprendizaje automático implementadas por otros participantes y aprender de la utilidad de las mismas.

A su vez, desde la óptica del marketing y gerencia de proyectos, se buscará analizar el beneficio que tienen los ejercicios de implementación de machine learning para la predicción de resultados de eventos deportivos en el marco empresarial, analizando los beneficios que pueden derivarse de la creación de estos modelos y su impacto en las métricas de diferentes sectores empresariales como lo son: casas de apuestas, sponsors, y equipos deportivos. A su vez, como este tipo de análisis pueden mejorar el “core” del negocio y por medio de campañas de marketing beneficiar el “engagement” del público objetivo con la marca.

1.1. Problema de negocio

Si bien el planteamiento del problema, se hace en el marco de un ejercicio académico derivado de un reto establecido en la plataforma kaggle.com, los ejercicios y modelos que se implementan por medio de la analítica de datos frente a los eventos deportivos permiten a clubes, casas de apuestas, inversionistas, entre otros, tener medidas objetivas para la toma de decisiones.

En el ámbito empresarial se pueden definir en primera instancia tres principales tipos de industrias que se pueden ver beneficiadas conforme a la implementación de modelos predictivos fundamentados en machine learning para la predicción de resultados deportivos, de esta manera, se relacionan las siguientes:

- Casas de apuestas: Para las casas de apuestas, es netamente importante contar con modelos predictivos que les permitan definir qué equipo tienen una mayor probabilidad de ganar frente a otro equipo y de esta manera, probabilísticamente distribuir los montos de las recompensas a entregar conforme a un acierto en una apuesta por parte de los usuarios.

- Sponsors: Para las empresas que buscan generar un patrocinio a un equipo deportivo, es netamente importante conocer de manera preliminar los equipos que tienen una mayor probabilidad de ganar, toda vez que, en función de esto se puede generar una inversión como sponsor para los mismos. Es habitual que, el equipo con mayor probabilidad de ganar sea el equipo que su patrocinio sea más costoso, sin embargo, utilizar modelos de machine learning permiten al sponsor visibilizar el abanico de posibilidades de inversión frente a otros equipos que su patrocinio no sea tan elevado y que su probabilidad de ganar sea alta.

- Equipos deportivos: En función de modelos predictivos basados en metodologías de machine learning, los equipos deportivos pueden visualizar la probabilidad de ganar a otro equipo. De esta manera, a nivel deportivo pueden ver que estrategias están implementado los otros equipos, así como el tipo de jugadores que tienen, y otras variables derivadas. De esta manera, a nivel directivo se podrán implementar estrategias que beneficien a los

equipos deportivos a la hora de enfrentar a otro equipo, conforme a la probabilidad de ganarle.

De esta manera, implementado conocimientos de analítica y ciencia de datos, se desarrollan metodologías que permiten atender diversas problemáticas en el contexto empresarial, para este caso, en la industria deportiva.

1.2. Aproximación desde la analítica de datos

El proyecto relacionado, se hace en el marco de la competencia de la plataforma de retos Kaggle.com, en donde con conocimientos en analítica de datos se busca dar respuesta a diversas problemáticas por medio de modelos de machine learning.

Para este caso, se buscará implementar un modelo predictivo de machine learning con el objetivo de determinar la probabilidad de que un equipo pueda ganarle a otro en el marco de los torneos de Basketball de la división 1 de la NCAA.

De esta manera, a través de un análisis estadístico descriptivo de las bases de datos, un correcto entendimiento de sus variables, la limpieza y manipulación de los datos, el entrenamiento de modelos de machine learning y su análisis de desempeño, se podrá determinar la probabilidad de que un equipo pueda ganarle a otro.

1.3. Origen de los datos

Se cuenta con las bases de datos de las temporadas jugadas desde el año 1985 para la categoría masculina y femenina, dicha data es obtenida de la plataforma Kaggle.com.

La data suministrada se compone principalmente de la recopilación de diferentes bases de datos en formato csv, que pueden llegar a ser relevantes para el análisis del problema e implementación de modelos de machine learning que cumplan con el objetivo de delimitar la probabilidad de ganarle a un equipo.

De esta manera, se tienen datos representativos del torneo NCAA tanto para la competencia de hombres como para mujeres los cuales se enmarcan en la temporada de 1985 para hombres (el primer año que la NCAA® tuvo un torneo masculino de 64 equipos) y la temporada de 1998 para mujeres y finalizan en el año 2022.

1.4. Métricas de desempeño

Teniendo en cuenta que el modelo predictivo se va hacer como una regresión, se implementarán métricas de desempeño como lo es MAE, SME los cuales se derivan de determinar la diferencia que existe entre los datos predichos frente a los reales. Adicionalmente, se analizarán otras métricas utilizadas en la literatura y academia.

- Mean Absolute Error - MAE.

"El error absoluto promedio proporciona el promedio de la diferencia absoluta entre la predicción del modelo y el valor objetivo". (IBM Error absoluto promedio, 2021)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|yi - \hat{y}i|$$

- Mean Squared Error - MSE.

"Representa la distancia al cuadrado entre los valores reales y predichos".

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i)^2$$

Desde la óptica del marketing y la inteligencia del modelo de negocio, es importante implementar métricas que permitan vislumbrar cómo dichas estrategias de uso de modelos de machine learning aumentan la productividad, engagement, Core y otras variables en el marco empresarial.

Las métricas de MAE y MSE son las más utilizadas en modelos de regresión y las cuales buscan calcular la distancia entre el valor calculado vs los reales, por ende, son útiles para este ejercicio. Sin embargo, se ha evidenciado que para este ejercicio el MSE es la medida más apta para realizar la medición.

Bajo la óptica del marketing, es sumamente importante generar acciones que permitan pasar del reconocimiento de la marca hasta la fidelización. A continuaciones algunas medidas, que pueden ser útiles desde el enfoque del mercadeo, derivado de la acción que se tome conforme a los resultados del modelo (Tonatiuh, 2020):

- **Reconocimiento.**

En la fase de reconocimiento empresarial, es fundamental poder medir métricas como lo son: Alcance, Frecuencia, views.

- **Consideración.**

Es la fase en la cual se genera contenido en pro de obtener reacciones y acercamiento con la marca, como métricas se pueden vislumbrar: likes, clics, reacciones, etc.

- **Conversión.**

Es la fase en la cual se percibe la intención de la implementación de campañas de marketing en donde los usuarios se adhieren a la marca, sus métricas: número de compras, descargar una app, entre otras.

- **Fidelización.**

Es una de las fases más importantes de las campañas de marketing, toda vez que fideliza al cliente, sus métricas: cupones usados, menciones de consumidores, etc.

- **Engagement**

De esta manera, calcular el engagement por parte de los sponsors que patrocinan algún equipo derivado del análisis mediante modelos de machine learning, permitirá evidenciar la utilidad de haber patrocinado a “X” o “Y” equipo. Existen diversas maneras de medirlo las cuales varían conforme a la campaña publicitaria que se implemente y el medio que se utilice, sin embargo se puede destacar:

$$Egagement = \frac{Interacciones}{Alcance} * 100$$

De esta manera, se podría visualizar por parte de las empresas conforme a la implementación de un patrocinio derivado de un análisis mediante machine learning, si el engagement aumenta conforme a las variaciones en el alcance y las interacciones que se presenten.

Todas las métricas anteriormente relacionadas, se deberán medir antes y después de la acción ejecutada por las empresas derivadas de la respuesta del modelo de machine learning, es decir, si el modelo arrojó que existen 5 equipos que tienen la mayor probabilidad de ganar un partido, la compañía podrá patrocinar uno de ellos, y de esta manera, se calcula antes y después del patrocinio las métricas anteriormente expuestas.

2. Objetivos

2.1.Objetivo general

Generar un modelo de machine learning de regresión que facilite la predicción e interpretabilidad de los resultados para torneos deportivos en competencias masculinas y femeninas en el marco de la competencia de la plataforma kaggle.com, comparándolo y alimentándose con los resultados de los competidores ganadores y analizando su beneficio desde una óptica de marketing y gerencia de proyectos.

2.2.Objetivos específicos

- Analizar la data histórica que se tiene para las series femenina y masculina de la división 1 de la NCAA, generando limpieza, extracción, concatenación, y todo el procesamiento que se requiera.
- Conocer el comportamiento y la relación de las variables que afectan el resultado final de un encuentro deportivo de las series masculina y femenina para los torneos derivados de la NCAA.
- Mejorar la interpretabilidad mediante la creación de variables adicionales que permitan una mejor representación e interpretabilidad de los datos a analizar permitiendo un modelamiento efectivo en aras de determinar la probabilidad de que un equipo pueda ganar a otro.
- Analizar el desempeño del modelo, comparándolo con los modelos generados por los equipos ganadores y alimentándose de los mismos, buscando establecer los beneficios que se derivan de la implementación de modelos de machine learning en el ámbito empresarial desde la óptica de marketing y gerencia de proyectos.

3. Datos

3.1. Datos originales

Se utilizaron las siguientes bases de datos de la totalidad de las bases suministradas por la plataforma kaggle, que permitieron crear el modelo de machine learning que se adaptara a la naturaleza del problema.

- MNCAATourneyCompactResults
- MNCAATourneySeeds
- MRegularSeasonCompactResults
- WNCAATourneyCompactResults
- WNCAATourneySeeds
- WRegularSeasonCompactResults
- SampleSubmission2023

Para cada una de las bases de datos, se comparte una corta descripción relacionada en kaggle:

MNCAATourneyCompactResults: Esta base de datos identifica los resultados juego por juego de muchas temporadas de datos históricos, comenzando con la temporada de 1985 para hombres (el primer año que la NCAA® tuvo un torneo masculino de 64 equipos) y la temporada de 1998 para mujeres.

MNCAATourneySeeds: Estos archivos identifican las cabezas de serie de todos los equipos en cada torneo de la NCAA, para todas las temporadas de datos históricos.

MRegularSeasonCompactResults - WRegularSeasonCompactResults: Estos archivos identifican los resultados juego por juego de muchas temporadas de datos históricos, comenzando con la temporada de 1985 para hombres (el primer año que la NCAA® tuvo un torneo masculino de 64 equipos) y la temporada de 1998 para mujeres.

WNCAATourneyCompactResults: Estos archivos identifican los resultados del torneo NCAA® juego por juego para todas las temporadas de datos históricos.

WNCAATourneySeeds: Estos archivos identifican las semillas de todos los equipos en cada torneo de la NCAA®, para todas las temporadas de datos históricos. Entiendase por semilla, un código que identifica la región.

SampleSubmission2023: Estos archivos ilustran el formato de archivo de presentación para la competencia de preparación y la competencia de 2023. Reflejan la sumisión más simple posible: se predice un porcentaje de victorias del 50% para cada enfrentamiento posible.

Las bases de datos utilizadas, se encuentran en formato csv, y a continuación se genera una descripción de las mismas mediante tablas:

Tabla 1. Descripción bases de datos

Base de datos	Número de variables	Número de Variables cuantitativas	Número de variables categóricas	Cantidad de registros	Tamaño del Archivo (kB)
MNCAATourneyCompactResults.csv	8	7	1	2384	71.67
MNCAATourneySeeds.csv	3	2	1	2490	37.49
MRegularSeasonCompactResults.csv	8	7	1	181682	5100
WNCAATourneyCompactResults.csv	8	7	1	1516	45.57
WNCAATourneySeeds	3	2	1	1608	24.16
WRegularSeasonCompactResults	8	7	1	126173	3690
SampleSubmission2023	2	1	1	130683	2610

3.2. Datsets

En pro de generar los análisis respectivos y la creación del modelo de machine learining necesario, se genera una concatenación de las bases de datos “Seeds” tanto para hombres como para mujeres generando un dataframe llamado df_seeds. Paralelamente, se genera una concatenación en aras de determinar un dataframe que consolide los resultados por temporada tanto de mujeres como de hombres.

```
df_season_results = pd.concat([

pd.read_csv("MRegularSeasonCompactResults.csv"),

pd.read_csv("WRegularSeasonCompactResults.csv"),
], ignore_index=True)

df_season_results.drop(['NumOT', 'WLoc'], axis=1, inplace=True)

df_season_results.head()
```

Tabla 2. Dataframe Season.

	Season	DayNum	WTeamID	WScore	LTeamID	LScore
0	1985	20	1228	81	1328	64
1	1985	25	1106	77	1354	70
2	1985	25	1112	63	1223	56
3	1985	25	1165	70	1432	54
4	1985	25	1192	86	1447	74
5	1985	25	1218	79	1337	78
6	1985	25	1228	64	1226	44
7	1985	25	1242	58	1268	56
8	1985	25	1260	98	1133	80
9	1985	25	1305	97	1424	89

Adicionalmente se genera una descripción del nuevo dataframe:

```
df_season_results.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307855 entries, 0 to 307854
Data columns (total 6 columns):
 #   Column   Non-Null Count   Dtype
---  ------   --------------   -----
 0   Season   307855 non-null  int64
 1   DayNum   307855 non-null  int64
 2   WTeamID  307855 non-null  int64
 3   WScore   307855 non-null  int64
 4   LTeamID  307855 non-null  int64
 5   LScore   307855 non-null  int64
dtypes: int64(6)
memory usage: 14.1 MB
```

Ilustración 1. Descripción resultados Dataframe

Ahora bien, en aras de poder realizar un modelo de machine learning que lograra generar una predicción objetiva, se crearon algunas variables que permitieran una mejor interpretabilidad y manipulación de los datos, teniendo en cuenta a su vez, cuáles variables habían beneficiado a otros competidores. De esta manera, se relacionan algunas de estas:

- **Diferencia Puntaje:** Esta variable lo que busca es calcular la diferencia del puntaje del equipo que gana frente al equipo que pierde.

```
df_season_results['Diferencia Puntaje'] = df_season_results['WScore'] - df_season_results['LScore']
df_season_results.head()
```

Así las cosas agrupando por equipos y sacando la media para cada uno, podríamos identificar cual es la diferencia de puntaje media cuando gana cada equipo. De esta manera, entre más grande sea esta media, indica que, el equipo normalmente gana por muchos puntos más que sus rivales. Al crear la nueva variable “Diferencia Puntaje”, se hace una prueba de chequeo para corroborar que quedó bien, puesto que, esta variable nunca puede ser menor que 0, si los equipos

empatan dicha variable adquiere un valor de 0, sin embargo si hay un ganador la variable siempre debe ser mayor que 0.

```
df_season_results[df_season_results['Diferencia Puntaje']<0].count()
```

```
Season               0
DayNum               0
WTeamID              0
WScore               0
LTeamID              0
LScore               0
Diferencia Puntaje   0
dtype: int64
```

Ilustración 2. Variables que cumplen condición <0

De esta manera se corrobora que la variable queda bien planteada. Además de manera preliminar, podría pensarse que, aquellos equipos cuya media de la variable “Diferencia Puntaje” sea mayor, tendrían en primera instancia mayor probabilidad de ganar. Por otra parte se hace la diferencia de puntaje para las partidas pérdidas.

La relevancia de las variables anteriormente calculadas se plasma mediante el siguiente ejemplo:

- ✓ Media diferencia Equipo A Gana = 12 puntos
- ✓ Media diferencia Equipo B Gana = 7 puntos
- ✓ Media diferencia Equipo A Pierde =2 puntos
- ✓ Media diferencia Equipo B Pierde = 5 puntos

De esta manera, en principio, en un enfrentamiento entre el equipo A y el B, el equipo A “ganaría", toda vez que, normalmente cuando gana hace más puntos que el equipo B y cuando pierde recibe menos puntos que este.

- **Partidas ganadas y perdidas.**

Con la función group.by se agrupa de tal manera que, se identifique por equipo cuántas partidas ha ganado y cuantas partidas ha perdido.

- **Proporción ganadas.**

Identifica de la totalidad de partidas jugadas, cuántas ha ganado, en teoría, entre mayor sea la proporción más posibilidad tiene de ganarle a otro equipo.

- **Winning Ratio o Promedio de la diferencia.**

Es una variable implementada por los competidores ganadores para tener en primera instancia un acercamiento a una "tasa de victoria", la cual es una variación a "Winning Percentage" (Wikipedia contributors, 2023), el promedio de la diferencia, es calculado de la siguiente manera:

```
df_features_season['Promedio diferencia'] = (

(df_features_season['Partidas    Ganadas']    *    df_features_season['Media
Diferencia Puntaje Partidas ganadas'] -

df_features_season['Partidas    Perdidas']    *    df_features_season['Media
Diferencia Puntaje perdidas'])

   / (df_features_season['Partidas Ganadas'] + df_features_season['Partidas
Perdidas'])
```

El denominador siempre será positivo porque es la totalidad de partidas jugadas por los equipos, sin embargo el numerador puede ser positivo o negativo.

Tanto la variable "Partidas Ganadas" como "Partidas Perdidas" es un entero positivo, por ende, cuando las partidas ganadas multiplicadas por la Diferencia Puntaje Partidas ganadas sea mayor a las partidas pérdidas multiplicadas por la Diferencia de Puntaje Partidas pérdidas será positivo. Por otra parte, cuando las partidas pérdidas multiplicadas por la Diferencia de Puntaje Partidas pérdidas sea más grande que las partidas ganadas multiplicadas por la Diferencia Puntaje Partidas ganadas, el valor será negativo.

Cuando la variable “Promedio diferencia” es positiva, significa que el equipo habitualmente gana en mayor proporción respecto a las veces que pierde y además, normalmente adquiere más puntos de ventaja cuando gana que puntos de desventaja cuando pierde, de manera viceversa cuando es negativo.

Para un equipo deportivo lo ideal es que la variable Promedio diferencia sea grande en magnitud y positiva, por el contrario cuando es negativa y en una gran magnitud denota que el equipo no tiene el mejor rendimiento deportivo.

- **Diferencia proporción ganadas “Proporcion_ganadasDiff”**

Esta variable calcula la diferencia de la proporción de partidas ganadas de un equipo “A” frente a un equipo “B”. Es decir, si el equipo A y B se enfrentan, si la “Proporcion_ganadasDiff” es positiva significa que el equipo A normalmente ha ganado más veces que el equipo B.

- **Promedio diferenciaDiff.**

Es una variable que calcula la diferencia entre “Winning Ratio ó Promedio de la diferencia” de un equipo A frente a un equipo B, así las cosas, si la diferenciaDiff entre el equipo A frente al B es positiva, indica que el equipo A tiene mayor probabilidad de ganarle al equipo B, toda vez que, habitualmente el equipo A gana en mayor proporción respecto a las veces que pierde y además, normalmente adquiere más puntos de ventaja cuando gana que puntos de desventaja cuando pierde.

- **Datos de entrenamiento - datos de prueba.**

Generalmente para realizar modelos de machine learning, los datos que se tienen se particionan para definir una cantidad que será utilizada para entrenar el modelo y otra que es utilizada para evaluarlo, habitualmente estos dos grupos son definidos como “data train” y “data

test". Para este caso de estudio, contamos con una base de datos denominada "SampleSubmission2023", la cual cuenta con una diversa cantidad de encuentros deportivos que deberemos validar.

Es importante precisar que para este caso de estudio, se generan particiones conforme se avanza en la temporada en aras de tener en cuenta la temporalidad de los datos, es decir, si se quisiera predecir un evento deportivo para el 2018, la data de entrenamiento será aquella que se tenga hasta el 2017, y la de validación serán los eventos deportivos que se tienen únicamente para el 2018.

La anterior condición permite utilizar únicamente los datos anteriores a la temporada en que se requiera evaluar, esto debido a que, no sería beneficioso utilizar datos de 2019 para predecir un evento del 2018, toda vez que, los jugadores y el rendimiento deportivo no es el mismo, además para 2018, las condiciones deportivas del año 2019 no eran conocidas, ni los jugadores que entrarían al equipo.

De esta manera, en aras de ejemplificar la anterior situación, si se quiere validar un evento deportivo para el 2004, se utilizaría la data hasta el 2003 y se validaría con la de 2004 respectivamente, de manera similar, si el evento deportivo fuera en el 2017, se entrenaría el modelo con la data hasta el 2016 y se validaría con los datos de 2017.

3.3. Analítica descriptiva

En función de generar un análisis descriptivo eficaz, primero se analizará las bases de datos suministradas por el ejercicio y posteriormente, se analizarán los datasets generados, así como las variables generadas que permiten una mayor interpretabilidad de los datos y modelado.

Una de las base de datos más importantes en la modelación, fue la denominada "MRegularSeadonCompactResults", la cual consolida los resultados de los encuentros deportivos para diferentes temporadas, al igual que para hombres existe una para los encuentros deportivos de mujeres, la cual es utilizada paralelamente.

- **MRegularSeasonCompactResults.csv**

Esta base de datos es de gran utilidad para el ejercicio, por ende queremos consultar la cantidad de datos que tiene, el tipo de variables que enmarca y sus características principales.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 181682 entries, 0 to 181681
Data columns (total 8 columns):
 #   Column   Non-Null Count   Dtype
---  ------   --------------   -----
 0   Season   181682 non-null  int64
 1   DayNum   181682 non-null  int64
 2   WTeamID  181682 non-null  int64
 3   WScore   181682 non-null  int64
 4   LTeamID  181682 non-null  int64
 5   LScore   181682 non-null  int64
 6   WLoc     181682 non-null  object
 7   NumOT    181682 non-null  int64
dtypes: int64(7), object(1)
memory usage: 11.1+ MB
```

Ilustración 3. Descripción MRegularSeason

De manera paralela, se visualiza su estructura como base de datos

Tabla 3. Estructura Base de datos M Regular Season

	Season	DayNum	WTeamID	WScore	LTeamID	LScore	WLoc	NumOT
0	1985	20	1228	81	1328	64	N	0
1	1985	25	1106	77	1354	70	H	0
2	1985	25	1112	63	1223	56	H	0
3	1985	25	1165	70	1432	54	H	0
4	1985	25	1192	86	1447	74	H	0

La mayoría de sus variables son categóricas así las mismas contengan datos numéricos, sin embargo es importante conocer las medidas estadísticas de las variables cuantitativas, por ende:

Tabla 4. Descripción Wscore y Lscore Hombres

	WScore	LScore
count	181682.000000	181682.000000
mean	76.746128	64.671569
std	11.897763	11.236768
min	34.000000	20.000000
25%	69.000000	57.000000
50%	76.000000	64.000000
75%	84.000000	72.000000
max	186.000000	150.000000

De esta manera, en principio para una apuesta deportiva para torneos masculinos se podría pensar en que, cuando gana un equipo, normalmente gana el partido con 77 puntos y cuando pierde normalmente lo hace con un puntaje de 65 puntos.

Análogamente se puede visualizar que, la media para la variable “WScore” es de 76,75 y su mediana es de 76, por lo que, dichos valores son bastante similares indicando que la distribución de la data, es simétrica, dicho situación ocurre de igual manera para la variable “LScore”, en donde dichos valores son 64,67 para su media y 64 para su mediana respectivamente. La condición de simetría indica que, al dividirla por el eje de simetría, la distribución se distribuiría aproximadamente un 50% a la izquierda y un 50% a la derecha, presentando una forma similar en ambos costados.

También se puede percibir que, los valores máximos son lejanos al correspondiente al 75%, sin embargo, esto se deriva en que, existen partidos en donde se marcan una gran cantidad de puntos.

El ejercicio es idéntico para la base de datos de las mujeres, por lo que, a continuación se relaciona al igual que para los hombres, su comportamiento estadístico.

Tabla 5.Descripción Wscore y Lscore Mujeres

	WScore	LScore
count	126173.000000	126173.000000
mean	71.849437	57.520856
std	11.338444	10.804215
min	30.000000	11.000000
25%	64.000000	50.000000
50%	71.000000	57.000000
75%	79.000000	64.000000
max	140.000000	130.000000

De esta manera, se puede percibir que habitualmente en los eventos deportivos para hombres cuando se gana, se gana por un puntaje mayor que en los eventos deportivos para las mujeres, mientras que, en los eventos deportivos para mujeres cuando se pierde se pierde con un puntaje inferior. De esta manera, se puede evidenciar que en los eventos deportivos masculinos, se marcan más cantidad de puntos.

En función de las anteriores bases de datos, se consolida un único dataframe que contiene los datos tanto de la categoría hombres como de mujeres. Como se evidencio en el apartado 3.2, se crearon una serie de variables que permitieron un mejor modelamiento del problema, así las cosas, a continuación se relaciona una descripción estadística del dataframe creado junto con dichas variables.

El Dataframe que consolida se denomina dftor, y contiene las siguientes variables:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5910 entries, 0 to 2954
Data columns (total 16 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Season                   5910 non-null   int64
 1   DayNum                   5910 non-null   int64
 2   TeamIdA                  5910 non-null   int64
 3   ScoreA                   5910 non-null   int64
 4   TeamIdB                  5910 non-null   int64
 5   ScoreB                   5910 non-null   int64
 6   SeedA                    5910 non-null   int64
 7   SeedB                    5910 non-null   int64
 8   Proporcion_ganadasA      5910 non-null   float64
 9   Promedio diferenciaA     5910 non-null   float64
 10  Proporcion_ganadasB      5910 non-null   float64
 11  Promedio diferenciaB     5910 non-null   float64
 12  Proporcion_ganadasDiff   5910 non-null   float64
 13  Promedio diferenciaDiff  5910 non-null   float64
 14  ScoreDiff                5910 non-null   int64
 15  WinA                     5910 non-null   int64
dtypes: float64(6), int64(10)
memory usage: 784.9 KB
```

Ilustración 4. Variables dataframe consolidado

A continuación se relaciona los primeros datos del dataframe Dftor, en lo que corresponde a las variables creadas:

Tabla 6. Primeros datos Dftor

Proporcion_ganadasA	Promedio diferenciaA	Proporcion_ganadasB	Promedio diferenciaB	Proporcion_ganadasDiff	Promedio diferenciaDiff	ScoreDiff	WinA
0.655172	10.551724	0.857143	17.071429	-0.201970	-6.519704	12	1
0.629630	4.481481	0.593750	4.968750	0.035880	-0.487269	3	1
0.827586	14.931034	0.620690	4.413793	0.206897	10.517241	31	1
0.939394	17.818182	0.750000	9.642857	0.189394	8.175325	12	1
0.870968	31.032258	0.620690	2.344828	0.250278	28.687430	51	1

El dataframe consolidado contiene diferentes variables, siendo las más representativas las creadas, las cuales son: Proporcion_ganadasA,Promedio DiferenciaA ,Proporcion_ganadasB, Promedio DiferenciaB, Proporcion_ganadasDiff, Promedio diferenciaDiff, ScoreDiff, WinA.

Así las cosas, se genera la comparación para las mujeres como para los hombres.

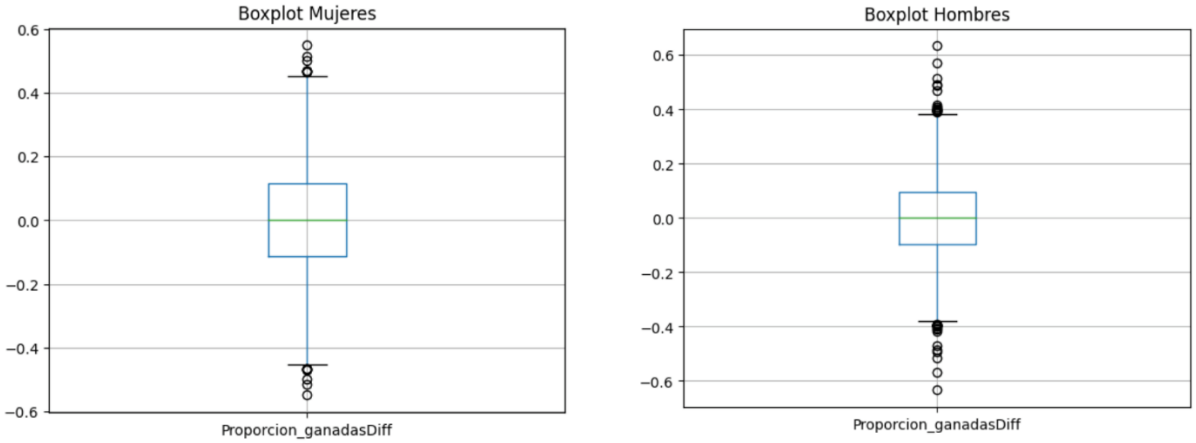

Ilustración 5. Comparación Proporcion_ganadasDiff Mujeres vs Hombres

La proporción de partidas ganadas, como se indicó anteriormente, evidencia frente al total de juegos, cuál es la proporción en que un equipo ganó. La "Proporcion_ganadasDiff", indica la diferencia de proporción de las partidas ganadas entre un equipo A y un equipo B.

La distribución de la data será simétrica, toda vez que, cuando se enfrenta el equipo A con el equipo B, la "Proporcion_ganadasDiff", será "x" valor positivo si A tiene una mayor proporción que B, sin embargo, cuando se enfrenta B contra A, la magnitud será la misma, pero el resultado será negativo, es decir "-x". De esta manera, el valor de la mediana tanto para hombres como para mujeres es de 0, esto debido a que, para cada valor positivo existirá uno negativo.

Para el caso de los hombres, existieron valores extremos superiores a 0.6 condición que evidencia que en algunos encuentros deportivos, persistía una gran diferencia de partidas ganadas entre un equipo y otro, es decir, para algunos partidos, era demasiado significativa la diferencia competitiva, lo cual en teoría evidenciaba una mayor probabilidad de ganar de un equipo frente a otro. El percentil 75% para el caso de los hombres indica que existen valores hasta 0.096, mientras

que para el caso de las mujeres, dicho valor es de 0.11. En el caso de los hombres, la mayor parte de los datos se agrupa en valores inferiores a 0.4, mientras que para las mujeres es un poco superior. Lo anterior indica que, en los enfrentamientos deportivos para el caso de las mujeres hay un mayor desbalance en cuanto a las proporciones ganadas, evidenciado a la hora de un enfrentamiento, mayor fortaleza de un equipo frente a otro.

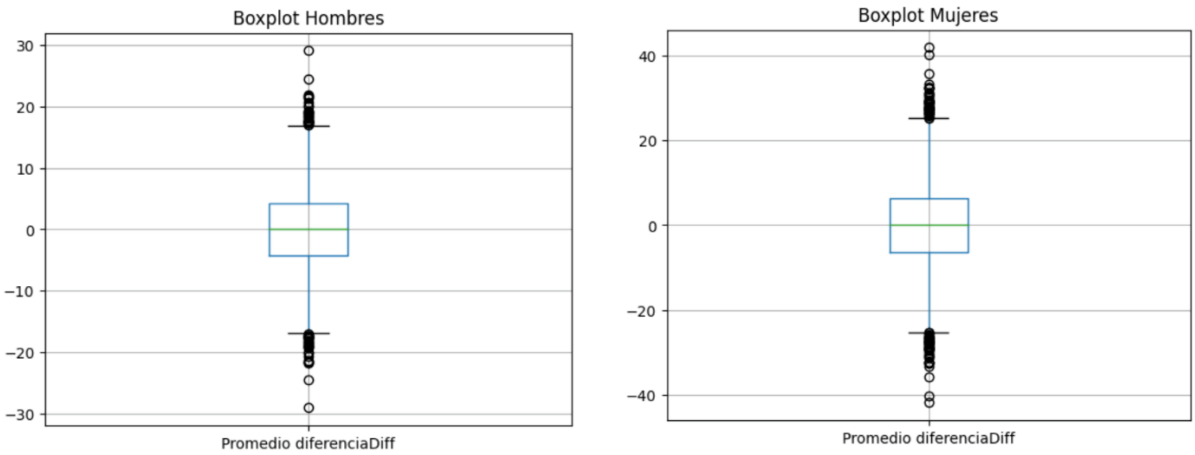

Ilustración 6. Comparación Promedio-Diff Mujeres vs Hombres

El "Promedio diferenciaDiff" calcula la diferencia entre los promedios de Winning Ratio o Promedio de la diferencia, lo cual expresa que a mayor valor, existe una mayor probabilidad de ganar de un equipo frente a otro.

En la comparación que se realiza, se puede evidenciar que existe una diferencia deportiva superior para la categoría de mujeres frente a la de los hombres, toda vez que, en la gráfica anteriormente expuesta, el percentil 75 en la categoría de mujeres se ubica en un valor del 6.31% y se perciben datos extremos superiores al 40%. Para el caso de los hombres, el 75% de los datos son inferiores a 4,22% y el dato más extremo es cercano al 30%.

Desde la perspectiva de marketing, en principio, un sponsor tendría más probabilidad de conocer quien va ganar un evento deportivo en la categoría de mujeres frente a la categoría de los hombres, de esta manera podría segmentar su patrón de inversión o tener una primera noción de qué categoría patrocinar suponiendo que sus recursos son limitados.

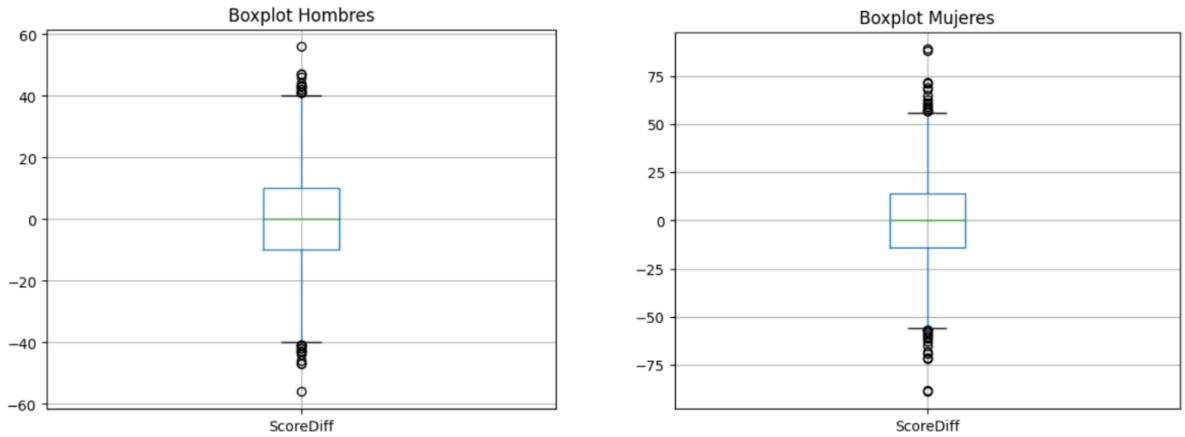

Ilustración 7. Comparación Score Diff Hombres vs Mujeres.

Para el caso de los hombres, en el 75% de los casos la diferencia de puntajes en los enfrentamientos deportivos es inferior a 10 puntos, en contraste para las mujeres la diferencia en los encuentros no sobrepasa los 14 puntos. En el caso de los hombres existen valores extremos de 40 puntos de diferencia y en las mujeres de 52 puntos. En los datos históricos, se visualiza que si bien es atípico, existieron jornadas en las que la diferencia de un partido de hombres pudo ser de 57 puntos, mientras que uno de mujeres es incluso superior a 75 puntos. De esta manera una casa de apuestas por ejemplo, podría identificar que si se trata de un evento deportivo masculino, normalmente en los partidos no existirán diferencias de puntajes superiores a 10 puntos, mientras que para las mujeres no es común diferencias superiores a 14 puntos.

4. Proceso de analítica

4.1. Pipeline principal

A continuación, se relaciona la ruta de atención que se le debe dar al proyecto en aras de cumplir con el objetivo.

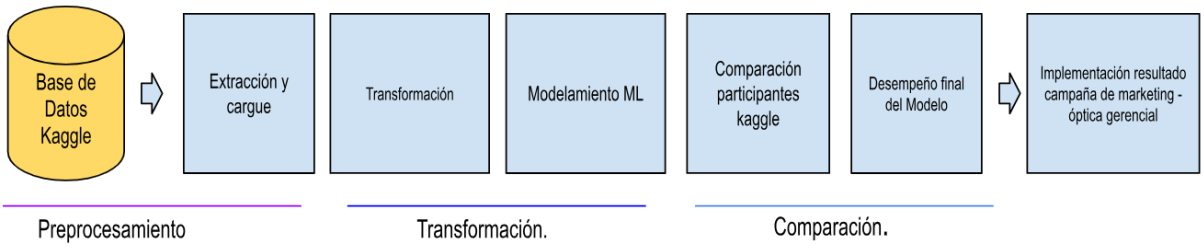

Ilustración 8. Pipeline principal

1. Como primer hito del desarrollo del pipeline, se ingresa a la plataforma kaggle.com y se consulta la base de datos para el desarrollo de la competencia kaggle mania.
2. Se genera la extracción y limpieza de las bases de datos y el cargue a Google Collab para poder trabajar en ella y generar el modelo predictivo.
3. Se generan los dataframes necesarios para poder generar el modelo predictivo.
4. Una vez generado el modelo predictivo se comparan los resultados con los generados por otros competidores de kaggle y se analiza el beneficio de su implementación.
5. Una vez se tiene el modelo, se analiza el desempeño del modelo y sus métricas
6. Se procede a validar el beneficio que trae la implementación del modelo desde la óptica gerencial y de marketing.

4.2. Preprocesamiento

En aras de poder generar el modelo de machine learning que se adhiriera a la necesidad de crear un modelo predictivo para eventos deportivos que permita aumentar el engagement empresarial, en primera instancia se buscó determinar el estado de la data.

Así las cosas, en primera instancia se validaron cada una de las bases de datos, sus características y si contaban con datos nulos o errados.

df_seeds.isnull().count()

Season 4098

Seed 4098

TeamID 4098

dtype: int64

Posteriormente, se crearon diferentes dataframe en aras de que el procesamiento pudiera generarse de una manera más eficiente, la agrupación a su vez se filtró para que quedara la información consignada de la manera más compacta posible.

```
df_features_season_ganadores = df_season_results.groupby(['Season',
'WTeamID']).count().reset_index()[['Season',
'WTeamID']].rename(columns={"WTeamID": "TeamID"})


df_features_season_perdedores = df_season_results.groupby(['Season',
'LTeamID']).count().reset_index()[['Season',
'LTeamID']].rename(columns={"LTeamID": "TeamID"})
```

```
dftor = pd.merge(
   dftor,
   df_features_season,
   how='left',
   left_on=['Season', 'LTeamID'],
   right_on=['Season', 'TeamID']
).rename(columns={
   'Partidas Ganadas': 'Partidas GanadasL',
   'Partidas Perdidas': 'Partidas PerdidasL',
   'Media Diferencia Puntaje Partidas ganadas': 'Media Diferencia Puntaje
Partidas ganadasL',
   'Media Diferencia Puntaje perdidas': 'Media Diferencia Puntaje
perdidasL',
   'Proporcion_ganadas': 'Proporcion_ganadasL',
   'Promedio diferencia': 'Promedio diferenciaL',
}).drop(columns='TeamID', axis=1)
```

A su vez, en pro de manejar todas las variables en una misma escala, se procedió a generar una normalización de las bases de datos, permitiendo generar un modelamiento efectivo, sin que la escalabilidad pueda afectar.

```
def minmax(caracteristica, df_train, df_val, df_test=None):
   min_ = df_train[caracteristica].min()
   max_ = df_train[caracteristica].max()

   df_train[caracteristica] = (df_train[caracteristica] - min_) / (max_ -
min_)
   df_val[caracteristica] = (df_val[caracteristica] - min_) / (max_ - min_)

   if df_test is not None:
       df_test[caracteristica] = (df_test[caracteristica] - min_) / (max_ -
min_)

   return df_train, df_val, df_test
```

Como ejercicio, se buscará generar una partición de los datos y generarla a través de las temporadas para tener en cuenta la temporalidad, debido a que es importante para los torneos deportivos analizar el rendimiento temporada a temporada, tal y como se expresó en el apartado 3.2, la partición de los datos se debe generar de tal manera que yo valide una temporada con los datos anteriores a ella.

De manera paralela se implementa la metodología de validación cruzada kfold, la cual permite generar subdivisiones en los datos, generando pliegues que se entrenan y validan. Permitiendo obtener un modelo más robusto.

Las bases de datos implementadas, se encontraban “limpias” y no contaban con errores de digitación o datos que pudieran interferir en el desarrollo del problema, en algunos casos, se encontraron outliers sin embargo, estos son naturales en los juegos deportivos cuando se presentan diferencias físicas amplias, por ejemplo, se genera la expulsión de uno o más jugadores, los jugadores titulares no juegan y se decide jugar con la suplencia, entre otras.

4.3.Modelos

En función de la naturaleza del problema, se decide por implementar un modelo de regresión.

"La regresión lineal es un método estadístico que trata de modelar la relación entre una variable continua y una o más variables independientes mediante el ajuste de una ecuación lineal. Tres de las limitaciones que aparecen en la práctica al tratar de emplear este tipo de modelos (ajustados por mínimos cuadrados ordinarios) son:

- *Se ven perjudicados por la incorporación de predictores correlacionados.*
- *No realizan selección de predictores, todos los predictores se incorporan en el modelo aunque no aporten información relevante. Esto suele complicar la interpretación del modelo y reducir su capacidad predictiva.*
- *No pueden ajustarse cuando el número de predictores es superior al número de observaciones.*

Una forma de atenuar el impacto de estos problemas es utilizar estrategias de regularización como ridge, Lasso o Elastic Net, que fuerzan a que los coeficientes del modelo tiendan a cero, minimizando así el riesgo de overfitting, reduciendo varianza, atenuado el efecto de la correlación entre predictores y reduciendo la influencia en el modelo de los predictores menos relevantes." (Regularización Ridge, Lasso y Elastic Net con Python y Scikitlearn, s. f.)

$$\mu_{y=}\beta_0 + \beta_1 X1 + \beta_2 X2 + \cdots + \beta_p Xp$$

El Elastic Net es un método de regresión que utiliza una mezcla de las penalizaciones L1 y L2 (Kyung et al., 2010, Li y Lin, 2010). La función de coste es equivalente a:

$$RSSelastic\ Net = \sum_{i=1}^{n} \big(yi - f(xi)\big)^2 + \alpha(\lambda \sum_{j=i}^{p} \beta^2 + (1-\lambda) \sum_{j=1}^{p} |\beta j|$$

El parámetro λ regula el peso dado a la regularización impuesta por Ridge y por Lasso. Desde este punto de vista Elastic Net es un superconjunto de ambos modelos." (Elastic Net | Interactive Chaos, s. f.)

Es importante contemplar la metodología predictiva de regresión logística, toda vez que, nuestro interés primordial está en identificar si un equipo le puede ganar a otro, y con esta metodología podríamos definir etiquetas de si o no gana y además la probabilidad. La regresión logística se deriva de una investigación de los autores Cornfield J, Gordon T, Smith WN (1961)

"La regresión logística es una técnica estadística para predecir variables categóricas mediante variables predictoras. Identifica factores que influyen en el resultado y es útil en análisis de datos.

Este tipo de análisis puede ayudar a predecir la probabilidad de que ocurra un eventoo de que se tome una decisión. Por ejemplo, es posible que desee conocer la probabilidad de que un visitante elija una oferta realizada en su sitio web, o no (variable dependiente). Su análisis puede observar las características conocidas de los visitantes, como los sitios de los que provienen, las visitas repetidas a su sitio, el comportamiento en su sitio (variables independientes). Los modelos de regresión logística ayudan a determinar una probabilidad de qué tipo de visitantes probablemente aceptarán la oferta, o no. Como resultado, puede tomar mejores decisiones sobre la promoción de su oferta o tomar decisiones sobre la oferta en sí." (¿Qué es la regresión logística? | IBM, s. f.)

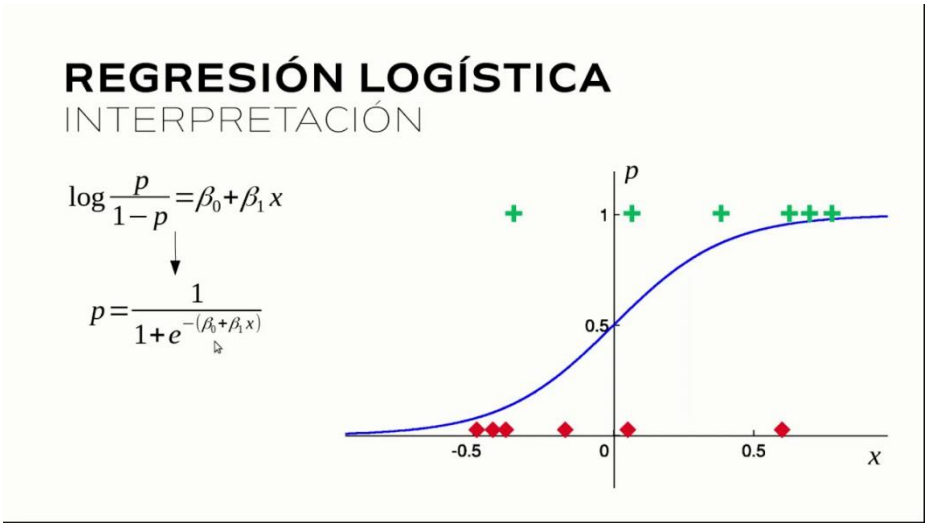

Ilustración 9. Regresión Logística

Extraída de https://www.youtube.com/watch?app=desktop&v=SeM4Rtoa4EU

Una vez utilizadas estas metodologías de machine learning, se implementará la metodología de validación cruzada en aras de buscar obtener mejores resultados.

Adicionalmente, como sugerencia realizada por la tutora de la monografía, se implementó la metodología de XGBoost.

XGBoost (potenciación del gradiente eXtreme) es una implementación de código abierto popular y eficiente del algoritmo de árboles aumentados de gradientes. La potenciación de gradientes es un algoritmo de aprendizaje supervisado que intenta predecir de forma apropiada una variable de destino mediante la combinación de un conjunto de estimaciones a partir de un conjunto de modelos más simples y más débiles. El algoritmo XGBoost funciona bien en competiciones de aprendizaje automático debido a su manejo robusto de una variedad de tipos de datos, relaciones, distribuciones y la variedad de hiperparámetros que puede ajustar. Puede usar XGBoost para problemas de regresión, de clasificación (binaria y multiclase) y de ranking." (Algoritmo XGBOOST - Amazon SageMaker, s. f.)

Las librerías de Python necesarias para la implementación de las metodologías anteriormente mencionadas, son: sklearn, xgboost, a continuación se evidencia el código implementado en Google Collab.

```
import sklearn
from sklearn.metrics import *
from sklearn.linear_model import *
from sklearn.model_selection import *
import xgboost
```

Ilustración 10. Librerías Python desarollo

4.4.Métricas

En primera instancia se analizará cómo se calculan las métricas desde la óptica de negocio, en la cual se buscará ver el beneficio de la implementación de modelos predictivos en las competencias deportivas en aras de materializar una campaña de promoción, pauta y marketing por parte de sponsors.

El cálculo del engagement posterior a la implementación del patrocinio derivado del modelo de machine learning, dependerá de la estrategia de pauta que se implemente por parte del sponsor, sí en función del modelo de machine learning se decide apoyar “X” o “Y” equipo, se podría establecer las siguientes formas de patrocinio:

1. El patrocinio se hace enfocado a redes sociales, será útil considerar:

- Cantidad de views de las historias antes y después del patrocinio.
- Número de seguidores antes y después del patrocinio.
- Número de interacciones antes y después del patrocinio.

2. El patrocinio se hace enfocado en la promoción de un producto para la venta, es útil considerar:

- Número de ventas de “X” producto antes y después del patrocinio.
- Interacciones sobre el producto en la página web antes y después del patrocinio.

3. El patrocinio se hace en función de incrementar las ventas en general de la empresa.

- Ingresos antes vs ingresos posteriores al patrocinio.
- Número de productos vendidos antes y después de la implementación.

Desde diversas perspectivas y métricas puede ser analizado el engagement del comportamiento posterior a la implementación de un patrocinio luego de generar una metodología de predicción de resultados para un evento deportivo.

Como fue relacionado anteriormente, a su vez se generará un análisis de las métricas de MAE y MSE, en primera instancia se busca que el MSE no supere el rango del 40% como se espera para las predicciones deportivas.

5. Metodología

5.1. Baseline

En primera instancia, se buscó generar la primera iteración únicamente corriendo un modelo de regresión en aras de determinar su comportamiento y tendencia, para posteriormente generar una comparación con los modelos que se han obtenido por algunos competidores de kaggle, y ver las variaciones que se podrían implementar en la materia. Para este caso, se utilizó un modelo de regresión lineal y utilizando la metodología de Kfold que es completamente indispensable.

```
Validating on season 2011
         -> MSE 0.203

Validating on season 2012
         -> MSE 0.207

Validating on season 2013
         -> MSE 0.213

Validating on season 2014
         -> MSE 0.213

Validating on season 2015
         -> MSE 0.204

Validating on season 2016
         -> MSE 0.215

Validating on season 2017
         -> MSE 0.205

Validating on season 2018
         -> MSE 0.214

Validating on season 2019
         -> MSE 0.201

Validating on season 2021
         -> MSE 0.216

Validating on season 2022
         -> MSE 0.218

 Local MSE is 0.205
```

Ilustración 11. Resultados primera iteración

El MSE local que se hace en función de la globalidad de las temporadas, es un MSE bueno, y la metodología de K-Fold propicia a tener un buen resultado, sin embargo, este ejercicio si bien se busca generar una predicción de probabilidades, se deberá buscar implementar a su vez una clasificación y buscar predecir si un equipo A puede ganarle a un equipo B.

En primera instancia, el primer modelo no podría clasificar si un equipo puede ganarle a otro, solo genera una cuantificación de la proporción de ganar, sin embargo, se deberá buscar que el MSE sea inferior implementando otras metodologías, además de visualizar las estrategias implementadas por los competidores de la plataforma y los beneficios que se obtienen.

5.2. Validación

En pro de realizar un ejercicio exhaustivo, se generó modelamiento mediante las metodologías de regresión lineal, elastic Net, Regresión logística, adicionalmente como ejercicio académico, se validó la implementación del algoritmo de XGBoost y su validación como modelo.

A su vez, se vislumbra una metodología implementada por los participantes la cual data de, particionar los datos conforme avanzan las temporadas deportivas, es decir, para validar los resultados de la temporada 2020, sería netamente beneficioso utilizar los datos anteriores a dicha temporada es decir, 2019 y anteriores, esto se puede ampliar en el apartado 3.2 del presente documento. En el ámbito deportivo los resultados y el nivel de los equipos varía conforme avanzan las temporadas, por ende generar una partición de los datos donde se contemple la temporalidad es correcto. De esta manera se tiene lo siguiente:

```
df_train= dftor[dftor['Season'] < season].reset_index(drop=True).copy()

df_val = dftor[dftor['Season'] == season].reset_index(drop=True).copy()
df_test = df_test_.copy()
```

Para este caso, el data test es suministrado por la competencia denominado “SampleSubmission2023.csv.

```
df_test = pd.read_csv("SampleSubmission2023.csv")
```

5.3. Iteraciones y evolución

En aras de generar un mayoy análisis y variar la metodología a implementar, se decidió utilizar la metodología de elastic-net utilizada por algunos competidores, para verificar la favorabilidad de la variación en la metodología de modelamiento.Así las cosas se obtuvo el siguiente resultado:

```
  if mode == "reg":
                model = ElasticNet(alpha=1, l1_ratio=0.5)
```

```
Validating on season 2011
        -> MSE 0.210

Validating on season 2012
        -> MSE 0.211

Validating on season 2013
        -> MSE 0.216

Validating on season 2014
        -> MSE 0.214

Validating on season 2015
        -> MSE 0.206

Validating on season 2016
        -> MSE 0.220

Validating on season 2017
        -> MSE 0.212

Validating on season 2018
        -> MSE 0.214

Validating on season 2019
        -> MSE 0.204

Validating on season 2021
        -> MSE 0.223

Validating on season 2022
        -> MSE 0.222

 Local MSE is 0.207
```

Ilustración 12. Primera iteración elastic Net

Como se puede percibir, la variabilidad del MSE entre la regresión lineal y la metodología de elastic Net no varía considerablemente.

Teniendo en cuenta la naturaleza del problema a resolver el cual busca determinar la probabilidad de "ganar" de un equipo frente a otro, se genera de manera paralela un modelo de machine learning implementando únicamente regresión logística, para ello se creó una variable denominada "Win A". La variable "Win A", es aquella variable que determina si se ganó un partido o no y será una variable binaria, es decir, entregará una respuesta de 0 ó de 1. Para esto, dicha variable fue calculada de la siguiente manera:

Equipo A puntaje = 70
Equipo B Puntaje =60

Se genera un condicional, que garantice que, si la diferencia del puntaje entre el equipo A y el B, es superior a 0, entonces "WinA"=1.

```
dftor['ScoreDiff'] = dftor['ScoreA'] - dftor['ScoreB']
dftor['WinA'] = (dftor['ScoreDiff'] > 0).astype(int)
```

Así las cosas, se implementa la metodología de regresión logística y este fue el cálculo del MSE temporada a temporada.

```
Validating on season 2009
         -> MSE 0.193

Validating on season 2010
         -> MSE 0.198

Validating on season 2011
         -> MSE 0.206

Validating on season 2012
         -> MSE 0.200

Validating on season 2013
         -> MSE 0.213

Validating on season 2014
         -> MSE 0.199

Validating on season 2015
         -> MSE 0.184

Validating on season 2016
         -> MSE 0.207

Validating on season 2017
         -> MSE 0.203

Validating on season 2018
         -> MSE 0.215

Validating on season 2019
         -> MSE 0.195

Validating on season 2021
         -> MSE 0.220

Validating on season 2022
         -> MSE 0.219

 Local MSE is 0.198
```

Ilustración 13. Regresión Logística

Como se percibió de manera preliminar, la data se concentra fuertemente posterior al año 2000, según se percibió en el siguiente diagrama de vela:

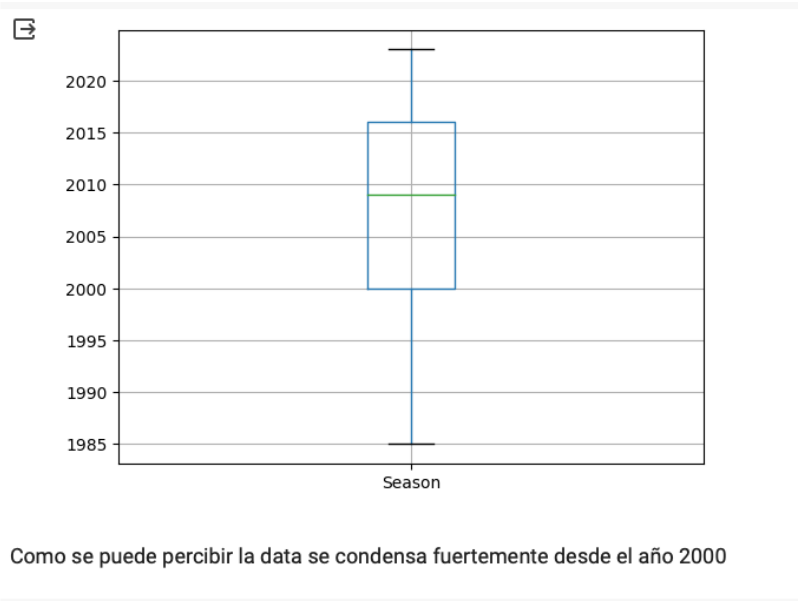

Ilustración 14. Distribución de la data

Seguido de ello se utilizan únicamente los valores superiores a 2016 como algunos participantes lo realizaron y estos fueron los resultados:

- Metodología Regresión lineal

```
Validating on season 2017
         -> MSE 0.205

Validating on season 2018
         -> MSE 0.215

Validating on season 2019
         -> MSE 0.201

Validating on season 2021
         -> MSE 0.216

Validating on season 2022
         -> MSE 0.219

 Local MSE is 0.211
```

Ilustración 15. Regresión lineal desde 2017

- Metodología ElasticNet

```
Validating on season 2017
         -> MSE 0.213

Validating on season 2018
         -> MSE 0.214

Validating on season 2019
         -> MSE 0.204

Validating on season 2021
         -> MSE 0.223

Validating on season 2022
         -> MSE 0.221

 Local MSE is 0.215
```

Ilustración 16. ElasticNet desde 2017

- Regresión logística.

```
Validating on season 2017
         -> MSE 0.205

Validating on season 2018
         -> MSE 0.212

Validating on season 2019
         -> MSE 0.196

Validating on season 2021
         -> MSE 0.218

Validating on season 2022
         -> MSE 0.217

 Local MSE is 0.209
```

Ilustración 17. Regresión Logística desde 2017

A su vez, como ejercicio académico y a solicitud de la asesora de la monografía, fue implementada la metodología de XGBoost.

- XGBoost

Se implementa la metodología de XGBoost

```
else:
    model = model = XGBClassifier()
```

Y se obtiene el siguiente resultado:

```
Validating on season 2010
         -> MSE 0.246

Validating on season 2011
         -> MSE 0.233

Validating on season 2012
         -> MSE 0.236

Validating on season 2013
         -> MSE 0.262

Validating on season 2014
         -> MSE 0.204

Validating on season 2015
         -> MSE 0.210

Validating on season 2016
         -> MSE 0.211

Validating on season 2017
         -> MSE 0.229

Validating on season 2018
         -> MSE 0.270

Validating on season 2019
         -> MSE 0.229

Validating on season 2021
         -> MSE 0.260

Validating on season 2022
         -> MSE 0.264

 Local MSE is 0.241
```

Ilustración 18. XGBoost como alternativa

Como se puede percibir, el XGBoost presenta un MSE bueno dado que es inferior a 0,4.

Luego de implementar diferentes tipologías de Machine learning, se visualiza que, el mejor modelo se obtiene con regresión logística utilizando kfold e incluyendo todas las temporadas, y es donde se encuentra el MSE global inferior.

```
Validating on season 2011
         -> MSE 0.206

Validating on season 2012
         -> MSE 0.200

Validating on season 2013
         -> MSE 0.213

Validating on season 2014
         -> MSE 0.199

Validating on season 2015
         -> MSE 0.184

Validating on season 2016
         -> MSE 0.207

Validating on season 2017
         -> MSE 0.203

Validating on season 2018
         -> MSE 0.215

Validating on season 2019
         -> MSE 0.195

Validating on season 2021
         -> MSE 0.220

Validating on season 2022
         -> MSE 0.219

 Local MSE is 0.198
```

Ilustración 19. MSE Regresión Logística

De esta manera, se predicen las probabilidades de los partidos que busca el torneo de kaggle predecir:

```
final_sub.head()
```

	ID	pred
0	2023_1101_1102	0.408234
1	2023_1101_1103	0.266898
2	2023_1101_1104	0.138591
3	2023_1101_1105	0.480422
4	2023_1101_1106	0.655318

Ilustración 20. Resultados.

5.4 Herramientas

Las herramientas implementadas para el desarrollo del ejercicio fueron la implementación de metodologías de machine learning: Regresión lineal, Elastic Net, Regresión logística y XGBoost. La teoría necesaria fue consultada en las notas de clase de la especialización en analítica y ciencia de datos de la Universidad de Antioquia, además la visualización de diversos videos de youtube para entender de una mejor manera la implementación de los diversos modelos y metodologías de machine learning. A su vez, se analizó los códigos de diferentes competidores de la plataforma kaggle los cuales son abiertos al público, entre ellos los códigos de los equipos ganadores.

A su vez, se consultó como utilizan los equipos deportivos herramientas de analítica de datos, para la toma de decisiones objetivas y se analizó como desde el marketing estas medidas repercuten directamente en el core de algunos negocios.

6. Resultados y discusión

Se generó el desarrollo de diferentes modelos de machine learning para la predicción de la probabilidad de que un equipo pueda ganarle a otro, las metodologías implementadas fueron, regresión lineal, elastic net, regresión logística y por último XGBoost. Sin embargo, como se observó anteriormente el mejor modelo se obtiene al implementar regresión logística utilizando todas las temporadas, en donde se obtiene un MSE del 0.198.

De esta manera, se ejecuta el modelo de regresión logística incluyendo todas las temporadas, y en función de la data suministrada por la plataforma kaggle, se valida uno a uno de los encuentros deportivos que se relacionan. A continuación se relaciona los datos suministrados por la plataforma kaggle.com, en donde todas las predicciones son del 50%, es decir no existe una diferencia entre un equipo y otro.

Tabla 7. Enfrentamientos para validar, plataforma kaggle.com.

	ID	Pred
0	2023_1101_1102	0.5
1	2023_1101_1103	0.5
2	2023_1101_1104	0.5
3	2023_1101_1105	0.5
4	2023_1101_1106	0.5

Una vez implementado el modelo de machine learning se obtienen los siguientes resultados:

Tabla 8. Modelado enfrentamientos Kaggle.com.

	ID	pred
0	2023_1101_1102	0.377654
1	2023_1101_1103	0.206033
2	2023_1101_1104	0.080283
3	2023_1101_1105	0.475150
4	2023_1101_1106	0.705072
5	2023_1101_1107	0.701167

De esta manera, se puede evidenciar claramente como al enfrentarse un equipo A y un equipo B, el modelo arrojará la probabilidad de que el equipo A gane. A manera de ejemplificación, se puede percibir que, para el primer enfrentamiento suministrado por la plataforma kaggle.com, la probabilidad de ganar del equipo “1101” al “1102” es del 37,8%.

Cuando se inicia con el ejercicio, todos los enfrentamientos se encuentran con una probabilidad del 50%, una vez que se corre el modelo, la probabilidad varía de acuerdo al desempeño deportivo de cada uno de los equipos enfrentados, por lo cual, a continuación, se relaciona una recopilación de las probabilidades obtenidas.

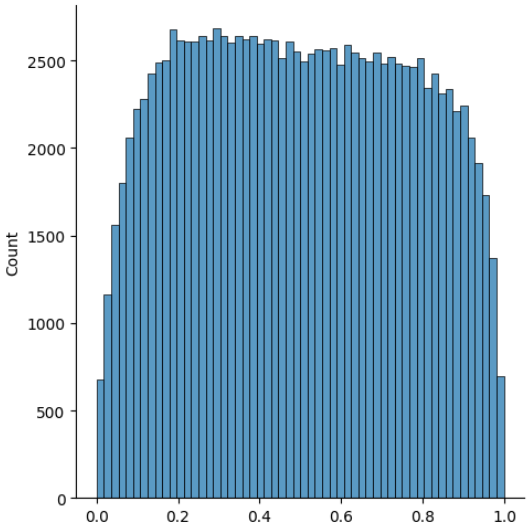

Ilustración 21. Distribución de las probabilidades.

Tal y como se percibe en la ilustración anterior, el modelo permitió obtener las diferentes probabilidades de victoria de un equipo a otro, con un error inferior al 20%, donde antes de generar la aplicación del modelo todos los valores serían del 50%, es decir los 130683 datos de testeo de la plataforma kaggle.com tienen un valor de 0.5, mientras que al aplicar el modelo de machine learning, las predicciones se distribuyen conforme a la ilustración anterior.

Desde la óptica del marketing, tener un modelo que permita obtener con este nivel de precisión el resultado de un enfrentamiento entre dos equipos, permite a un sponsor determinar

cuán beneficioso es patrocinar a un equipo. A manera de ejemplificación, una empresa podría negociar el patrocinio de un equipo inclusive únicamente por un partido o una serie de los mismos, y no toda la serie, toda vez que, promocionar por toda la temporada puede ser costoso.

Es de esa manera en que, desde el core del negocio un sponsor puede determinar qué tan beneficio o no puede llegar a ser el patrocinio de toda la temporada de un equipo, o a su vez, cuales son los partidos más estratégicos para realizar alianzas con un equipo u otro equipo. Las campañas de marketing, buscan poder generar un amplio engagement de la marca con una cantidad limitada de recursos. Es así como, con este modelo de machine learning a la hora de generar un patrocinio por parte un sponsor, se delimitarán cuales equipos son los más óptimos para patrocinar.

De manera paralela, los equipos deportivos al tener un modelo de machine learning de estas características que les permite predecir cuando tienen una mayor probabilidad de ganar, pueden tomar decisiones deportivas que beneficien al equipo en su rendimiento deportivo y/o económico. En aras de ejemplificar, un club puede determinar cuáles son los partidos que va jugar durante una temporada, y teniendo en cuenta el modelo de machine learning de regresión logística analizar la probabilidad de ganar o perder con cada uno de sus rivales. De esta manera podría, en aquellos partidos donde cuente con una mayor probabilidad de ganar, dejar descansar a sus jugadores más prestigiosos para que se encuentren en excelentes condiciones físicas para los eventos en los que tienen menor probabilidad de ganar. De igual manera, determinar según la probabilidad de ganar, cómo distribuir los equipos cuando debe jugar de local y a la vez internacionalmente en la misma semana. Por otra parte, según la probabilidad de ganar como distribuir la alineación deportiva si ir a la defensiva u a la ofensiva. Otra estrategia, antes de iniciar la temporada, vislumbrar si requiere de refuerzos para enfrentarse a “X” o “Y” equipo.

Como se puede identificar, las casas de apuestas podrán tomar decisiones de una manera más certera y objetiva, y es así como, podrán distribuir los premios según los porcentajes que le entregue el modelo, es decir, si la probabilidad de que un equipo A le gane a un equipo B, es muy alta, la tasa de ganancias por apostar por el equipo A será inversamente proporcional, es decir la tasa será baja, aplica de manera viceversa. A manera de ejemplificar, en un enfrentamiento entre el equipo “1101” y “1106”, tal y como se percibe en las predicciones realizadas por medio del

modelo, existe un probabilidad del 70% en que gane el equipo 1101, por lo cual, la casa de apuestas debe pagar en mayor medida por el equipo contrario, mientras que, se debe pagar un menor monto por acertar en que el equipo “1101” gane.

Así las cosas, se puede vislumbrar como los modelos de machine learning implementados, pueden ser útiles desde diferentes ámbitos generando beneficios económicos, estratégicos, organizacionales, entre otros a diferentes tipologías de empresas.

6.1. Métricas

Las métricas obtenidas en la ejecución de los diferentes modelos fueron plasmadas a lo largo del desarrollo del presente documento, sin embargo se evidencia que las mejores métricas se obtienen al implementar regresión logística y fue de 0,198.

A continuación se relaciona una tabla resumen con el MSE global, implementando las diferentes metodologías y temporalidades.

Tabla 9. Resultados MSE

Metodología	MSE
Regresión Lineal todas las temporadas.	0.205
Regresión Lineal temporadas mayores al 2017	0.211
Elastic Net todas las temporadas	0.207
Elastic Net temporadas mayores al año 2017	0.215
Regresión Logística todas las temporadas.	0.198
Regresión Logística temporadas mayores al año 2017	0.209
XGBoost todas las temporadas.	0.241

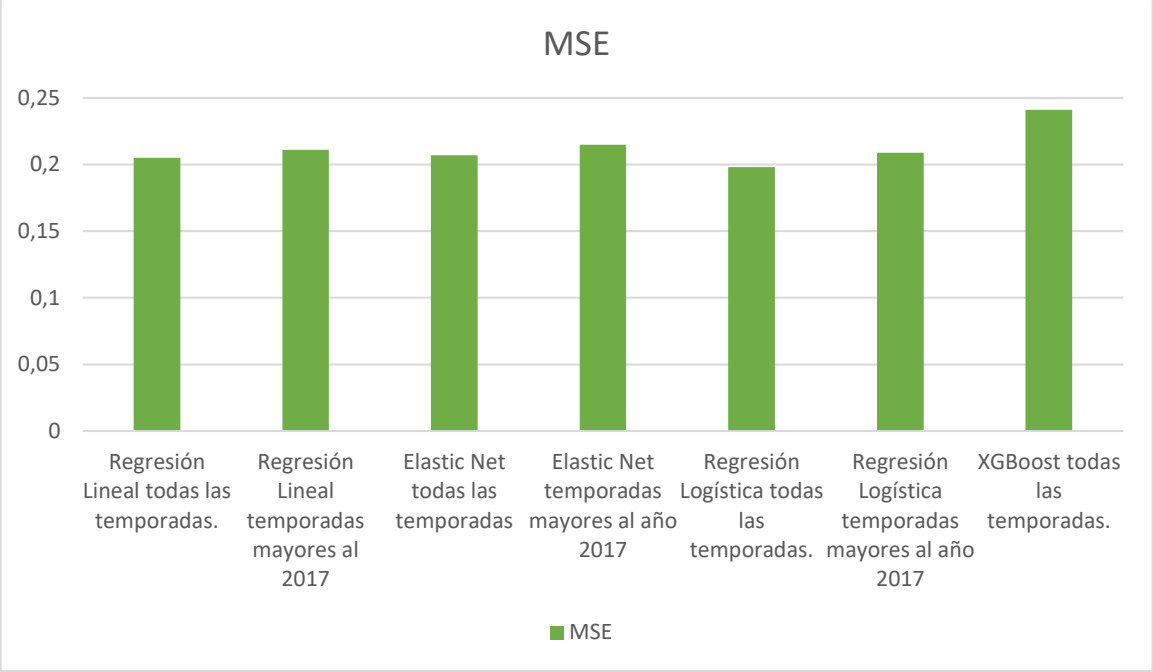

Ilustración 22. MSE para los diferentes modelos

Así las cosas, desde la óptica gerencial y de marketing una vez teniendo el modelo y conociendo que el error cuadrado medio es de 0,198 se pueden tomar decisiones frente al patrocinio a cierto equipo y de esta manera una vez implementada la estrategia de patrocinio, calcular una a una de las métricas relacionadas en el numeral 1.4. A su vez, se calculó el MAE para este modelo el cual fue de 0.392, sin embargo, la métrica utilizada fue el erro cuadrático medio, tal y como lo generan algunos de los competidores.

Si bien no existe una fluctuación considerablemente alta entre los modelos, para el ámbito predictivo deportivo, una variación por mínima que sea, es representativa toda vez que, denota mayor precisión de un modelado teniendo en cuenta la variabilidad deportiva que puede existir.

6.2.Evaluación cualitativa

Como se percibió anteriormente, los resultados que se obtuvieron son el consolidado de la variación entre las diferentes metodologías de machine learning implementadas, en donde el mejor resultado se encuentra asociado a la implementación de la regresión logística.

La regresión logística es comúnmente utilizada en este tipo de problemas de clasificación de resultados deportivos, y esta misma permite la predicción de la probabilidad que permite determinar la factibilidad de ganar de un equipo a otro. Cómo se relaciona anteriormente, dichos modelos son comúnmente utilizados por casas de apuestas, sponsors deportivos, equipos deportivos entre otros.

Ahora bien las métricas de ML obtenidas están asociadas al Error cuadrático Medio, el cual permite establecer la distancia que existe entre el dato predecido y el dato real, de esta manera, se obtiene un valor que expone que tan alejado se está de la predicción precisa de un resultado.

Para este caso, el MSE obtenido data de aproximadamente un 0,2 lo cual implica que, se puede predecir en una buena medida la probabilidad de que un equipo pueda ganarle a otro, un error cuadrático medio cercano a 0, podría indicar un sobreajuste, sin embargo, el resultado obtenido data de una buena predicción con un error habitual para eventos deportivos, considerándose una métrica de buena calidad.

El modelo que se generó permitirá a las empresas tomar decisiones objetivas que permitan aumentar su rendimiento deportivo, el engagement de un producto, o tomar decisiones frente a la implementación de recursos en campañas de expansión y mercadeo.

6.3. Consideraciones de producción

El modelo generado se realizó en función de la data suministrada por la plataforma kaggle, la cual es la recopilación de data para el torneo NCAA de estados Unidos, sin embargo, a manera de implementar esta metodología en un ambiente de producción, se podría generar un streaming que permita consolidar los resultados de diferentes torneos para diferentes deportes a lo largo del mundo, y poder predecir en tiempo real el resultado de eventos deportivos.

A su vez, se podría generar una validación constante conforme se desarrollen los eventos deportivos, es decir, validar y alimentar en tiempo real el modelo con cientos de datos, obtenidos

de diferentes eventos deportivos alrededor del mundo para diferentes torneos, tipología de deportes, categoría, temporalidad, entre otros.

Un modelo adaptado a esta gran cantidad de datos con servicios de streaming, permitiría en tiempo real predecir eventos deportivos que se evalúan constantemente y mejoran sus métricas día a día. La actualización permanente de la base de datos de entrenamiento garantiza que, se puedan evaluar eventos deportivos a lo largo del tiempo, dado que para el caso de estudio, se podría evaluar eventos deportivos del 2023, pero no sería ideal predecir eventos deportivos para el 2026, toda vez que, existiría una brecha de tiempo no analizada.

Este servicio de streaming sería de gran utilidad para casas de apuestas, inclusive para apuestas deportivas en tiempo real, además la data de los partidos que se juegan en vivo podrían alimentar el modelo permitiéndole ser más robusto.

Se deberán generar los cálculos de todas las métricas de negocio una vez se genere el patrocinio de un equipo fundamentado en el modelamiento de machine learning.

7. Conclusiones

Para la manipulación adecuada de los datos y la modelación de los mismos, se deben generar variables que permitan una mejor comprensión del trabajo a analizar. Para este caso particular, generar variables deportivas como la proporción de partidas ganadas, el promedio de la diferencia en proporción, el Winning Ratio, etc., generan un mejor entendimiento de la realidad deportiva de cada uno de los equipos.

El rendimiento deportivo de los equipos en cuanto a la cantidad de partidas ganadas juega un papel realmente importante en el marco de la predicción por parte de los modelos de predicción deportiva así como la media de las puntuaciones tanto en las partidas ganadas como en las partidas pérdidas, por ende generar una correcta recopilación de estos datos es beneficioso para la modelación.

Los retos en la plataforma kaggle.com, permiten generar una comparación entre el desarrollo que han tenido otros competidores, y de esta manera generar un modelo que pueda tener en cuenta distintas perspectivas en la manera en que se aborda la solución a un problema.

La metodología de regresión logística fue el modelo que mejor se ajustó para el desarrollo del ejercicio en el marco de la competencia de kaggle.com. Dicha metodología es comúnmente empleada para abordar predicciones deportivas dado que permite clasificar en victoria o derrota.

Es de gran utilidad contemplar modelos predictivos de machine learning para eventos deportivos por parte de sponsors, casas de apuestas y equipos deportivos, toda vez que, fundamentados en la estadística y la probabilidad de ocurrencia de un evento, pueden implementar y/o financiar ciertas campañas, actividades y tomar decisiones objetivas buscando mejorar la rentabilidad, engagement, visualización, de una empresa y/o producto.

El MSE obtenido mediante la regresión logística, fue de aproximadamente 0,2 lo cual representa la distancia cuadrada media entre un valor predecido y un valor real, de esta manera

mediante la implementación de este modelo, se pueden obtener valores predecidos confiables a la hora de estimar qué equipo ganará en un duelo.

Implementar metodologías de streaming pueden robustecer los modelos de predicción deportivas y a su vez incluir variables que pueden generarse en tiempo real, como lo son, tiros libres, tiros de esquina, posesión del balón, entre otras, podría mejorar sus métricas. Es de esta manera en que, las metodologías de predicción deportiva serán estrategias que deberán ser consolidadas y constantemente actualizadas conforme avancen diferentes encuentros deportivos.

8. Recomendaciones

En el marco de una implementación para una empresa que busque usufructuarse económicamente de un modelamiento en base a metodologías de machine learning para la predicción deportiva, sería completamente beneficioso implementar metodologías de streaming que actualice el modelamiento de una manera constante y que tengan en consideración eventos deportivos de toda índole, es decir, golf, tenis, futbol, entre otros. De esta manera se tendrá una valoración global de todos los deportes y se podrá tomar decisiones con recursos actualizados y para diferentes competencias deportivas.

Como recomendación a futuros competidores, se podría incluir el rendimiento deportivo de todos los jugadores alrededor de todas las temporadas, y de esta manera, generar modelos que tengan presente la relatividad que existe en un enfrentamiento deportivo dependiendo de cuales jugadores entran a debutar al campo de juego.

A futuro, se podría tener en cuenta los datos denominados “Seeds”, debido a que estos describen un territorio, por lo cual, sus condiciones climatológicas, geográficas, podrían tomar relevancia en el desarrollo de un partido. Lo anterior, se puede ver claramente ejemplificado en las implicaciones físicas de los jugadores cuando disputan un partido en otra latitud.

Se podría investigar cómo mejora el engagement de un producto o empresa derivado de la implementación de una metodología de machine learning para predecir eventos deportivos y en función de ellos, tomar la decisión del patrocinio a un equipo A o un equipo B.

Se podría analizar el beneficio de incluir diferentes tipologías de deportes y a su vez considerar variables adicionales y particulares de cada evento deportivo, como lo podría ser en el caso del golf, la velocidad del viento, la humedad, entre otros. De esta manera, se podría verificar particularmente por cada deporte cuales son las características realmente importantes, para la predicción en cada tipología de evento deportivo.

9. Referencias

March Machine Learning Mania 2023. (2023). kaggle.com. https://www.kaggle.com/competitions/march- machine-learning-mania-2023

Error absoluto promedio. (2021). IBM. https://www.ibm.com/docs/es/cloud-paks/cp-data/3.5.0?topic=overview-mean-absolute-error

March Machine Learning Mania 2023 Notebooks Code. (2023). Kaggle Mania 2023 Code. https://www.kaggle.com/competitions/march-machine-learning-mania-2023/code

Wikipedia contributors. (2023, 21 agosto). Winning percentage. Wikipedia. https://en.wikipedia.org/wiki/Winning_percentage

Regularización Ridge, Lasso y Elastic Net con Python y Scikitlearn. (s. f.). https://cienciadedatos.net/documentos/py14-ridge-lasso-elastic-net-python

Elastic Net | Interactive Chaos. (s. f.). https://interactivechaos.com/es/manual/tutorial-de-machine-learning/elastic-net

¿Qué es la regresión logística? | IBM. (s. f.). https://www.ibm.com/mx-es/topics/logistic-regression

Calderón Perez, L (2017). Predicción de resultados deportivos con técnicas de Machine Learning aplicado al fútbol. Universidad Carlos lll de Madrid.

Jeff Sonas, Maggie, Will Cukierski. (2023). March Machine Learning Mania 2023. Kaggle. https://kaggle.com/competitions/march-machine-learning-mania-2023

March Machine Learning Mania 2023. (2023). kaggle.com. https://www.kaggle.com/competitions/march-machine-learning-mania-2023/leaderboard

Tonatiuh, A. (2020, 18 agosto). ¿Cómo calcular el «engagement» en redes sociales? - Indat. Indat. https://indat.mx/2019/07/11/como-calcular-el-engagement-en-redes-sociales/

Algoritmo XGBOOST - Amazon SageMaker. (s. f.). https://docs.aws.amazon.com/es_es/sagemaker/latest/dg/xgboost.html

10. Anexos

Github: https://github.com/FelipeRam22/Monografiafinal