Universidad de Alcalá

Master Data Science Universidad de Alcala (2022-2023)

Trabajo Fin de Master

DATA SCIENCE EN EL GOLF

Enrique Alderete Heras

Madrid 2023

Índice

Trabajo realizado por **Enrique Alderete Heras**
Contacto: Enriquealdereteheras@gmail.com; Tel.: +34678947198
Todos los datos obtenidos y el código de este estudio está disponible Open Source en el repositorio:
https://github.com/Kikote71/TFM-GOLF-STADISTICS-

1. Introducción

Este trabajo de investigación se centra en el análisis del rendimiento de jugadores de golf mediante la aplicación de técnicas de Data Science. Utilizando un conjunto de datos recopilados a través de web scraping desde la página oficial del PGA Tour, que abarca más de 18,000 registros estadísticos, se busca revelar patrones y tendencias significativas en el desempeño de los jugadores. En este contexto, se contextualiza la importancia de las estadísticas en la evaluación del rendimiento y se resalta el propósito del estudio: explorar las posibilidades de mejora a través del análisis de datos.

Los métodos aplicados se basan en análisis exploratorio de datos y técnicas de visualización, así como en la implementación de modelos de aprendizaje automático para predecir el rendimiento futuro. Los resultados principales abarcan la identificación de correlaciones entre variables clave y el rendimiento de los jugadores, la identificación de patrones en los datos históricos y la creación de modelos predictivos con un nivel de precisión notable.

Las conclusiones sugieren que el análisis de datos a través de herramientas de Data Science puede proporcionar una comprensión más profunda del rendimiento de los jugadores de golf, lo que a su vez puede respaldar la toma de decisiones estratégicas para jugadores, entrenadores y seguidores. Este enfoque podría tener aplicaciones valiosas en la optimización del entrenamiento, la elaboración de estrategias y la mejora general del juego.

El golf, un deporte que combina habilidad, precisión y estrategia, ha evolucionado considerablemente en los últimos años, y junto con esta evolución ha surgido la importancia de la recopilación y análisis de datos para comprender y mejorar el rendimiento de los jugadores. La creciente disponibilidad de datos detallados proporciona una oportunidad única para explorar las variables que influyen en el éxito de los golfistas en el escenario competitivo. En este contexto, la aplicación de técnicas de Data Science ha emergido como una herramienta poderosa para desentrañar patrones ocultos y tendencias dentro de los datos del golf.

A medida que los deportes en general evolucionan hacia un enfoque más orientado a los datos, el golf no es una excepción. El análisis de datos puede desbloquear una comprensión más profunda de cómo se desarrolla un juego, cuáles son las fortalezas individuales y las áreas que requieren atención. Esta investigación contribuye a la creciente literatura que demuestra cómo la aplicación inteligente de Data Science puede transformar la forma en que se aborda el rendimiento deportivo. En las secciones siguientes, se describen los métodos utilizados, los resultados obtenidos y las conclusiones que emergen de este análisis de datos en el contexto del golf de alto nivel.

2. Materiales y Métodos: Recopilación y Análisis de Datos Estadísticos del PGA Tour

En este apartado, se describe en detalle la metodología utilizada para la recopilación y el análisis de los datos estadísticos del PGA Tour a lo largo de los últimos cuatro años. Estos datos fueron obtenidos mediante web scraping desde la página oficial del PGA Tour y constituyen la base de nuestro estudio.

Recopilación de Datos:

Se empleó un enfoque de web scraping para extraer información estadística relevante de la sección de estadísticas del PGA Tour en su página oficial. Los datos recopilados abarcan un período de cuatro años y comprenden una amplia gama de variables que se detallaran más adelante, que van desde el promedio de golpes por ronda hasta las estadísticas de precisión en el juego largo y corto. El proceso de web scraping fue manual ya que no pudimos automatizar la toma de datos a través de herramientas específicas.

Análisis Exploratorio de Datos:

Una vez recopilados los datos, se llevó a cabo un análisis exploratorio para comprender la naturaleza y la distribución de las variables. Se generaron gráficos y visualizaciones para identificar tendencias y patrones preliminares en los datos. Además, se realizó una limpieza y transformación de los datos para garantizar la coherencia y la calidad de la información recopilada.

Análisis Avanzado y Modelización:

Se aplicaron técnicas de análisis de datos avanzadas, incluyendo análisis de correlación, regresión y agrupación, para descubrir relaciones subyacentes entre las variables y el rendimiento de los jugadores. Se implementaron modelos de aprendizaje automático para predecir el rendimiento futuro en función de las estadísticas pasadas. Se utilizaron bibliotecas de Python como pandas, numpy y scikit-learn para llevar a cabo estos análisis y modelados.

Disponibilidad de Datos y Código:

Todos los datos recopilados, así como el código utilizado para el análisis, estarán disponibles para su acceso público y revisión a través de https://github.com/Kikote71. Los datos recopilados se alojarán en este repositorio, y el código estará disponible para garantizar la transparencia y la reproducibilidad de los resultados.

En conjunto, este enfoque de recopilación y análisis de datos ofrece una perspectiva valiosa sobre el rendimiento de los jugadores de golf a lo largo de los últimos cuatro años en el PGA Tour. Las técnicas del análisis exploratorio y modelado de datos proporcionan una base sólida para la obtención de información relevante y la identificación de patrones clave que pueden influir en el desempeño de los jugadores.

3. Estadísticas

3.1. Estadísticas Utilizadas

3.1.1. Los Identificadores:

Las variables "**PLAYER_ID**" y "**PLAYER_NAME**" son identificadores únicos y nombres de los jugadores, respectivamente. Aunque no son estadísticas en sí mismas, desempeñan un papel crucial en el análisis y la interpretación de los datos

Estos identificadores nos son muy útiles:

Agrupación y Segmentación: Podemos agrupar y segmentar los datos en función de los jugadores utilizando sus identificadores. Esto nos permitirá analizar el rendimiento de jugadores específicos y comparar sus estadísticas con otros jugadores.

Visualización Personalizada: Podemos crear visualizaciones personalizadas que muestren el rendimiento de jugadores individuales a lo largo del tiempo. Esto nos sirve para elaborar que representen cómo las estadísticas cambian para un jugador en particular en diferentes torneos.

Análisis de Jugadores Destacados: Podemos identificar jugadores destacados al examinar sus nombres y estadísticas.

Comparación de Jugadores: Utilizando los nombres y las estadísticas, podemos comparar el rendimiento de varios jugadores en diferentes áreas del juego.

Relación entre Identificador y Rendimiento: Podemos investigar si existe alguna correlación entre los identificadores únicos ("PLAYER_ID") y el rendimiento de los jugadores. Esto podría ayudarte a comprender si ciertos grupos de jugadores tienen un rendimiento generalmente mejor o peor en comparación con otros.

Validación y Control de Calidad: Los identificadores únicos ("PLAYER_ID") pueden ser útiles para verificar la integridad de tus datos. Siempre es importante asegurarse de que los datos estén correctos y que no haya duplicados o errores.

En resumen, aunque "PLAYER_ID" y "PLAYER_NAME" no son estadísticas en sí mismas, son esenciales para contextualizar y analizar las estadísticas de los jugadores de manera significativa. Pueden ayudarte a individualizar el análisis y a comprender mejor cómo cada jugador contribuye al conjunto de datos y al rendimiento general en el golf.

3.1.2 Las Estadísticas Utilizadas:

TOTAL_ROUNDS: El número total de rondas de golf jugadas por el jugador en el período considerado. Indica la actividad y la participación del jugador en torneos.

AVG_FAIRWAY: El promedio de golpes que el jugador realiza desde el tee y aterriza en la calle (fairway) en un hoyo dado. Representa la precisión del jugador para mantener la bola en el fairway desde el tee de salida. Esta estadística se obtiene:

$$AVG_{FAIRWAY} = \frac{FAIRWAY\ HITS}{POSSIBLE\ FAIRWAYS}\ X\ 100$$

Cuanto mayor sea este valor representa un mejor rendimiento en alcanzar los fairways desde el tee

AVG_SCRAMBLING: El promedio de golpes que el jugador realiza desde fuera del green y logra terminar el hoyo con un putt. Muestra la habilidad del jugador para recuperarse de situaciones difíciles. Esta estadística se obtiene:

$$AVG_{scrambling} = \frac{PAR\ OR\ BETTER}{MISSED\ GIR}\ X\ 100$$

La fórmula divide el número de veces que el jugador terminó el hoyo con par o mejor después de haber estado fuera del green (PAR OR BETTER) entre el número de veces que el jugador no logró alcanzar el green en regulación (MISSED GIR). Este estadístico es mejor cuanto mayor sea su valor

AVG_TOTAL_FOR_ROUND: El promedio total de golpes realizados por el jugador en cada ronda de golf. Es una medida del rendimiento general del jugador en términos de golpes por ronda.

No es necesario un cálculo especial para esta estadística, ya que "AVG Total for Round" es simplemente el promedio de los golpes realizados por un jugador en cada ronda. En este estadístico, un valor menor supone un mejor rendimiento en cada ronda, por lo que convertiremos en negativo este estadístico en el data set

AVG_DRIVER_DISTANCE: La distancia promedio alcanzada por el jugador con su golpe de salida (drive) desde el tee. Muestra la potencia y la capacidad del jugador para lograr distancias más largas.

No es necesario realizar un cálculo especial para esta estadística, ya que "AVG Driver Distance" es simplemente el promedio de las distancias alcanzadas por un jugador en sus golpes de salida. Un valor alto supone un mejor rendimiento en la distancia alcanzada por el drive

AVG_SG_APPROACH: (Promedio de Golpes Ganados o Perdidos en Tiros de Aproximación) se refiere al promedio de golpes ganados o perdidos por un jugador en comparación con el promedio en tiros de aproximación al green. Esta estadística mide la habilidad del jugador en tiros de aproximación precisos que se ejecutan desde una distancia considerable al green.

$$AVG_{SG\_APPROACH} = \frac{TOTAL\ SG{:}\ APP}{MEASURED\ ROUNDS}\ X\ 100$$

El AVG_SG_Approach se calcula restando el promedio de los golpes ganados o perdidos en tiros de aproximación (TOTAL SG:APP) entre el jugador y el promedio del campo, y luego dividiendo ese valor por el número de rondas consideradas (MEASURED ROUNDS).

Este cálculo proporciona un valor que indica cuántos golpes por ronda el jugador gana o pierde en comparación con el promedio del campo en tiros de aproximación al green. Un valor positivo indica que el jugador gana golpes en tiros de aproximación, y un valor negativo indica que pierde golpes en promedio

AVG_GIR: (Promedio de Greens en Regulación) se refiere al promedio de greens en regulación alcanzados por un jugador durante una ronda de golf. Un "green en regulación" se alcanza cuando un jugador completa un hoyo en el número estándar de golpes o menos. Esta estadística mide la frecuencia con la que un jugador logra alcanzar el green en el número estándar de golpes.

$$AVG_{GIR} = \frac{GREEN\ HITS}{HOLES}$$

Este cálculo proporciona el porcentaje de veces que un jugador logra alcanzar el green en regulación en relación con el número total de hoyos jugados en una ronda de golf.Un valor alto supone un mejor rendimiento en el GIR

AVG_PUTT: (Promedio de Putts) se refiere al promedio de putts realizados por un jugador en cada green durante una ronda de golf. Esta estadística mide la habilidad del jugador para finalizar con éxito los hoyos con un número mínimo de putts.

$$AVG_{PUTT} = \frac{GIR\ PUTTS}{GREEN\ HITS}$$

El "AVG Putt" se calcula dividiendo el número total de putts realizados por el jugador (GIR PUTTS) entre el número total de greens en regulación alcanzados (GREENS HIT) y luego redondeando el resultado a dos decimales.

Un valor más bajo en "AVG Putt" indica que el jugador tiene la habilidad de finalizar los hoyos con un menor número de putts, lo que es favorable para lograr buenos resultados en el golf. Por ello, convertiremos en negativo este estadístico en el data set

AVG_SG_ATG: (Promedio de Golpes Ganados o Perdidos en Tiros al Green y Putts) se refiere al promedio de golpes ganados o perdidos por un jugador en comparación con el promedio del campo en tiros al green y putts. Esta estadística combina tanto los tiros de aproximación como las estadísticas de putt en un solo número, proporcionando una medida integral del rendimiento del jugador en estos aspectos del juego.

$$AVG_{SG\_ATG} = \frac{TOTAL\ SG{:}\ ARG}{MEASURED\ ROUNDS}$$

Este cálculo proporciona un valor que indica cuántos golpes por ronda el jugador gana o pierde en comparación con el promedio del campo en tiros al green y putts. Un valor positivo indica que el jugador gana golpes en estas áreas, mientras que un valor negativo indica que pierde golpes en promedio

AVG_SG_PUTTING: (Promedio de Golpes Ganados o Perdidos en Putts) se refiere al promedio de golpes ganados o perdidos por un jugador en comparación con el promedio del campo en putts. Esta estadística mide la destreza del jugador en el juego corto, específicamente en el aspecto de putts.

$$AVG_{SG\_PUTTING} = \frac{TOTAL\ SG{:}\ PUTTING}{MEASURED\ ROUNDS}$$

Este cálculo proporciona un valor que indica cuántos golpes por ronda el jugador gana o pierde en comparación con el promedio del campo en putts. Un valor positivo indica que el jugador gana golpes en putts, mientras que un valor negativo indica que pierde golpes en promedio.

AVG_SG_TOTAL: (Promedio de Golpes Ganados o Perdidos en Total) en el golf se refiere al promedio de golpes ganados o perdidos por un jugador en comparación con el promedio del campo en todos los aspectos del juego combinados. Esta estadística engloba el rendimiento global del jugador y considera tanto los tiros al green, los tiros desde fuera del green y los putts.

$$AVG_{SG\_TOTAL} = \frac{TOTAL\ SG{:}T}{MEASURED\ ROUNDS}$$

Este cálculo proporciona un valor que indica cuántos golpes por ronda el jugador gana o pierde en comparación con el promedio del campo en todos los aspectos del juego combinados. Un valor positivo indica que el jugador gana golpes en promedio, mientras que un valor negativo indica que pierde golpes en promedio.

AVG_SG_TTG: (Promedio de Golpes Ganados o Perdidos desde el Tee hasta el Green) en el golf se refiere al promedio de golpes ganados o perdidos por un jugador en comparación con el promedio del campo en todos los aspectos del juego desde el tee hasta el green. Esta estadística abarca los tiros desde el tee (drive), tiros de aproximación al green y tiros desde fuera del green.

$$AVG_{SG\_TTG} = \frac{(SG{:}OTT + SG{:}APR + SG{:}ARG) - AVG}{MEASURED\ ROUNDS}$$

Este cálculo proporciona un valor que indica cuántos golpes por ronda el jugador gana o pierde en comparación con el promedio del campo en todos los aspectos del juego desde el tee hasta el green. Un valor positivo indica que el jugador gana golpes en promedio en esta categoría, mientras que un valor negativo indica que pierde golpes en promedio

PAR_5_AVG: Esta estadística representa el promedio de golpes que el jugador realiza en los hoyos de par 5. Los hoyos de par 5 son los más largos en un campo de golf y generalmente se espera que los jugadores realicen más golpes en ellos. Un valor bajo en esta estadística indica que el jugador es eficaz en los hoyos largos y puede hacer menos golpes para completarlos. Por ello, convertiremos en negativo este estadístico en el data set

PAR_4_AVG: Esta estadística se refiere al promedio de golpes realizados por el jugador en los hoyos de par 4. Los hoyos de par 4 son de longitud moderada y presentan un equilibrio entre distancia y precisión. Un bajo valor en esta estadística sugiere que el jugador es efectivo en enfrentar los hoyos de longitud moderada. Por ello, convertiremos en negativo este estadístico en el data set

PAR_3_AVG: Representa el promedio de golpes realizados por el jugador en los hoyos de par 3. Los hoyos de par 3 son los más cortos en un campo y generalmente requieren precisión en los tiros de tee. Una baja cantidad de golpes en esta estadística indica que el jugador tiene buena precisión en tiros cortos. Por ello, convertiremos en negativo este estadístico en el data set

WINS: Indica el número de torneos ganados por el jugador durante el período considerado. Esta estadística refleja la capacidad del jugador para llevarse la victoria en eventos competitivos, lo que es un indicador clave de su éxito en el circuito.

TOP10: Representa la cantidad de veces que el jugador se ubicó en el top 10 en torneos durante el período considerado. Esta estadística muestra con cuánta frecuencia el jugador está entre los diez mejores competidores, lo que sugiere una consistencia en su desempeño.

YEAR: Esta estadística simplemente proporciona el año al que pertenecen los datos registrados. Es útil para contextualizar los resultados y comprender en qué período se aplican las estadísticas.

FEDEX_POINTS: Indica la cantidad de puntos FedEx ganados por el jugador, los cuales influyen en su posición en el ranking de la FedEx Cup. Los puntos FedEx son asignados en función del desempeño del jugador en diferentes torneos y pueden ser determinantes en las clasificaciones generales.

MONEY: Esta estadística representa la cantidad total de ganancias monetarias obtenidas por el jugador en los torneos. Refleja el éxito financiero del jugador y su capacidad para obtener premios en efectivo basados en su rendimiento en los eventos.

4 El Analisis del Dato:

En primer lugar, debemos detectar que variables estamos analizando:

Categóricas: PLAYER_ID (Identificador único del jugador), PLAYER_NAME (Nombre del jugador)

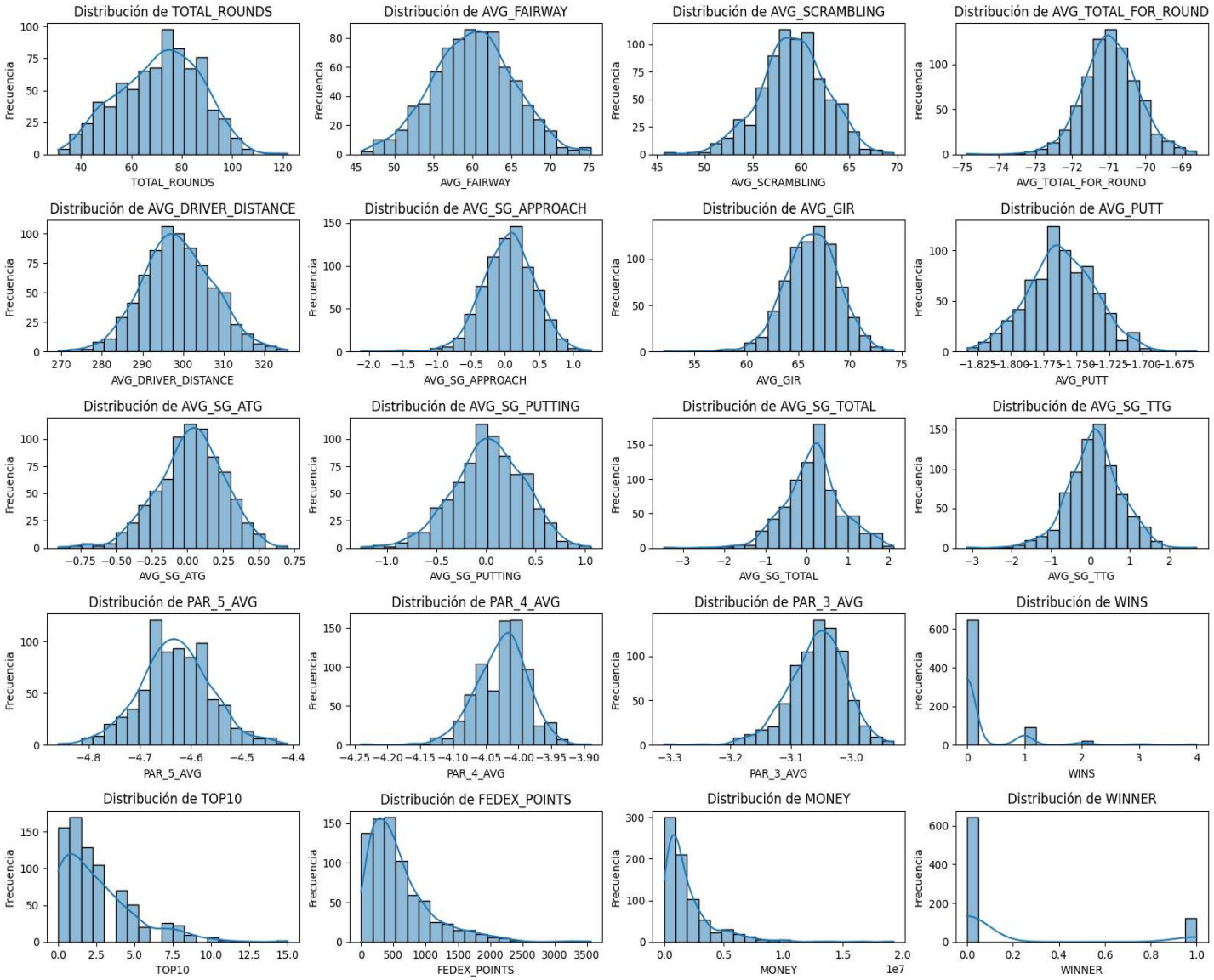
Numéricas Estadisticas: AVG_FAIRWAY (Promedio de golpes en el fairway),AVG_SCRAMBLING (Promedio de recuperación de situaciones difíciles),AVG_TOTAL_FOR_ROUND (Promedio total de golpes por ronda),AVG_DRIVER_DISTANCE (Promedio de distancia del golpe de salida),AVG_SG_APPROACH (Promedio de golpes ganados o perdidos en tiros de aproximación),AVG_GIR (Promedio de greens en regulación),AVG_PUTT (Promedio de putts realizados),AVG_SG_ATG (Promedio de golpes ganados o perdidos en tiros al green y putts),AVG_SG_PUTTING (Promedio de golpes ganados o perdidos en putts),AVG_SG_TOTAL (Promedio de golpes ganados o perdidos en total),AVG_SG_TTG (Promedio de golpes ganados o perdidos desde el tee hasta el green),PAR_5_AVG (Promedio de golpes en hoyos de par 5),PAR_4_AVG (Promedio de golpes en hoyos de par 4),PAR_3_AVG (Promedio de golpes en hoyos de par 3)

Numéricas (Métricas):TOTAL_ROUNDS (Número total de rondas de golf jugadas por el jugador),TOP10 (Número de veces en el top 10 en torneos)

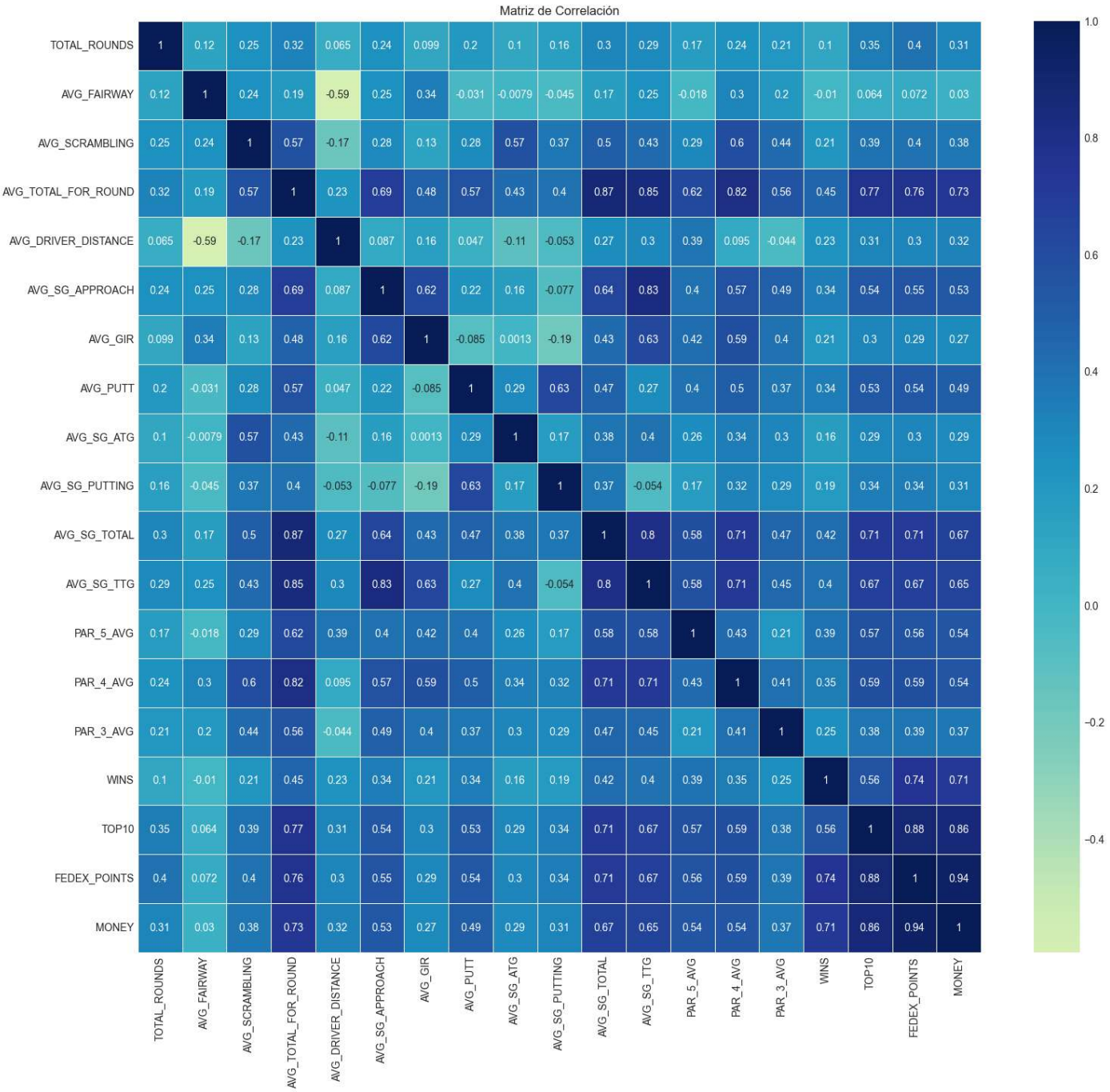
Categóricas (Variables Descriptivas): YEAR (Año al que pertenecen los datos)

Numéricas (Variables de Logro): WINS (Número de torneos ganados), FEDEX_POINTS (Cantidad de puntos FedEx ganados), MONEY (Cantidad total de ganancias monetarias)

Las que mas nos interesan analizar son las variables numéricas estadisticas, para su correcto análisis, sacamos las distribuciones y podemos comprobar como todas las distribuciones son “Normales” algo fundamental para el correcto análisis del dato



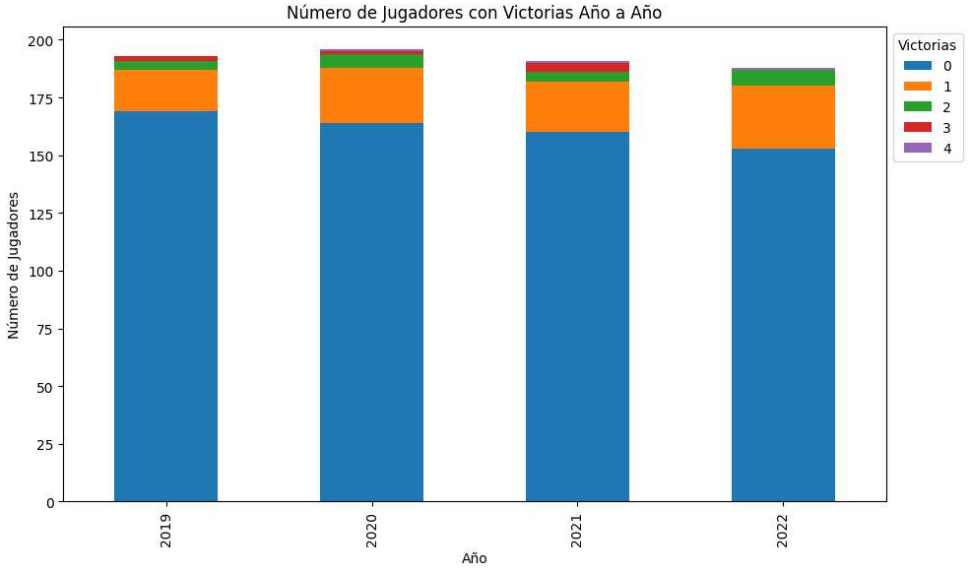
Una vez que hemos visto que los datos son buenos para ser analizados, vamos a obtener una matriz de correlacion para ver la correlacion existente entre las diferentes variables:



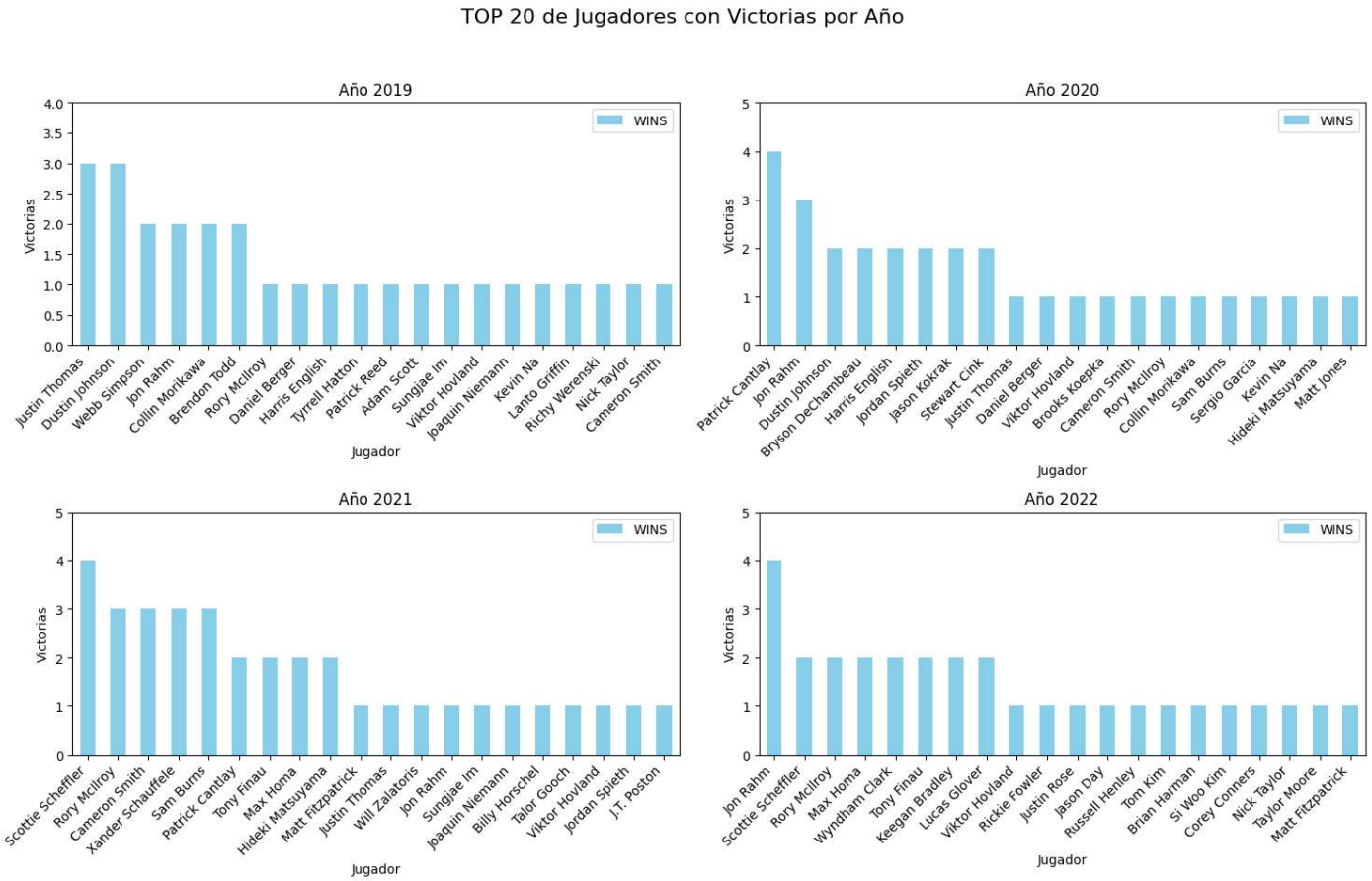
4.1 Las Victorias

Como en cualquier deporte, el factor determinante entre el existo y el fracaso es la victoria, en el golf, al ser un deporte individual en la mayoría de los formatos/torneos profesionales solo puede ganar un jugador. Semanalmente se celebran torneos del PGA donde participan unos 180 jugadores y donde solo gana uno

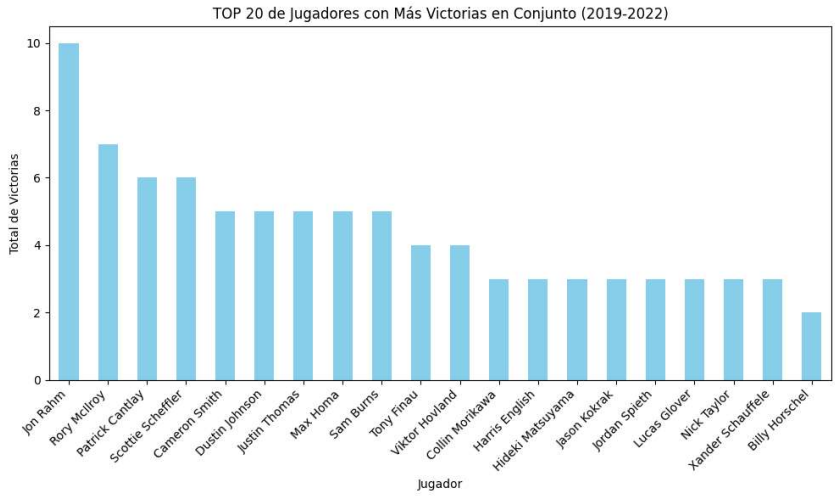
En nuestro análisis valoramos las victorias como lo que son, algo extraordinario una vez que se producen, pero también damos valor a la varibale TOP10, es decir, cuando un jugador queda dentro de los 10 primeros puestos en un torneo



En el grafico anterior es importante que destaquemos lo difícil que es ganar un torneo, y mucho mas jugadores que ganan hasta 4 torneos en un mismo año



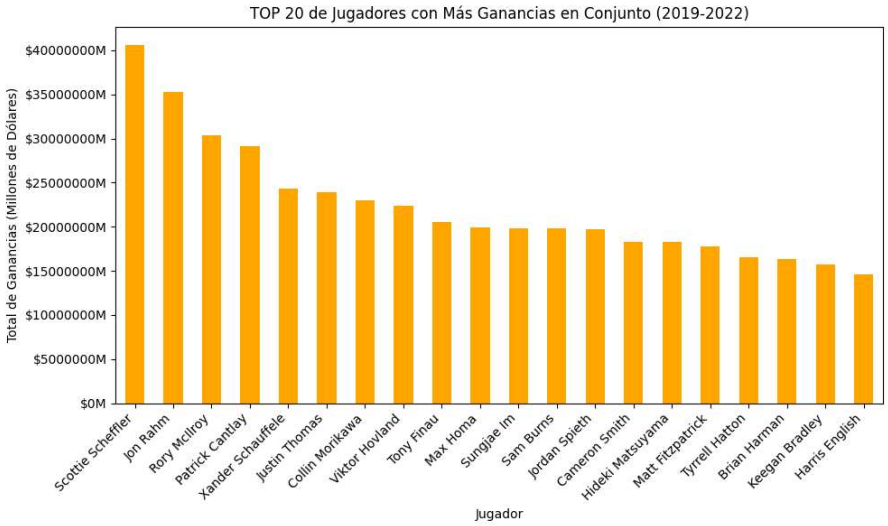
El Golf es un deporte tan difícil que es muy complicado que un jugador este en su “prime” temporada tras temporada, se puede ver cómo hay muy pocos jugadores que repitan año tras año victorias en el PGA TOUR



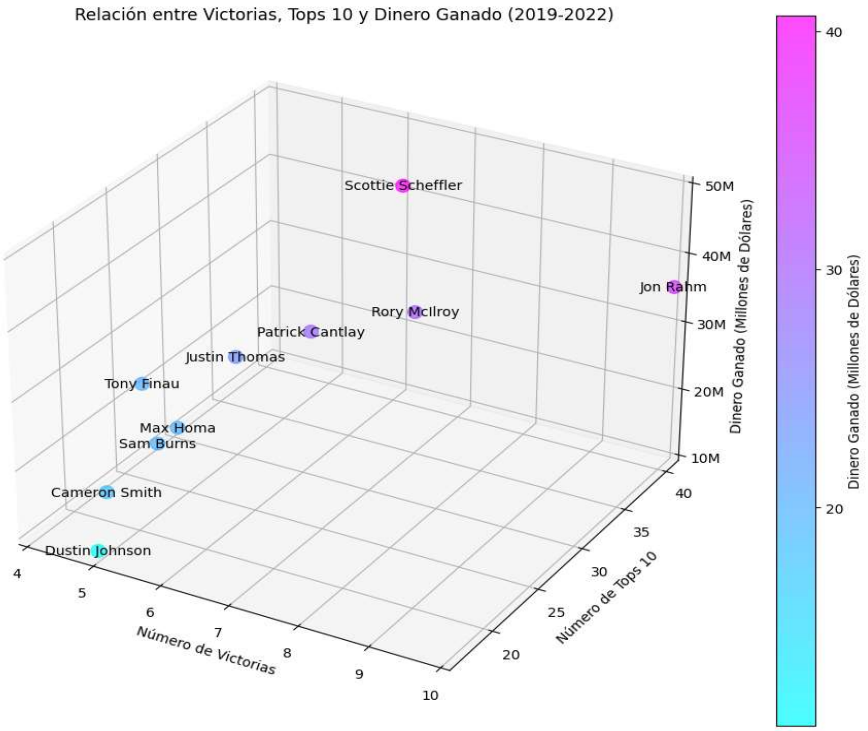
Es necesario destacar al jugador español Jon Rahm, sin duda un “Outlier” el cual ha ganado 10 torneos en los 4 últimos años en el tour

4.2 Ganancias Económicas

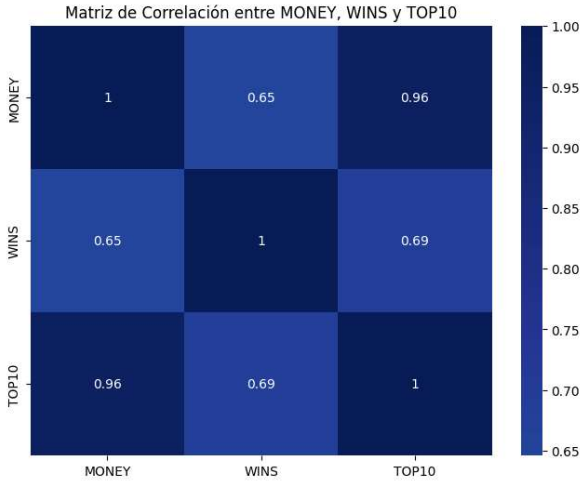
El Golf es uno de los deportes que mas premios reparte entre sus jugadores



Como se puede observar en el grafico anterior, no siempre el jugador que mas victorias tiene en el tour es el que más premios percibe, ya que quedar en el Top10 de un torneo supone percibir cantidades importantes en premios, vemos la relación existente entre el dinero ganado, las victorias y los Top10



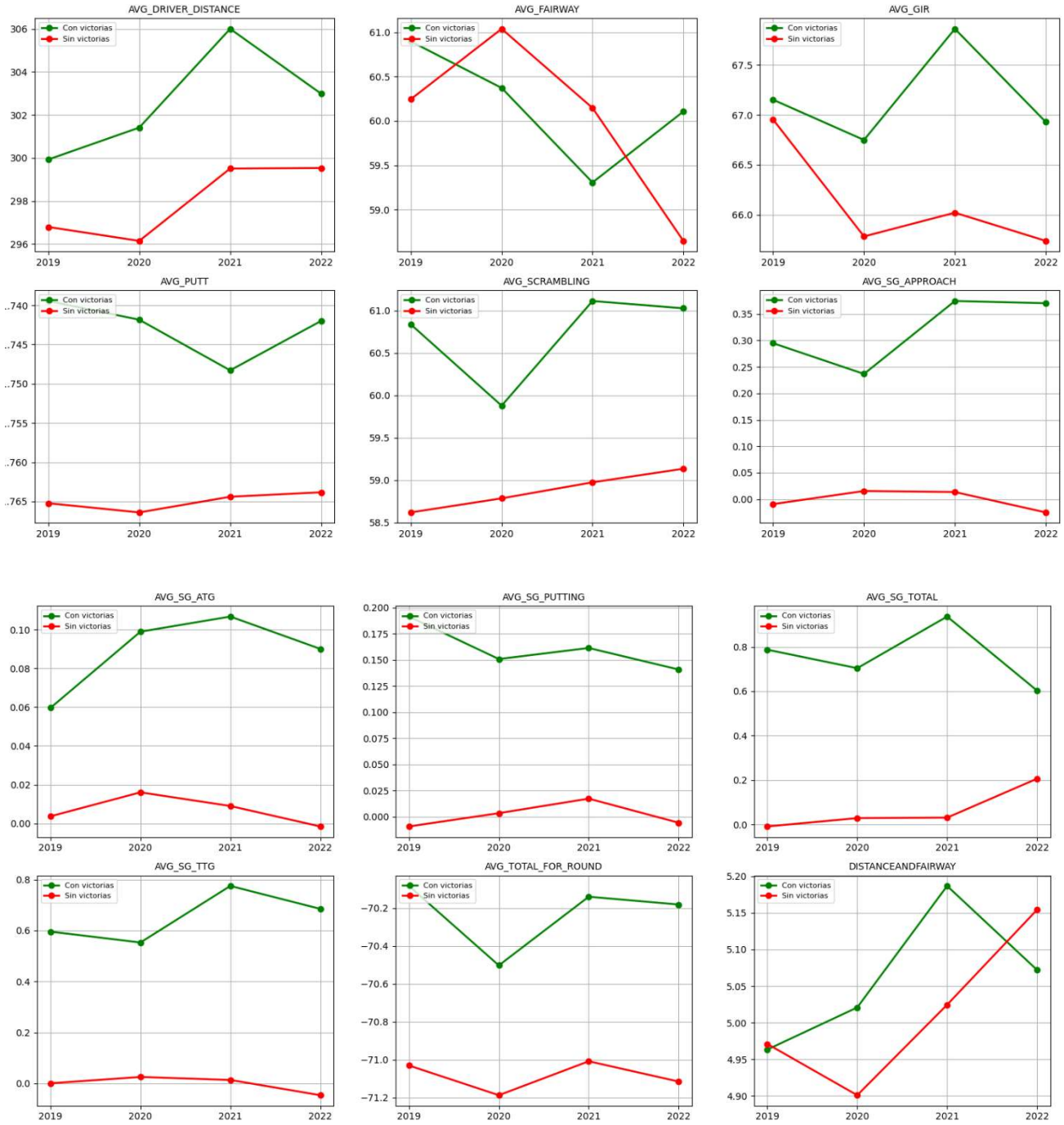
Veamos cuanto de importante es cada variable respecto a las victorias:

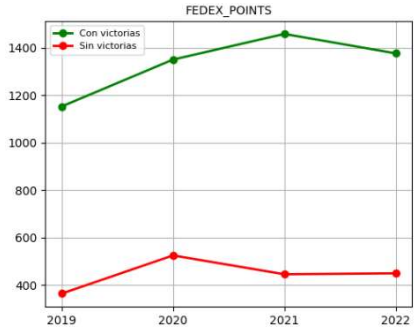


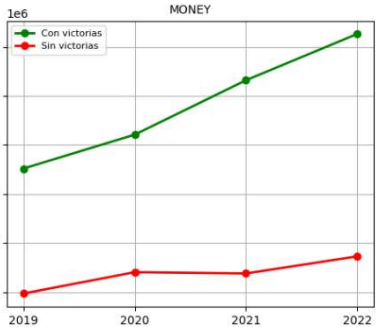
Como podemos ver en la matriz de correlación, las 2 variables Top10 y Wins están muy correlacionadas con la variable Money, pero esta mucho mas correlacionada la variable Top10 con el dinero que las victorias con el dinero, esto se debe a que es más factible alcanzar Top10 en un torneo que ganar el propio torneo y en consecuencia el obtener más dinero por alcanzar de manera recurrente ese Top10 que por tener victorias

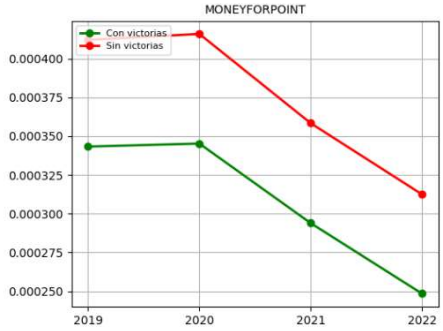
4.3 Jugadores con Victorias VS Jugadores sin Victorias

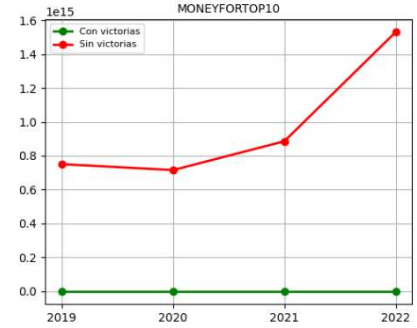
En este punto vamos a analizar los rendimientos de los jugadores en las variables estadísticas en función de los jugadores que cuentan en su haber con victorias y los jugadores que no

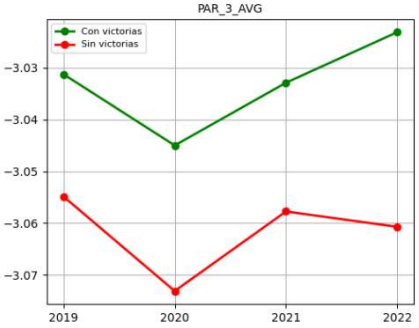


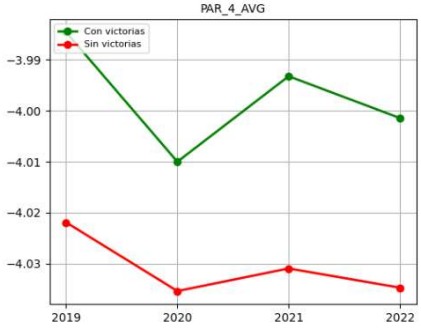

FEDEX_POINTS
Con victorias
Sin victorias
1400
1200
1000
800
600
400
2019
2020
2021
2022

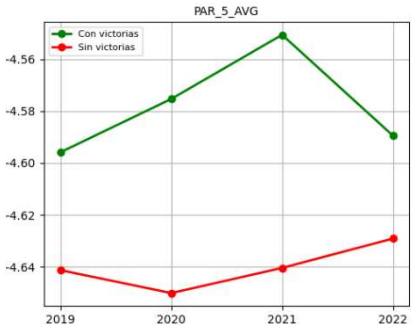

MONEY
1e6
Con victorias
Sin victorias
6
5
4
3
2
1
2019
2020
2021
2022

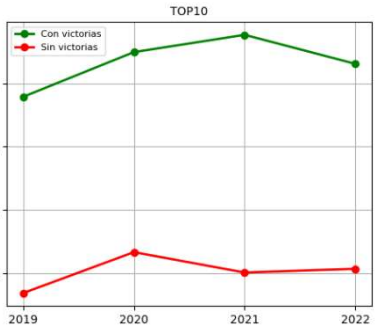

MONEYFORPOINT
Con victorias
Sin victorias
0.000400
0.000375
0.000350
0.000325
0.000300
0.000275
0.000250
2019
2020
2021
2022

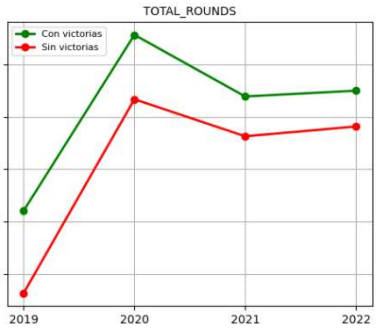

MONEYFORTOP10
1e15
Con victorias
Sin victorias
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
2019
2020
2021
2022


PAR_3_AVG
Con victorias
Sin victorias
−3.03
−3.04
−3.05
−3.06
−3.07
2019
2020
2021
2022

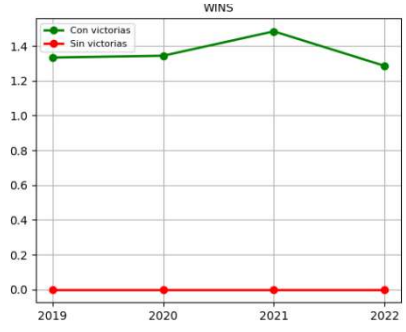

PAR_4_AVG
Con victorias
Sin victorias
−3.99
−4.00
−4.01
−4.02
−4.03
2019
2020
2021
2022


PAR_5_AVG
Con victorias
Sin victorias
-4.56
-4.58
-4.60
-4.62
-4.64
2019
2020
2021
2022


TOP10
Con victorias
Sin victorias
5
4
3
2
2019
2020
2021
2022


TOTAL_ROUNDS
Con victorias
Sin victorias
80
75
70
65
60
2019
2020
2021
2022


WINNER
Con victorias
Sin victorias
1.0
0.8
0.6
0.4
0.2
0.0
2019
2020
2021
2022


WINS
Con victorias
Sin victorias
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
2019
2020
2021
2022

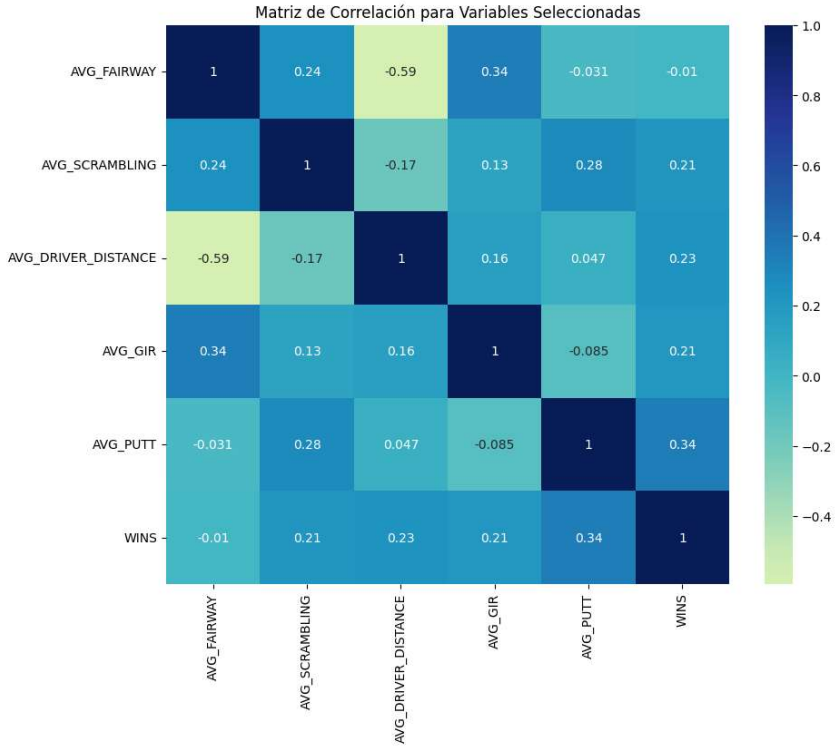
Como se puede ver en los gráficos anteriores, el rendimiento de los jugadores con victorias es muy superior en todos los estadísticos (Excepto en el AVG_DRIVER) que los jugadores sin victoria. Es normal que se dé esto ya que los jugadores que tienen victorias por supuesto, son los que mejor rendimiento van a tener en los diferentes aspectos del juego

Es importante destacar varios aspectos:

Se ve como en la variable AVG_FAIRWAY (El promedio de golpes que el jugador realiza desde el tee y aterriza en la calle (fairway) en un hoyo dado) hay 2 años en los que los jugadores sin victoria tienen mejores rendimientos que los jugadores con victoria, esto se debe a que esta variable, despende mucho del desempeño que se tenga con el Driver, aquí no se valora la distancia que se alcanza, si no el que la bola llegue a calle o no, es decir la precisión en el golpe, siendo el golpe de salida o de Driver uno de los aspectos más difícil de este deporte ya que a mayor distancia en el golpe es mayor la dispersión del golpe

Hay que valorar también la variable Top10, donde se puede ver que lo jugadores con victorias tienen mucho mejor desempeño que los jugadores que no lo tienen. Hay que destacar en este caso que es razonable, ya que a lo largo del año puede haber 35/40 jugadores con victoria, que suelen estar presentes en muchos torneos también en el Top10

4.4 Análisis del juego



Vemos como en el grafico anterior analizamos la correlación entre las estadísticas que únicamente tienen en cuenta la media de golpes en cada aspecto del juego y no otras variables que valoran el desempeño de cada aspecto del juego que veremos más adelante (Las estadísticas "Stroke Gained"

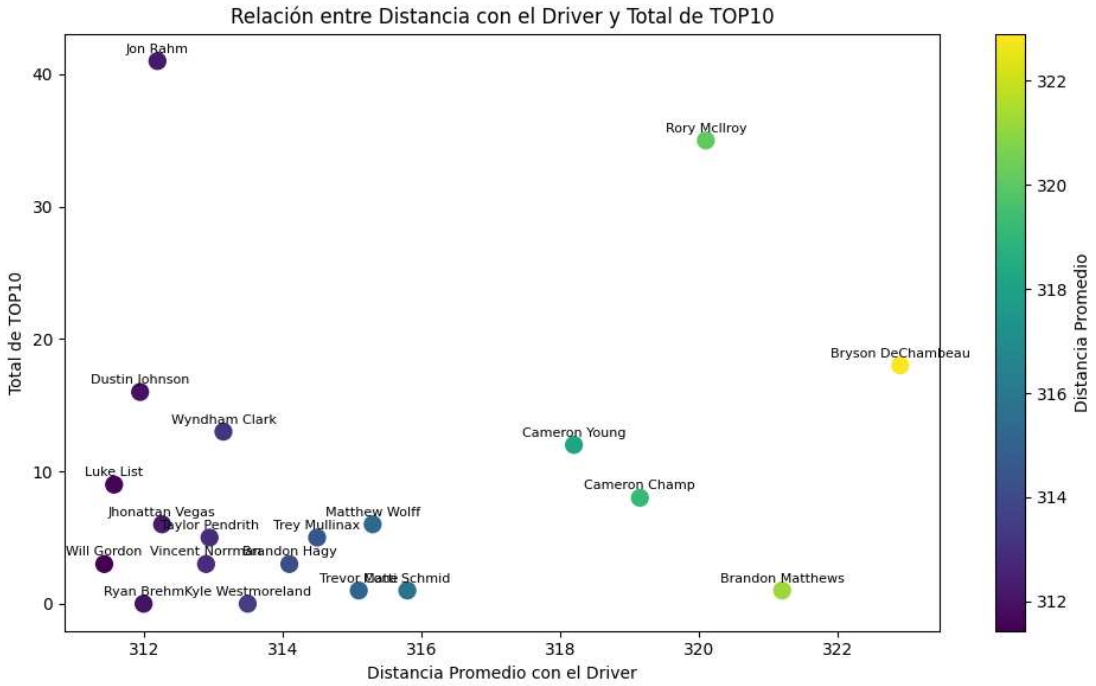
Podemos destacar varios aspectos:

Según estos datos, la variable mas importante para las victorias es el resultado que se obtiene con AVG_PUT que mide la media de golpes obtenidos en el green (Es algo que siempre se ha comentado en este deporte, es famoso el dicho *"El Driver te da la gloria y el putt la victoria"*)

Podemos ver también como la siguiente variable mas correlacionada con las victorias es la distancia que se alcanza con el Driver, seguido del desempeño de AVG_SCRAMBLING así como de AVG_GIR en partes iguales, y el valor AVG_FAIRWAY es el que menor correlación tiene con las victorias (Ya comentamos en un punto anterior el motivo)

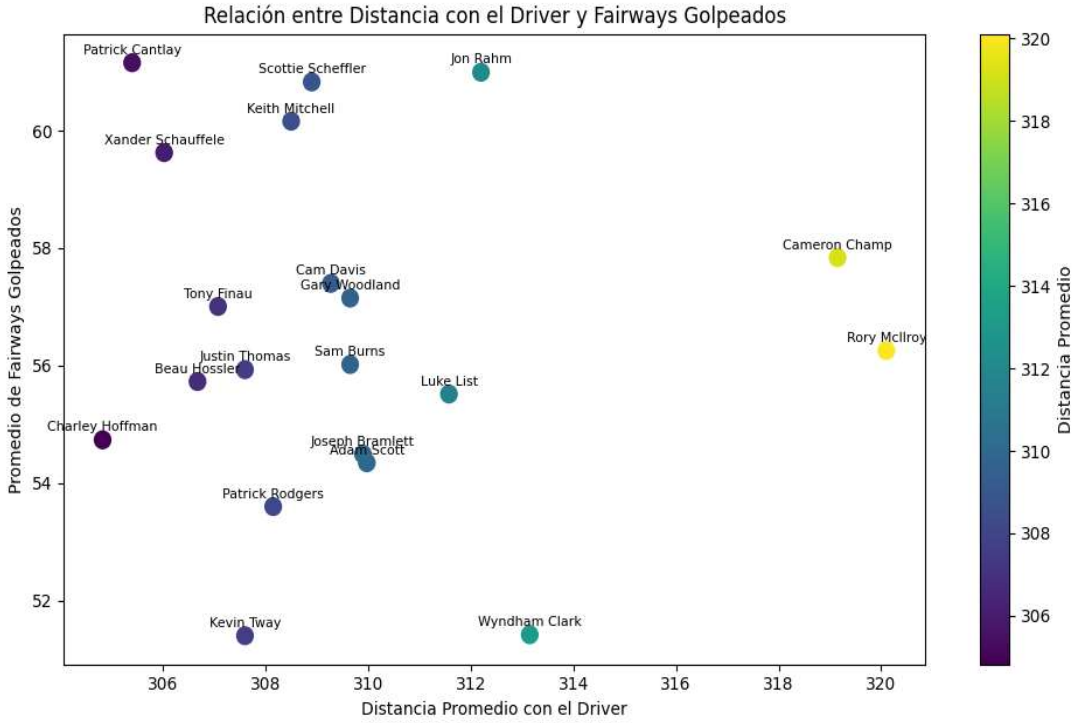
4.4.1 El Driver

El golpe desde el tee con el palo que mas distancia alcanza la bola, muchos expertos dicen que el control de este golpe o de este palo es prácticamente un deporte aparte dentro del propio golf, vamos a ver cuánto de importante es el desempeño de este palo



Podemos observar cómo no siempre los jugadores más pegadores en distancia son los que más TOP10 alcanzan a lo largo de la temporada

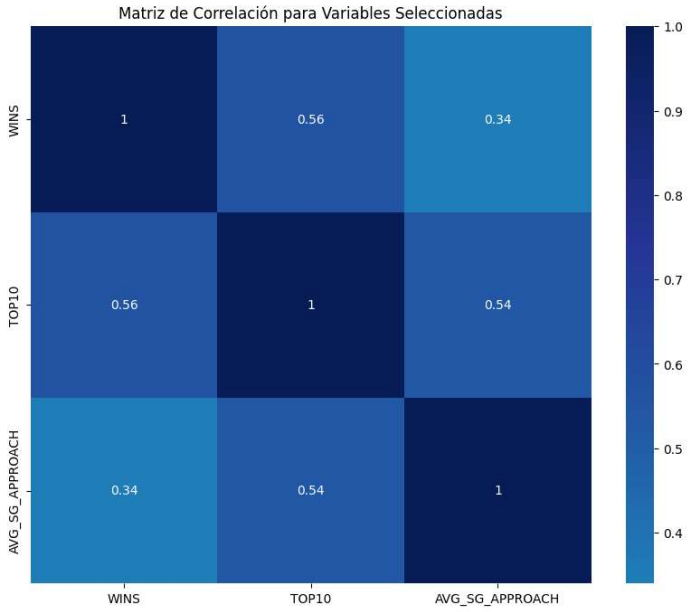
Vamos a analizar cuanto es de importante es la relación entre la distancia alcanzada con el Driver respecto a la precisión, medida por los Fairways alcanzados con este palo



Vemos que a excepción de 2 jugadores muy pegadores como son Cameron Champ y uno de los mejores jugadores históricos Rory Mcilroy la mayoría de jugadores están en promedios de menos de 312 yardas y menos de 58 fairway promedio alcanzados, podríamos destacar como “Outliers” a Jon Rahm, Scottie Scheffer o Keith Michel, los cuales tienen mas de 60 Fairways alcanzados, con golpes que de distancia de media superan las 308 yardas

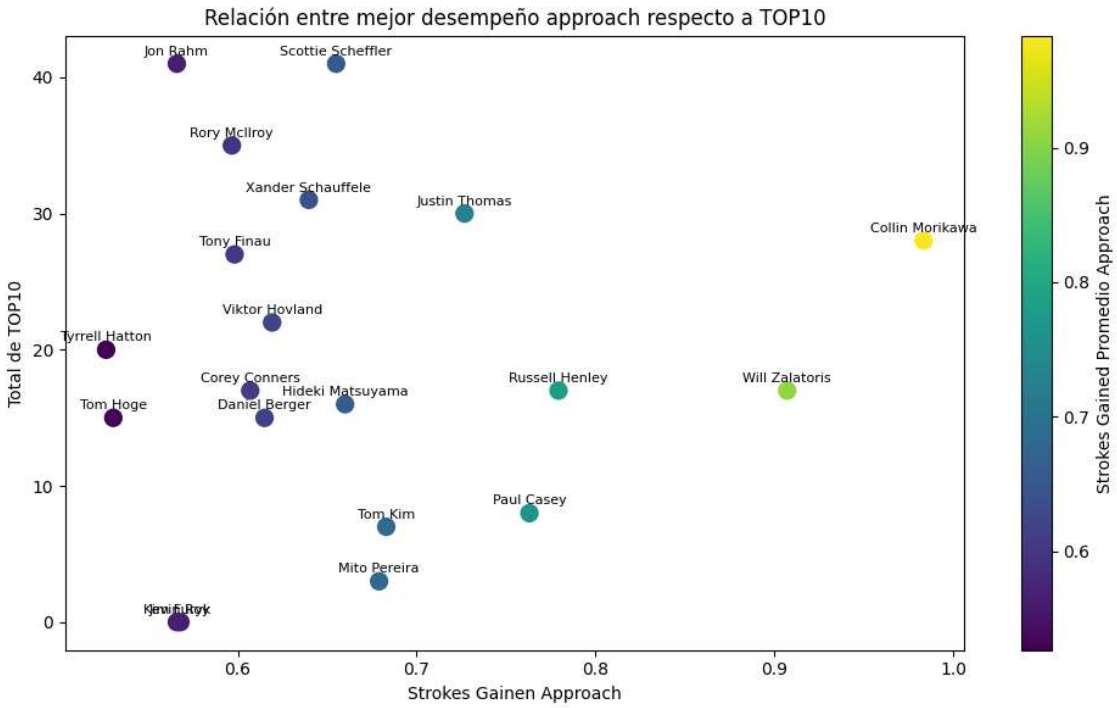
4.4.2 El Approach

El "approach" en golf se refiere a los tiros que un golfista realiza para acercarse al green, específicamente cuando está a una distancia moderada del hoyo. El objetivo principal del approach es dejar la pelota lo más cerca posible del hoyo para facilitar el siguiente putt (golpe en el green) y aumentar las posibilidades de hacer un buen score en ese hoyo.

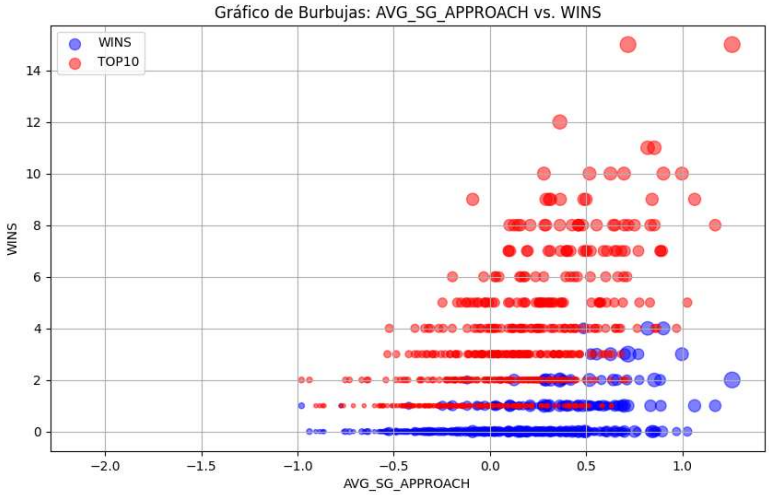


Se puede ver como el rendimiento en este golpe esta muy correlacionado con las Victorias y los Top10

Para valorar el mejor desempeño de este golpe, vamos a tener en cuenta la estadística AVG_SG_APPROACH



Podemos observar como los jugadores con mas “Strokes Gained” en este aspecto del juego son los que mejor rendimiento obtienen en la variable Top10

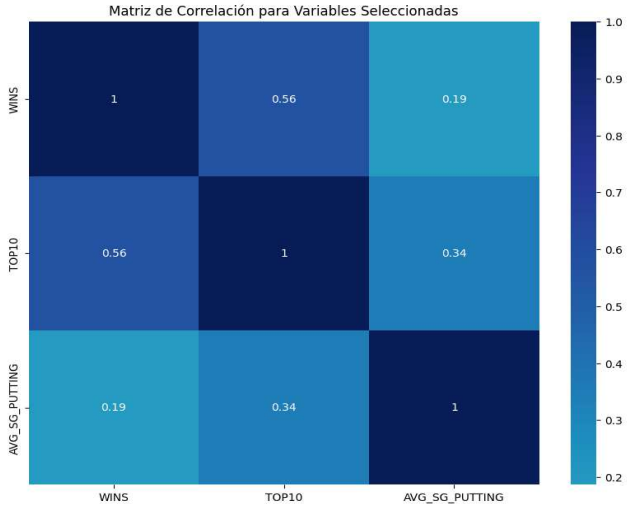


En este gráfico de burbujas se puede observar como los Top10 y las Wins, se dan en jugadores que tienen rendimientos positivos en este aspecto del juego, y como hay pocos jugadores que en este estadístico estando por debajo de -0.5 obtengan victorias o Top10

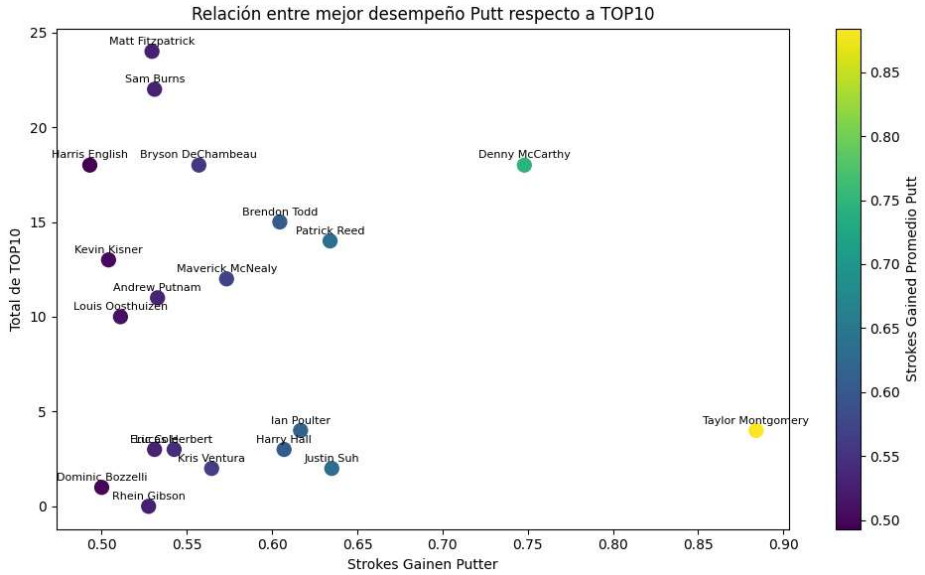
4.4.3 Análisis del Putt

Los golpes de putt se caracterizan por ser tiros de corta distancia y velocidad baja, diseñados para una mayor precisión en lugar de distancia. El objetivo principal del putt es completar un hoyo en el menor número de golpes posible, ya que la mayoría de los hoyos se completan con uno o 2 putt una vez que la pelota ha alcanzado el green.

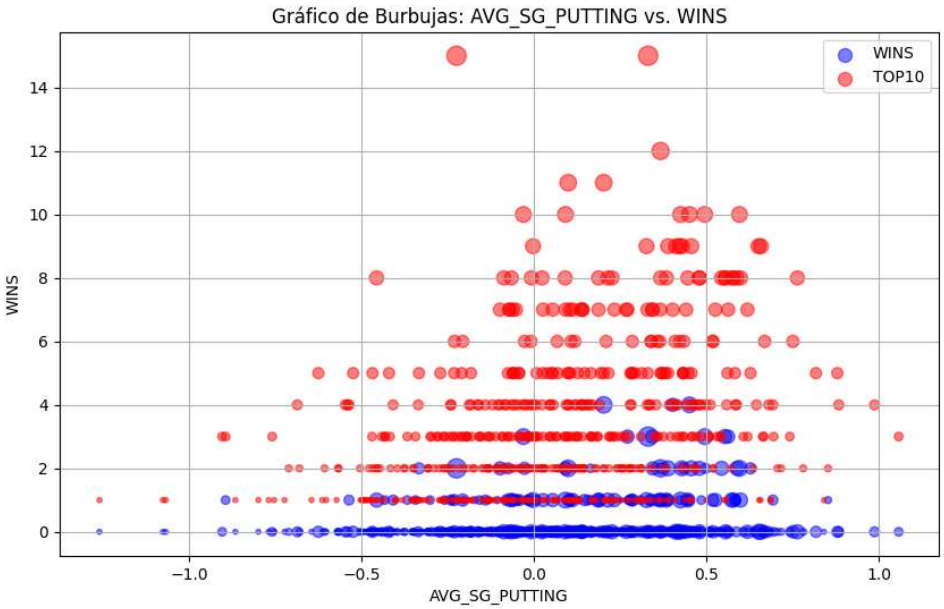
Aquí también vamos a valorar el rendimiento en este golpe con el estadístico AVG_SG_PUTT



Podemos observar como el rendimiento en este golpe esta muy correlacionado con Wins y Top10



En este grafico podemos relacionar el rendimiento del Putt respecto a los Top10 obtenidos aquí podemos observar como los nombres de jugadores cambian drásticamente respecto las otras estadísticas estudiadas, esto se debe a que hay jugadores con rendimientos realmente muy buenos en este aspecto del juego pero que no son tan consistentes en otros aspectos

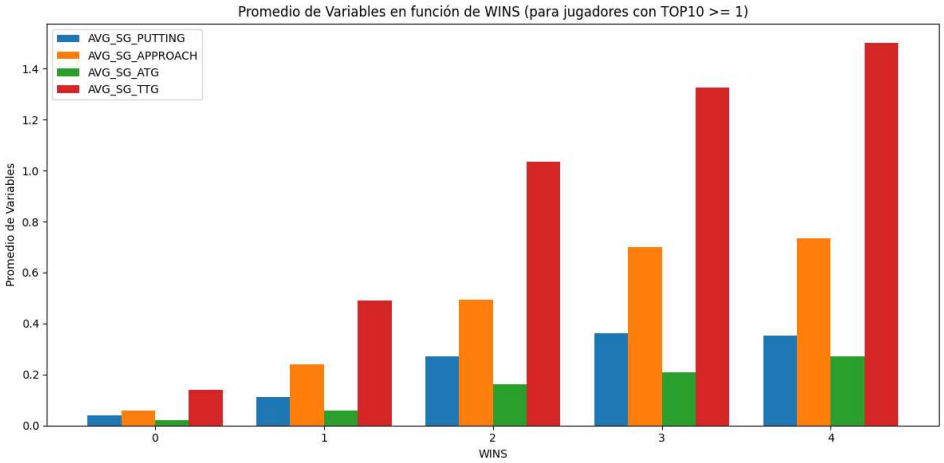


Como vimos en el estadístico anterior, aquí también queda claro que a mejores rendimientos en este aspecto del juego, mejores resultados, aunque en este aspecto se ve como jugadores con rendimientos negativos en el stroke gained, también alcanzan victorias y top10.

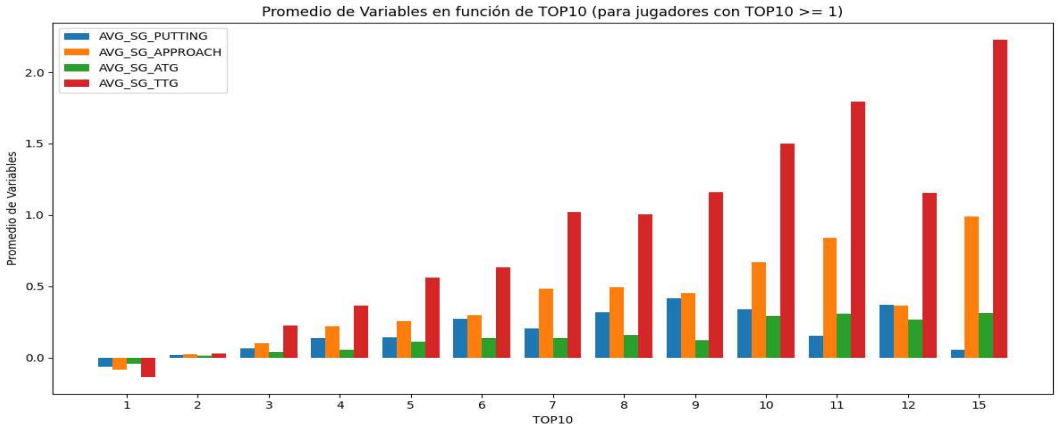
4.4.4 Análisis global de estadísticos Stroke Gained

Este tipo de estadísticos miden el desempeño en un aspecto del juego concreto, veamos cual de estos estadísticos es mas importante para poder alcanzar la victoria

Queda claro que AVG_SG_TTG (golpes ganados o perdidos por un jugador en comparación con el promedio del campo en todos los aspectos del juego desde el tee hasta el green. Esta estadística abarca los tiros desde el tee (drive), tiros de aproximación al green y tiros desde fuera del green) es estadístico mas importante y a continuación AVG_SG_APPROACH (Promedio de Golpes Ganados o Perdidos en Tiros de Aproximación)



Podemos ver como si comparamos estos estadisticos con el Top10 se mantiene lo indicado:



5 Modelos Predictivos

Una vez que hemos analizado todas las estadísticas y los aspectos más relevantes del juego, vamos a desarrollar unos modelos predictivos hasta que demos con el mas óptimo para que nos ayude a predecir los rendimientos en diferentes estadísticas por parte de los jugadores

Para empezar, quiero indicar que como solo hay determinados jugadores que han ganado (un porcentaje pequeño) y nuestra variable objetivo, lo que más nos importa por encima de todo son las victorias, para desarrollar el mejor modelo predictivo vamos a analizar jugadores que tienen alguna victoria y los que no.

(En el documento técnico que hay en el re4positorio hemos estudiado varios modelos predictivos, pero en este trabajo solo hacemos constancia del mejor modelo obtenido)

De esta manera creamos un problema de clasificación donde una vez que hemos creado varios modelos el mejor rendimiento nos lo da el modelo llamado **RF_MODEL**

Usamos el algoritmo de aprendizaje automático ampliamente utilizado en la clasificación y regresión llamado Random Forest.

Es una técnica basada en ensambles que combina múltiples árboles de decisión para obtener un modelo más robusto y preciso, y por lo tanto mejores predicciones

El rendimiento del mejor modelo obtenido:

$$Accuracy\ of\ Random\ Forest\ classifier\ on\ training\ set: \mathbf{0.92}$$
$$Accuracy\ of\ Random\ Forest\ classifier\ on\ test\ set: \mathbf{0.92}$$
$$ROC\ AUC\ Score: \mathbf{0.8591597796143251}$$

A través de este modelo, también hemos determinado cuales son las características mas importantes para obtener victorias en el PGA Tour

$$Top\ 10\ Feature\ Importance:$$
$$Index(['AVG\_SG\_TOTAL', 'AVG\_SG\_TTG', 'PAR\_5\_AVG', 'AVG\_SG\_APPROACH',$$
$$'MONEYFORTOP10', 'MONEYFORPOINT', 'AVG\_PUTT', 'AVG\_SG\_PUTTING',$$
$$'PAR\_4\_AVG', 'PAR\_3\_AVG'])$$

Vemos como los rendimientos del modelo de clasificación en el set de prueba y el set para test son del 0.92, esto supone un dato realmente bueno ya que el modelo clasifica correctamente el 92% de las muestras tanto del conjunto de entrenamiento como de prueba

El Área Bajo la Curva ROC (ROC AUC) es una métrica utilizada para evaluar la capacidad de un modelo para distinguir entre clases positivas y negativas. Un valor de 0.859 sugiere que el modelo tiene un buen rendimiento en la clasificación binaria

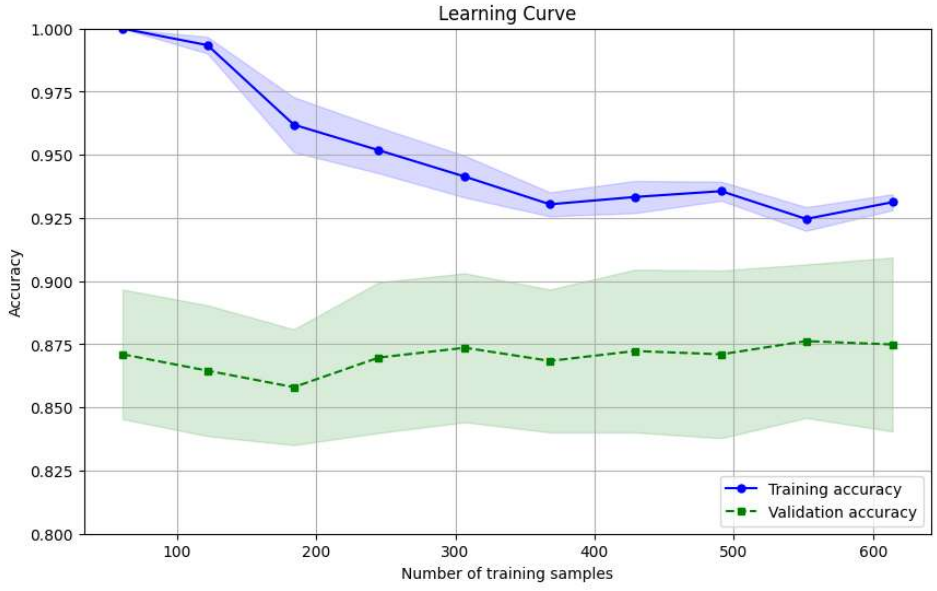
La importancia de las características se calcula en función de cómo contribuyen al rendimiento del modelo

El rendimiento es tan bueno, que nos quedaban dudas se si existía sobreajuste, el sobreajuste hace referencia a una situación en la que un modelo de machine learning se ajusta demasiado a los datos de entrenamiento y, como resultado, pierde su capacidad de generalización en datos nuevos y no vistos.

En otras palabras, el modelo se "memoriza" los datos de entrenamiento en lugar de aprender patrones subyacentes que se pueden aplicar a otros conjuntos de datos.

Hemos obtenido un valor Overfitting score de 0.01 en el modelo lo que nos indica que el modelo esta balanceado y equilibrado y que no existe sobreajuste

Como vemos en el grafico siguiente, la curva de aprendizaje del modelo ajusta la curva de entrenamiento según va entrenando el modelo y se ajusta su accuracy, cercano al set de validación



Para ver el desempeño de nuestro modelo, vamos a analizar que jugadores han ganado realmente en 2022 y que jugadores según el modelo ganarían en este 2022 según su rendimiento:

Cuando consultamos al modelo sus predicciones de jugadores ganados y los jugadores que realmente han ganado torneos nos indica lo siguiente:

PLAYER_NAME	PREDICTED_WINNER_2022	WINS
Scottie Scheffler	1	2
Rory McIlroy	1	2
Jon Rahm	1	4
Viktor Hovland	1	1
Rickie Fowler	1	1
Max Homa	1	2
Justin Rose	1	1
Wyndham Clark	1	2
Jason Day	1	1
Tony Finau	1	2
Russell Henley	1	1
Tom Kim	1	1
Brian Harman	1	1
Keegan Bradley	1	2
Corey Conners	1	1
Nick Taylor	1	1
Matt Fitzpatrick	1	1
Sam Burns	1	1
Akshay Bhatia	1	1
Sepp Straka	1	1

Vemos como en 20 jugadores acierta y determina que ganarían algún torneo en 2022 y asi ha sido, pero para ver el correcto desempeño del modelo, necesitamos saber que jugadores no ha predicho el modelo que ganarían, y si lo hicieron

Jugadores que ganaron en 2022 pero no los predijo el modelo:

PLAYER_NAME	WINS
Si Woo Kim	1
Taylor Moore	1
Chris Kirk	1
Emiliano Grillo	1

Lee Hodges	1
Kurt Kitayama	1
Adam Svensson	1
Nick Hardy	1
Seamus Power	1
Vincent Norrman	1
Lucas Glover	2
Mackenzie Hughes	1
Matt Wallace	1
Davis Riley	1
Nico Echavarria	1

Considero que el resultado del modelo de machine learning es muy bueno ya que el número de jugadores totales que ganaron algún torneo en 2022 es de 35 jugadores, y el modelo ha acertado 20 de ellos. El resto, los 15 jugadores que el modelo no ha predicho no se pueden considerar outliers, pero si quiero destacar que excepto un par de jugadores, el resto no tienen un bagaje relevante histórico en el tour.

6 Conclusiones

El análisis de datos desempeña un papel fundamental en las estadísticas deportivas, y puede tener un impacto significativo en el deporte profesional.

A continuación, presento algunas conclusiones sobre la importancia del análisis de datos en el golf y cómo puede afectar al deporte profesional:

Mejora del Rendimiento Individual: El análisis de datos permite a los golfistas profesionales evaluar su propio rendimiento de manera más objetiva. Al recopilar y analizar estadísticas sobre sus tiros, putts, distancia y precisión, los jugadores pueden identificar áreas de mejora y ajustar su enfoque de entrenamiento. Esto puede ayudar a los golfistas a perfeccionar sus habilidades y aumentar su competitividad.

Optimización de Estrategias de Juego: Los datos pueden proporcionar información valiosa sobre cómo abordar un campo de golf específico. Los profesionales pueden utilizar análisis previos para desarrollar estrategias de juego más efectivas, como seleccionar los palos adecuados para cada hoyo o evitar áreas de alto riesgo. Esto puede marcar la diferencia en términos de resultados y puntuaciones.

Evaluación de Condiciones de Campo: El análisis de datos también se utiliza para evaluar las condiciones del campo, como la velocidad del green, la calidad del césped, la dirección

del viento y otros factores ambientales. Esta información ayuda a los golfistas a adaptar su juego en tiempo real y tomar decisiones informadas sobre cómo jugar en un campo específico.

Apoyo a la Toma de Decisiones de los Entrenadores: Los entrenadores y equipos técnicos de golf utilizan el análisis de datos para evaluar el rendimiento de sus jugadores y tomar decisiones sobre la estrategia de entrenamiento. Pueden identificar las fortalezas y debilidades de un jugador y diseñar programas de entrenamiento personalizados para mejorar el juego.

Avance en la Investigación Científica: El análisis de datos en el golf también ha impulsado la investigación científica en áreas como la biomecánica y la cinemática del swing, la psicología deportiva y la fisiología del ejercicio. Esto ha llevado al desarrollo de equipos y técnicas de entrenamiento más avanzados.

Aumento de la Emoción y la Participación de los Aficionados: Los datos y las estadísticas en tiempo real están disponibles para los aficionados al golf a través de diversas plataformas y aplicaciones. Esto ha aumentado la emoción en los torneos y ha involucrado a los espectadores de manera más activa, ya que pueden seguir y comprender mejor el rendimiento de los jugadores.

Toma de Decisiones Estratégicas para Equipos y Patrocinadores: Los equipos de golf y los patrocinadores utilizan el análisis de datos para tomar decisiones estratégicas, como la selección de jugadores, la inversión en equipos y la planificación de estrategias de marketing. Los datos pueden influir en la dirección de un equipo y en las inversiones en el deporte.

En resumen, el análisis de datos desempeña un papel esencial en el golf profesional al mejorar el rendimiento de los jugadores, optimizar las estrategias de juego, brindar información valiosa a entrenadores y equipos técnicos, y aumentar la participación de los aficionados. Además, contribuye a la evolución continua del deporte y su crecimiento en términos de competitividad y audiencia. En un mundo cada vez más impulsado por los datos, su importancia en el golf y otros deportes profesionales es innegable.

Referencias:

Libro: Johnson, M. L. (2018). Statistical Analysis in Golf. CRC Press.

Articulo: Smith, J. D., & Brown, A. R. (2019). "Analyzing Golfer Performance: A Bayesian Approach." Journal of Sports Analytics, 5(2), 87-102.

Libro: Zeng, J., & Wang, Y. (2017). Advanced Golf Analytics. Springer.

Artícul: Garcia, R., & Kim, S. (2020). "Predicting Golfer Performance using Machine Learning Techniques." Journal of Sports Sciences, 38(6), 643-655.

Libro: Madden, J. (2016). Data Science in Golf: Course Management and Performance Analysis. CreateSpace Independent Publishing Platform.

Artículo: White, L. S., & Jackson, P. L. (2018). "Exploring the Relationship Between Driving Accuracy and Putting Performance in Professional Golf." International Journal of Sports Science & Coaching, 13(6), 898-907.

Libro: Johnson, A. (2019). Golf Statistics Journal: A Comprehensive Guide to Golf Data Analytics. Independently published.

Libro: Kim, Y. J., & Chen, S. (2015). Applied Statistics in Golf. Wiley.

Artículo de Revista: Lee, C. K., & Park, J. H. (2021). "A Data-Driven Approach to Golfer Performance Prediction: A Case Study of the European Tour." Journal of Sports Analytics, 7(3), 197-212.

Terminos relevantes del jugego – Glorario

Green: El área cuidadosamente preparada y corta del campo de golf con césped más corto y suave alrededor del hoyo, donde los golfistas intentan terminar cada hoyo con el menor número de golpes posibles.

Fairway: La parte bien mantenida del campo de golf que se encuentra entre el tee de salida y el green. Es el área donde los golfistas tratan de golpear sus tiros desde el tee.

Tee: La ubicación desde donde se inicia cada hoyo en el campo de golf. Puede ser una pequeña plataforma elevada con un marcador de tee.

Hoyo: El objetivo del juego en cada etapa del campo de golf. Cada hoyo tiene un número designado de golpes para completarlo.

Putt: Un tiro suave y corto que se realiza en el green con un putter. El objetivo es rodar la pelota hacia el hoyo.

Driver: El tiro más largo y potente que se realiza desde el tee de salida en un hoyo, generalmente con un driver.

Handicap: Una medida de la habilidad de un golfista que se utiliza para igualar a jugadores de diferentes niveles. Cuanto menor es el handicap, mejor es el golfista.

Bunker: Un obstáculo de arena en el campo de golf, también conocido como trampa de arena. Los golfistas intentan evitar caer en ellos ya que pueden dificultar el siguiente tiro.

Birdie: Terminar un hoyo en un golpe menos que el par del hoyo.

Eagle: Terminar un hoyo en dos golpes menos que el par del hoyo.

Bogey: Terminar un hoyo en un golpe más que el par del hoyo.

Par: El número estándar de golpes que se espera que un golfista profesional necesite para completar un hoyo o un campo de golf completo.

Caddie: Una persona que lleva los palos y brinda asesoramiento y apoyo a un golfista durante una ronda de golf.

Rough: El área de césped más largo y sin cortar que rodea el fairway en un campo de golf. Los tiros desde el rough suelen ser más difíciles de controlar.

Putter: Un palo especializado diseñado para hacer putts en el green. Tiene un ángulo de loft bajo.

Slice: Un tiro que se desplaza hacia la derecha (para un golfista diestro) de manera incontrolada.

Hook: Un tiro que se desplaza hacia la izquierda (para un golfista diestro) de manera incontrolada.

Green Fee: El costo que se debe pagar para jugar en un campo de golf.