Trabajo Fin de Grado

Ingeniería de las Tecnologías Industriales

Predicción de resultados de Formula 1 mediante técnicas de Machine Learning

Autor: Ignacio Marín Torres

Tutor: Alicia Robles Velasco

Dpto. de Organización Industrial y Gestión de Empresas II

Escuela Técnica Superior de Ingeniería

Universidad de Sevilla

Sevilla, 2023

Trabajo Fin de Grado
Ingeniería de Tecnologías Industriales

Predicción de resultados de Formula 1 mediante técnicas de Machine Learning

Autor:
Ignacio Marín Torres

Tutor:
Alicia Robles Velasco
Profesor Ayudante Doctor

Dpto. de Organización Industrial y Gestión de Empresas II
Escuela Técnica Superior de Ingeniería
Universidad de Sevilla

Sevilla, 2023

Trabajo Fin de Grado: Predicción de resultados de Formula 1 mediante técnicas de Machine Learning

Autor: Ignacio Marín Torres
Tutor: Alicia Robles Velasco

El tribunal nombrado para juzgar el Proyecto arriba indicado, compuesto por los siguientes miembros:

Presidente:

Vocales:

Secretario:

Acuerdan otorgarle la calificación de:

Sevilla, 2023

El Secretario del Tribunal

A mi familia

A mis amigos

A mi Piña

A todos los que me han acompañado estos años

Agradecimientos

A mis padres por ser mi fuerza de empuje y son las personas que más confían en mí, cada logro mío les pertenece tanto como a mí. A las amistades tan bonitas que he creado en estos años en la escuela, han sido el motor, la motivación y la desconexión tan necesaria a veces en toda la carrera universitaria. A las personas que ya no están a mi lado que han sido tan importantes como las demás, han sido también parte de la etapa posiblemente más bonita de mi vida hasta ahora. A mi tutora Alicia Robles por su apoyo, paciencia y comprensión. Y a mí, por haber luchado muchas veces solo, incluso peleado contra la desmotivación, momentos duros y dolorosos, me he demostrado mucho a mí mismo.

Ignacio Marín Torres

Sevilla, 2023

Resumen

La Fórmula 1, más que un deporte, es una amalgama de pasión, tecnología y estrategia, capturando la atención de millones de aficionados a nivel mundial. En 2023, la popularidad de la F1 ha alcanzado nuevos horizontes, evidenciado por el impresionante número de 480,000 asistentes en el Gran Premio Británico, reflejando un interés creciente más allá de los niveles previos a la pandemia. Esta tendencia ascendente también se refleja en su salud financiera, con ingresos que aumentaron a $2,573 millones en 2022, marcando un retorno significativo desde los desafíos de 2020.

Los aficionados de la F1 no son meros espectadores; son una parte integral de su cultura, viviendo intensamente cada carrera. La audiencia televisiva acumulada en 2021 fue de 1.55 mil millones, teniendo picos de audiencias como en el apasionante final de temporada en Abu Dabi, que atrajo a 108.7 millones de espectadores.

En este contexto, nuestro estudio se enfoca en la incorporación de tecnologías inteligentes, como el Machine Learning, en la Fórmula 1. La integración de estas tecnologías no solo enriquece la experiencia de los espectadores y aficionados, sino que también ofrece herramientas valiosas para los equipos, combinando el deporte físico con la ingeniería de vanguardia. Exploramos cómo el Machine Learning, que ha logrado avances significativos siendo una rama de la inteligencia artificial, puede ser aplicado para predecir resultados de carreras, analizando patrones y tendencias en datos históricos y ofreciendo un modelo predictivo preciso y profundo. Este enfoque destaca la singularidad de la Fórmula 1 como un deporte que fusiona intensamente la emoción humana con la precisión tecnológica.

Abstract

Formula 1, more than a sport, is an amalgam of passion, technology and strategy, capturing the attention of millions of fans worldwide. In 2023, F1's popularity has reached new heights, evidenced by an impressive 480,000 attendees at the British Grand Prix, reflecting a growing interest beyond pre-pandemic levels. This upward trend is also reflected in its financial health, with revenues rising to $2,573 billion in 2022, marking a significant return from the challenges of 2020.

F1 fans are not mere spectators; they are an integral part of its culture, living each race intensely. The cumulative TV audience in 2021 was 1.55 billion, with peak audiences such as the thrilling season finale in Abu Dhabi attracting 108.7 million viewers.

In this context, our study focuses on the incorporation of smart technologies, such as Machine Learning, in Formula 1. The integration of these technologies not only enriches the experience for spectators and fans, but also offers valuable tools for teams, combining physical sport with cutting-edge engineering. We explore how Machine Learning, which has made significant advances as a branch of artificial intelligence, can be applied to predict race outcomes, analysing patterns and trends in historical data and providing accurate and insightful predictive modelling. This approach highlights the uniqueness of Formula 1 as a sport that intensely fuses human emotion with technological precision.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1 INTRODUCCIÓN Y OBJETIVOS

Desde su concepción en la década de 1950, la Fórmula 1 ha capturado la imaginación de entusiastas del automovilismo de todo el mundo. Este deporte es considerado la cúspide del automovilismo, no sólo por la adrenalina que desencadena en sus seguidores, sino también por su complejidad técnica y estratégica. Representa no solo un escaparate de habilidad y destreza de los pilotos, sino también un crisol de innovación tecnológica, donde ingenieros y científicos buscan constantemente romper las barreras de lo posible. A lo largo de las décadas, hemos sido testigos de cómo la F1 ha transformado y reinventado la percepción de las carreras, incorporando avances tecnológicos y estratégicos en su tejido mismo. Estas incorporaciones han añadido una capa de impredecibilidad, haciendo que cada carrera sea una experiencia única en sí misma, un espectáculo donde la estrategia y la tecnología se fusionan en un ballet de velocidad. Sin embargo, esta impredecibilidad ha generado un interés creciente en el desarrollo de métodos que puedan anticipar, con cierta precisión, los posibles resultados de las carreras.

La Fórmula 1, siendo el pináculo del deporte automovilístico, ha atraído a millones con su promesa de velocidad y competencia. No obstante, esta impredecibilidad, aunque es un fuerte atractivo para los espectadores y un testimonio del nivel de competencia, ha generado un deseo palpable de desarrollar metodologías que puedan prever los posibles desenlaces de las carreras. En un mundo donde la precisión y la anticipación pueden marcar la diferencia entre la victoria y la derrota, la capacidad de predecir resultados se ha convertido en una herramienta deseada e imprescindible, en constante desarrollo y mejora, y cuya perfección podría revolucionar la forma en que se abordan las estrategias en la F1.

Entramos ahora en una era caracterizada por una explosión de datos. Cada acción, cada movimiento, cada decisión está siendo registrada, creando un tesoro de información esperando ser desentrañado. Esta acumulación masiva de datos representa una oportunidad sin precedentes para comprender mejor y predecir eventos complejos. Con el avance exponencial de la tecnología, herramientas como el Machine Learning (ML) han emergido como esperanza en diversos campos. Desde su aplicación en el análisis financiero, donde se predice el movimiento de los mercados, hasta la medicina personalizada, donde se configuran tratamientos basados en genomas individuales, el ML está remodelando nuestra percepción del futuro. En este contexto, es natural explorar cómo estas tecnologías pueden ser aplicadas en el ámbito deportivo, y más concretamente, en la Fórmula 1, donde la combinación de habilidades humanas y avances técnicos produce resultados fascinantes.

Figura 1: Información mostrada en una carrera con predicciones basadas en ML

Dado este contexto, es inevitable preguntarse: ¿Cómo puede el Machine Learning revolucionar el mundo del deporte? Y más específicamente, ¿cómo puede influir en el ámbito de la Fórmula 1? Esta investigación surge de tales cuestionamientos, pero también de un interés por entender los matices y las dinámicas que configuran este deporte espectacular.

La propuesta de este trabajo es sencilla en su declaración, pero profunda en sus implicaciones. Busca fusionar dos mundos: el creciente interés por el deporte motor y la meticulosa precisión de la ciencia de datos. La aspiración es desarrollar un modelo de Machine Learning robusto y preciso que pueda predecir los resultados de las carreras de Fórmula 1. Pero no es solo una cuestión de predicción; es una exploración para descifrar patrones ocultos, tendencias subyacentes y variables no reconocidas que influyen en los desenlaces de las carreras. Es un reto para entender y, potencialmente, anticipar el futuro de las carreras de F1, y en última instancia, ofrecer una perspectiva fresca y renovada sobre cómo la ciencia de datos puede ser un valioso aliado en el mundo del deporte.

Para concretar este ambicioso objetivo, es fundamental establecer metas claras y específicas. Primero, se debe recopilar y tratar un conjunto de datos históricos de carreras de F1 que abarquen, si es posible, desde sus inicios en 1950 hasta la actualidad. Esta base de datos será la piedra angular sobre la que se construirá el modelo. Además, será esencial realizar un análisis exhaustivo de estos datos, identificando variables clave, tendencias y patrones que puedan influir en los resultados de las carreras. Con estos objetivos claros en mente, en este proyecto nos embarcamos en una interesante mirada hacia el futuro de las predicciones en la Fórmula 1.

El objetivo primordial de este proyecto es diseñar y poner en práctica un modelo predictivo basado en técnicas avanzadas de Machine Learning que pueda prever con precisión los resultados de las carreras de Fórmula 1. Si bien el objetivo general puede parecer amplio, el enfoque principal es predecir qué pilotos tendrán el honor de terminar en el podio, una hazaña que es el resultado de una combinación de habilidad, estrategia, e incluso de suerte. Esta meta no es trivial, dada la multitud de factores que influyen en cada carrera. Para alcanzar este objetivo principal, se han establecido los siguientes objetivos específicos:

1. Recopilación y limpieza de un conjunto de datos históricos de carreras de Fórmula 1, abarcando desde 1950 hasta la actualidad si es posible.
2. Análisis exploratorio de los datos para identificar patrones, tendencias y posibles variables influyentes en los resultados de las carreras.

3. Selección y aplicación de técnicas adecuadas de preprocesamiento de datos para prepararlos para el entrenamiento de modelos de ML.
4. Diseño, entrenamiento y validación de varios modelos de Machine Learning, con el fin de seleccionar el modelo que ofrezca las mejores predicciones.
5. Evaluación del modelo final en términos de precisión, robustez y aplicabilidad en escenarios reales.

Es importante destacar que este trabajo se centra en una exploración académica de la capacidad de las técnicas de Machine Learning para predecir resultados en un deporte tan complejo y multifacético como la Fórmula 1. En lugar de intentar proporcionar una herramienta definitiva para la industria, este proyecto se adentra en el reto de discernir patrones en un deporte donde convergen innumerables variables. El objetivo subyacente es ampliar el conocimiento en el campo de la ciencia de datos y su aplicabilidad en contextos deportivos, aprender sobre los desafíos inherentes a este tipo de predicciones e investigar cómo diferentes variables y factores pueden influir en los resultados. Con este trabajo, se aspira a contribuir al ámbito académico, proporcionando una visión y comprensión más profunda de la intersección entre el deporte y la ciencia de datos.

2 FÓRMULA 1

La Fórmula 1 es considerada la cúspide del automovilismo, donde la tecnología y la habilidad humana se entrelazan en un espectáculo de velocidad, estrategia y precisión. Desde su inicio oficial en 1950, la Fórmula 1 ha sido un escaparate de innovación y talento, atrayendo a millones de espectadores en todo el mundo en cada carrera. No es solo una competición entre pilotos, sino también una lucha entre los equipos de ingenieros y técnicos que diseñan y optimizan los monoplazas. Cada detalle, desde la aerodinámica hasta los neumáticos y la estrategia de combustible, puede ser la diferencia entre ganar y perder. Es un deporte donde la toma de decisiones en fracciones de segundo por parte de los pilotos en la pista, así como del equipo en los boxes, determinan el resultado final de cada carrera.

2.1 Breve historia y evolución

La Fórmula 1, iniciada oficialmente en 1950, evolucionó a partir de las carreras de automóviles de gran premio europeas del siglo XX. El término "Fórmula" denota el conjunto de reglas establecidas y supervisadas por la FIA (Federación Internacional del Automóvil) que deben cumplir todos los participantes y automóviles.

El dominio inicial de Alfa Romeo cedió en la década de 1950 a marcas como Ferrari, Mercedes y Maserati. La década de 1960 marcó un período de innovación tecnológica rápida, incluyendo el traslado de los motores a la parte trasera de los coches, la introducción de alerones para mejorar la aerodinámica, y avances en los neumáticos y sistemas de suspensión. En la Figura 2 se muestran los equipos ganadores del mundial de constructores en todos los años comprendidos entre 1958 y 2020. Se puede comprobar que el equipo que más campeonatos por equipo posee históricamente es la escudería italiana Ferrari.

Figura 2: Ganadores del mundial de constructores 1958-2023. Fuente: reddit.com

En las décadas de 1970 y 1980, la Fórmula 1 se profesionalizó y comercializó aún más. La seguridad se convirtió en una preocupación primordial debido a varios accidentes mortales. Bernie Ecclestone reorganizó la Fórmula 1 en una entidad comercial, aumentando los acuerdos de televisión y patrocinio.

La era moderna ha visto a los equipos convertirse en operaciones de alta tecnología que emplean a cientos de personas. Los avances tecnológicos han continuado con la introducción de sistemas híbridos de recuperación de energía y mejoras en aerodinámica y electrónica.

La Fórmula 1 se ha convertido en un deporte global, atrayendo a millones de espectadores por carrera. Los equipos y pilotos compiten en diversos circuitos alrededor del mundo, cada uno con sus propias características únicas. En la Figura 3 se muestra en rojo los países donde actualmente se disputa algún Gran Premio (GP) y en amarillo los países donde alguna vez lo hubo.

Figura 3: Países donde se ha competido en F1. Fuente: statista.com

Históricamente, la Fórmula 1 ha probado nuevas tecnologías que finalmente han llegado a los automóviles de producción. El éxito depende tanto de la habilidad del piloto como de la eficacia del equipo y la calidad del coche. El siglo XXI ha traído evoluciones constantes a este deporte, tanto tecnológicas como reglamentarias. Se han implementado medidas para aumentar la competitividad y emoción, como restricciones en las pruebas y el desarrollo de los coches durante la temporada, y la introducción del DRS (Sistema de Reducción de Resistencia) para facilitar los adelantamientos. En la Figura 4 se puede ver la comparación del Alfa Romeo 158 (1950) de Farina y del Aston Martin AMR23 (2023) de Alonso

Figura 4: Diferencias entre los primeros F1 y los actuales. Fuente: Elaboración propia

La sostenibilidad ha ganado importancia, con la introducción de motores híbridos y el objetivo de la Fórmula 1 de ser carbono neutral para 2030, lo que ha supuesto cambios en las reglas y la introducción de nuevas tecnologías, planteando nuevos desafíos para los equipos.

2.2 Estructura de un equipo de F1

Un equipo de Fórmula 1 es una organización compleja compuesta por cientos de personas, cada una de ellas especializada en su propia área. En la cima de la jerarquía de un equipo de F1 está el director del Equipo, quien es responsable de la toma de decisiones estratégicas y la gestión global.

Debajo del director, la estructura se divide en diferentes departamentos. El departamento técnico, liderado por el director técnico, se encarga del diseño y desarrollo del coche. Este departamento a menudo se divide en subsecciones más pequeñas, cada una enfocada en un área específica como aerodinámica, chasis, motor y electrónica. El departamento de operaciones en pista gestiona todo lo que sucede durante un fin de semana de carrera, desde los ingenieros de carrera y los mecánicos hasta los estrategas y los analistas de datos. El departamento de logística se encarga de transportar el equipo y el equipamiento a las carreras en todo el mundo, mientras que el departamento comercial se encarga de los patrocinios y las relaciones con los medios.

Finalmente, los pilotos son la cara visible del equipo, pero su papel va más allá de solo conducir el coche. También trabajan estrechamente con los ingenieros para desarrollar y mejorar el coche, proporcionando comentarios vitales basados en su experiencia en la pista. Cada constructor puede contar con un máximo de dos pilotos, donde ambos disponen de un coche idéntico pero cada uno tiene su equipo de ingenieros encargados de poner el coche a punto con los reglajes óptimos (altura del coche, posición de alerones, reparto de peso, etc.). Por tanto, los compañeros de equipo son el mejor indicador donde medir la calidad del piloto, cuanta más diferencia de resultados entre los dos pilotos, mayor diferencia de calidad entre los dos.

2.3 Desarrollo de un fin de semana de Gran Premio

Un fin de semana de Gran Premio de Fórmula 1 es un evento de tres días que incluye prácticas, clasificación y la carrera. Cada fase tiene su propia importancia y estrategia, jugando un papel crucial en el resultado final del Gran Premio.

2.3.1. Entrenamientos Libres

Los fines de semana de Gran Premio comienzan con las sesiones de entrenamientos libres. Estas son oportunidades para que los equipos y los pilotos se familiaricen con el circuito y recojan datos vitales para la configuración del coche. Hay tres sesiones de entrenamientos libres: dos el viernes y una el sábado por la mañana.

Durante estas sesiones, los equipos experimentan con diferentes configuraciones de coche, prueban los diferentes compuestos de neumáticos disponibles y recopilan datos sobre el rendimiento del coche en diferentes condiciones de pista y de combustible. Los pilotos también usan este tiempo para aprender el trazado de la pista, identificando los puntos de frenado, las líneas de carrera ideales y cómo gestionar el desgaste de los neumáticos.

Los datos recopilados durante los entrenamientos libres son analizados en detalle por los ingenieros y los analistas de datos del equipo para preparar la estrategia para la clasificación y la carrera. Aunque los tiempos de vuelta en los entrenamientos libres no determinan la parrilla de salida, son un indicador útil de la velocidad relativa de los equipos y pueden proporcionar una visión preliminar de cómo se desarrollará el resto del fin de semana.

2.3.2. Clasificación

La clasificación tiene lugar el sábado por la tarde y determina el orden de la parrilla de salida para la carrera del domingo. Es un momento crítico del fin de semana, ya que una buena posición en la parrilla puede proporcionar una ventaja significativa en la carrera.

La sesión de clasificación se divide en tres partes: Q1, Q2 y Q3. En Q1, todos los coches tienen 18 minutos para establecer su tiempo más rápido. Los cinco coches más lentos son eliminados y ocupan los últimos lugares de la parrilla. Q2 sigue el mismo formato, pero solo dura 15 minutos y los diez coches más rápidos pasan a Q3. Finalmente, en Q3, estos diez coches tienen 12 minutos para luchar por la primera posición, del inglés *pole position*.

En la Figura 5 se muestra un ejemplo de clasificación con los mejores tiempos de cada piloto en cada sesión. Se puede observar cómo los 5 pilotos más lentos de la Q1 y Q2 quedan eliminados, no compiten en las sesiones posteriores y sus posiciones quedan fijadas.

Qualifying Session Provisional Classification

	NO	DRIVER	NAT	ENTRANT	Q1	LAPS	%	TIME	Q2	LAPS	TIME	Q3	LAPS	TIME
1	44	Lewis HAMILTON		Mercedes-AMG Petronas F1 Team	1:28.343	5	100.000	17:09:56	1:27.586	5	17:42:16	**1:27.264**	6	18:07:38
2	77	Valtteri BOTTAS		Mercedes-AMG Petronas F1 Team	1:28.767	5	100.479	17:10:26	1:28.063	5	17:42:24	**1:27.553**	6	18:07:58
3	33	Max VERSTAPPEN		Aston Martin Red Bull Racing	1:28.885	5	100.613	17:06:14	1:28.025	4	17:41:15	**1:27.678**	6	18:08:11
4	23	Alexander ALBON		Aston Martin Red Bull Racing	1:28.732	6	100.440	17:18:54	1:28.749	6	17:41:26	**1:28.274**	6	18:08:16
5	11	Sergio PEREZ		BWT Racing Point F1 Team	1:29.178	6	100.945	17:10:13	1:28.894	8	17:41:35	**1:28.322**	6	18:08:43
6	3	Daniel RICCIARDO		Renault DP World F1 Team	1:29.005	6	100.749	17:18:42	1:28.648	5	17:47:11	**1:28.417**	6	18:08:23
7	31	Esteban OCON		Renault DP World F1 Team	1:29.203	3	100.973	17:11:03	1:28.937	5	17:45:53	**1:28.419**	6	18:08:31
8	10	Pierre GASLY		Scuderia AlphaTauri Honda	1:28.971	3	100.710	17:10:19	1:29.008	5	17:47:17	**1:28.448**	6	18:07:46
9	4	Lando NORRIS		McLaren F1 Team	1:29.464	5	101.268	17:10:38	1:28.877	6	17:41:22	**1:28.542**	6	18:08:06
10	26	Daniil KVYAT		Scuderia AlphaTauri Honda	1:29.158	6	100.922	17:18:25	1:28.944	5	17:47:07	**1:28.618**	6	18:07:52
11	5	Sebastian VETTEL		Scuderia Ferrari	1:29.142	6	100.904	17:18:04	**1:29.149**	5	17:47:33			
12	16	Charles LECLERC		Scuderia Ferrari	1:29.137	6	100.898	17:18:19	**1:29.165**	5	17:47:24			
13	18	Lance STROLL		BWT Racing Point F1 Team	1:28.679	6	100.380	17:18:49	**1:29.557**	5	17:47:39			
14	63	George RUSSELL		Williams Racing	1:29.294	8	101.076	17:17:57	**1:31.218**	3	17:47:30			
15	55	Carlos SAINZ		McLaren F1 Team	1:28.975	3	100.715	17:10:55	DNF	2				
16	99	Antonio GIOVINAZZI		Alfa Romeo Racing ORLEN	**1:29.491**	6	101.299	17:18:28						
17	7	Kimi RAIKKONEN		Alfa Romeo Racing ORLEN	**1:29.810**	6	101.660	17:18:32						
18	20	Kevin MAGNUSSEN		Haas F1 Team	**1:30.111**	6	102.001	17:18:09						
19	8	Romain GROSJEAN		Haas F1 Team	**1:30.138**	6	102.031	17:18:15						
20	6	Nicholas LATIFI		Williams Racing	**1:30.182**	6	102.081	17:19:15						

POLE POSITION LAP

44	Lewis HAMILTON	Mercedes-AMG Petronas F1 Team	1:27.264	223.267 KM/H

FASTEST LAP

44	Lewis HAMILTON	Mercedes-AMG Petronas F1 Team	1:27.264	223.267 KM/H

Figura 5: Resultados de la sesión de clasificación del GP de Baréin 2020. Fuente: fia.com

Durante la clasificación, los pilotos suelen utilizar el compuesto de neumáticos más rápido disponible para maximizar su velocidad. Incluso, con la normativa hasta 2020 debían tener en cuenta que, los pilotos que pasaban a Q3 debían comenzar la carrera con el set de neumáticos con el que lograron su tiempo más rápido en Q2. Esto influía en la estrategia de los equipos durante la sesión de clasificación para la elección del primer juego de neumáticos de carrera.

2.3.3. Carrera

El clímax del fin de semana de Gran Premio es la carrera, que se celebra el domingo. Esta es donde se otorgan los puntos y se determinan los ganadores. Las carreras de Fórmula 1 son pruebas de velocidad, estrategia y resistencia, con cada carrera durando alrededor de dos horas o una distancia predeterminada, lo que ocurra primero. Los pilotos comienzan en la parrilla de salida en el orden determinado por la clasificación. La salida es un momento crítico de la carrera, donde los pilotos pueden ganar o perder varias posiciones en la primera curva. En la Figura 6 se puede observar cómo segundos después de realizar la salida, los pilotos están muy pegados y un choque entre dos coches podría desencadenar más accidentes.

Figura 6: Salida en el GP de Indianápolis 1990 (EEUU). Fuente: inmotion.dh

Durante la carrera, los pilotos deben manejar una serie de desafíos, desde lidiar con el tráfico y adelantar a los competidores, hasta gestionar el desgaste de los neumáticos y el consumo de combustible. También deben hacer al menos una parada en boxes para cambiar de neumáticos, y la elección del momento para hacerlo es una parte crucial de la estrategia de la carrera.

Las carreras pueden ser afectadas por una serie de factores imprevistos, como el clima, los accidentes y los coches de seguridad. Los equipos y los pilotos deben ser capaces de adaptarse rápidamente a estas circunstancias cambiantes para aprovechar cualquier oportunidad que se presente.

El objetivo final de cada piloto es cruzar la línea de meta en la posición más alta posible. Actualmente, los diez primeros clasificados en cada carrera reciben puntos, que se suman a lo largo de la temporada para determinar los campeonatos de pilotos y constructores.

El sistema de puntuación en la Fórmula 1 ha evolucionado a lo largo de la historia del deporte, adaptándose a los cambios en el número de equipos y pilotos, así como a la necesidad de mantener la competencia emocionante y equitativa.

En el sistema actual, que se ha utilizado desde 2010, los puntos se otorgan a los diez primeros clasificados en cada carrera de la siguiente manera: el ganador recibe 25 puntos, el segundo lugar recibe 18 puntos, el tercer lugar 15, y luego 12, 10, 8, 6, 4, 2 y 1 punto respectivamente para los puestos del cuarto al décimo. Además, desde la temporada 2019, se otorga un punto adicional al piloto que establece la vuelta más rápida en la carrera, siempre que ese piloto termine dentro de los diez primeros. Esto añade un elemento adicional de estrategia a las carreras, ya que los equipos pueden optar por intentar asegurar este punto adicional, incluso si no están en posición de ganar la carrera.

2.3.3.1. Carrera Sprint

La FIA introdujo en 2021 la "Carrera Sprint" para aumentar la emoción y participación de los fans durante los fines de semana de carrera. Estas carreras más cortas, celebradas en eventos seleccionados, determinan la parrilla de salida para la carrera principal del domingo. Inicialmente, la clasificación del viernes establecía la parrilla para la Carrera Sprint del sábado, pero ahora se experimenta con clasificaciones independientes para cada

carrera. Con el nuevo reglamento, se distribuyen más puntos entre los primeros ocho pilotos de la Carrera Sprint, en lugar de solo los tres primeros, como en 2021.

2.4 Componentes clave en la F1

La Fórmula 1 es un deporte que requiere la integración de varios componentes clave para garantizar el éxito en la pista. Los pilotos, los equipos, los circuitos y las variables climáticas son factores críticos que determinan el rendimiento en las carreras.

2.4.1 Piloto

El piloto es un componente esencial en cualquier equipo de F1. No sólo son responsables de manejar el coche durante la carrera, sino que también desempeñan un papel crucial en el desarrollo del vehículo a lo largo de la temporada. La habilidad, la resistencia y la capacidad de tomar decisiones rápidas son algunas de las cualidades esenciales de un piloto de F1. Ellos deben ser capaces de procesar la información a altas velocidades, mantener la calma bajo presión y, lo más importante, tener un profundo conocimiento de cómo su coche responde en diferentes situaciones de carrera.

2.4.2 Equipo

Un equipo de F1 es una organización compleja que incluye no sólo al piloto y al copiloto, sino también a los ingenieros, mecánicos, estrategas y personal de apoyo. Todos trabajan juntos para optimizar el rendimiento del coche en la pista. Los ingenieros y mecánicos se encargan de ajustar y reparar el coche, mientras que los estrategas planifican la estrategia de carrera, incluyendo las paradas en boxes y la elección de neumáticos. El personal de apoyo puede incluir a todos, desde entrenadores físicos hasta psicólogos deportivos, que ayudan a los pilotos a mantenerse en su mejor forma física y mental.

2.4.3 Circuito

Cada circuito de F1 presenta sus propios desafíos únicos. Algunos circuitos, como el de Mónaco, son estrechos y sinuosos, lo que requiere una gran habilidad de conducción y un coche con buena tracción y manejo. Otros, como el de Monza (Italia), son rápidos y abiertos, favoreciendo a los coches con alta velocidad máxima y eficiencia aerodinámica. Los equipos deben adaptar sus coches y estrategias a cada circuito para maximizar su rendimiento.

2.4.4 Variables climáticas

Las variables climáticas, como la lluvia y la temperatura, son un factor crucial en las carreras de F1. Pueden afectar la adherencia a la pista, la visibilidad, el rendimiento del motor y la durabilidad de los neumáticos. Los equipos deben estar preparados para adaptarse rápidamente a estas condiciones cambiantes.

3 METODOLOGÍA

La capacidad de generar predicciones precisas en el mundo de la Fórmula 1 fusiona el desafío del análisis detallado de datos con la intrínseca volatilidad del deporte motorizado. Esta sección tiene como propósito principal detallar el enfoque metodológico empleado para anticipar los desenlaces de las carreras de Fórmula 1, recurriendo a métodos sofisticados de Machine Learning (ML).

La metodología empleada en esta investigación puede ser desglosada en tres etapas esenciales: el análisis descriptivo, la preparación de datos y el análisis predictivo. Durante la etapa de análisis descriptivo, se lleva a cabo una inspección meticulosa de las variables del *dataset*, buscando entender su composición y comportamiento. Posteriormente, en la fase de preparación de datos, se efectúan las modificaciones requeridas para asegurar que el conjunto de datos esté listo y optimizado para el modelo de ML que se implementará.

La fase de análisis predictivo representa el epicentro de este enfoque metodológico. En este segmento, se despliegan herramientas de ML con el objetivo de construir un modelo que pueda anticipar los resultados de las carreras de Fórmula 1 con el máximo nivel de exactitud. Para lograr esta meta, nos adentraremos en un abanico de algoritmos y metodologías, identificando aquel que presente una sintonía óptima con nuestros datos y produzca las predicciones más acertadas. En el contexto del ML, se llevará a cabo un estudio detallado de trabajos previos y literatura que aborden la aplicación de estas herramientas en el dominio de la Fórmula 1.

En el transcurso de esta investigación, recurriremos al lenguaje de programación Python, complementado con librerías reconocidas como *Scikit-learn* y *Pandas* para la construcción del modelo y el procesamiento de datos. La preferencia por Python radica en su versatilidad, el extenso repertorio de herramientas para análisis de datos que propone y su renombre en el mundo de la ciencia de datos. A medida que avancemos en esta sección, ahondaremos en las particularidades y méritos de este lenguaje y sus librerías afines.

3.1 Machine Learning

En el ámbito de la ciencia de los datos, el aprendizaje automático o *Machine Learning* se ha consolidado como una herramienta esencial en el análisis predictivo. Este campo de estudio, que es una rama de la inteligencia artificial, permite a los sistemas informáticos aprender de los datos y mejorar su rendimiento sin la necesidad de ser explícitamente programados para hacerlo. Mediante el uso de algoritmos y modelos, seremos capaces de detectar patrones en conjuntos de datos y hacer predicciones o tomar decisiones basadas en ellos.

Existen diferentes tipos de aprendizaje automático que se aplican en función del tipo de problema que se quiera resolver y la naturaleza de los datos disponibles. Los tres principales son el aprendizaje supervisado, el aprendizaje no supervisado y el aprendizaje por refuerzo.

El aprendizaje supervisado es el más utilizado y se aplica cuando se cuenta con datos etiquetados, es decir, cuando se conoce la variable de salida que se quiere predecir. En este tipo de aprendizaje, el modelo se entrena utilizando un conjunto de datos de entrada y sus correspondientes salidas, con el objetivo de que aprenda a mapear las entradas con las salidas. Una vez que el modelo está entrenado, se puede utilizar para predecir la salida de nuevos datos de entrada. Ejemplos comunes de técnicas de aprendizaje supervisado son la regresión lineal y logística, los árboles de decisión y las redes neuronales.

Por otro lado, el aprendizaje no supervisado se utiliza cuando no se cuenta con datos etiquetados. En este caso, el objetivo es que el modelo aprenda la estructura subyacente de los datos. Las técnicas comunes de aprendizaje no supervisado incluyen el *clustering*, que agrupa los datos en base a su similitud, y la reducción de la dimensionalidad, que se utiliza para reducir el número de variables en un conjunto de datos.

Finalmente, el aprendizaje por refuerzo es un tipo de aprendizaje donde un agente aprende a tomar decisiones basándose en la recompensa que recibe por las acciones que realiza. Este tipo de aprendizaje es comúnmente utilizado en problemas de robótica y en juegos.

El éxito de la aplicación de estas técnicas depende en gran medida de la calidad de los datos disponibles y de la selección del modelo adecuado para el problema. Por lo tanto, el procesamiento de los datos y la elección del modelo constituyen dos aspectos críticos en cualquier proyecto de *Machine Learning*.

Para este trabajo, diversas técnicas serán consideradas, cada una con sus propias ventajas y desventajas, y realizaremos un estudio cuidadoso para seleccionar la más apropiada. Las técnicas que se explorarán incluyen, pero no se limitan a, la regresión lineal y logística, los árboles de decisión y las redes neuronales.

A continuación, haremos una revisión de la literatura existente sobre el uso de *Machine Learning* en el ámbito de la Fórmula 1 para proporcionar un contexto adecuado para esta investigación. Este examen de la literatura servirá para identificar los enfoques que se han tomado previamente y los resultados que se han obtenido, y ayudará a informar la elección de la técnica que se aplicará en este estudio.

3.1.1 Revisión de la literatura científica relacionada

El artículo titulado "*Machine Learning framework for Formula 1 Race Winner and Championship Standings Predictor*" [SICOE, 2022] se centra en investigar la predicción de ganadores de carreras utilizando algoritmos de Machine Learning supervisados como una solución de software inteligente. Además, ofrece un análisis crítico de la literatura realizada en *Machine Learning* en relación con las predicciones de resultados deportivos, destacando importantes procesos metodológicos (fuentes de datos, recolección, implementación y evaluación).

El trabajo presentado en el artículo se centra en el uso de técnicas de *Machine Learning* para predecir los ganadores de las carreras y las posiciones en el campeonato de la Fórmula 1. Al igual que nuestro estudio, este artículo reconoce la importancia de los numerosos factores tanto internos como externos que pueden influir en el resultado de un evento deportivo. Asimismo, el artículo destaca la aplicación creciente de software inteligente en el campo de las analíticas deportivas y cómo estas aplicaciones han revolucionado el deporte al permitir aprovechar al máximo el poder de las técnicas de *Machine Learning*.

Similar a nuestro enfoque, este artículo también utiliza técnicas de aprendizaje supervisado y pone énfasis en los métodos de conjunto y regresión. El objetivo del trabajo es predecir las posiciones del campeonato de la Fórmula 1 para la temporada 2021 basándose en datos históricos, lo cual coincide con nuestro objetivo de predecir los resultados en la Fórmula 1.

No obstante, aunque existen similitudes en los objetivos y técnicas utilizadas, también hay diferencias y limitaciones en relación con nuestro trabajo. Una de las principales diferencias es el enfoque en la predicción de los ganadores de las carreras y las posiciones en el campeonato. En nuestro caso, nos centraremos en una serie de variables adicionales que pueden influir en los resultados de las carreras, como la experiencia del piloto, las características de la pista y el rendimiento del constructor, lo que puede proporcionar una perspectiva más matizada de los factores que determinan el éxito en la Fórmula 1.

El estudio "*The Future of Formula 1 Racing: Neural Networks to Predict Tyre Strategy*" [Piccolomini , 2022] se enfoca en el desarrollo e implementación de redes neuronales, específicamente LSTM y GRU, para prever la estrategia de neumáticos durante una carrera de Fórmula 1. La elección de los neumáticos durante una carrera es un factor crítico en el rendimiento de un equipo, y este artículo examina el uso de algoritmos de aprendizaje profundo para prever con precisión cuándo y qué cambios de neumáticos se requieren durante una carrera.

Esta investigación comparte con nuestro trabajo el enfoque en la Fórmula 1 y la importancia de la estrategia de neumáticos como un factor determinante en el rendimiento de un equipo. Además, ambas investigaciones buscan aplicar algoritmos de aprendizaje automático y analizar grandes cantidades de datos para optimizar la estrategia y mejorar el rendimiento en la carrera.

Sin embargo, existen diferencias claras en términos de la especificidad del enfoque y las técnicas de aprendizaje automático utilizadas. Este trabajo se centra exclusivamente en la predicción de la estrategia de neumáticos, utilizando redes neuronales y, en particular, las LSTM y las GRU. Nuestro trabajo, por otro lado, tiene un enfoque más amplio que no se limita a la estrategia de neumáticos y busca considerar una gama más amplia de factores que pueden influir en el resultado de una carrera de Fórmula 1. Además, el enfoque en las redes neuronales puede limitar la aplicación de los hallazgos a contextos en los que estas técnicas son apropiadas, y puede no ser relevante para situaciones en las que otros enfoques de aprendizaje automático sean más adecuados. En contraste, nuestro trabajo busca aplicar un enfoque más flexible para prever los resultados de las carreras de Fórmula 1.

3.1.2 Regresión logística

La regresión logística es un método estadístico utilizado para modelar la relación entre una variable dependiente binaria y una o más variables independientes. A diferencia de la regresión lineal, que predice valores continuos, la regresión logística predice la probabilidad de que un evento ocurra, siendo especialmente útil para la clasificación. Es decir, el resultado de la predicción es una variable binaria que toma el valor de uno cuando la probabilidad de que ocurra el evento es mayor a 0,5 y de 0 en caso contrario.

Originada en el ámbito de la investigación biométrica, la regresión logística ha encontrado aplicaciones en numerosos campos, desde la medicina hasta las ciencias sociales y el marketing. Su principal ventaja radica en su capacidad para predecir eventos dicotómicos (por ejemplo, si un cliente comprará o no un producto) a partir de variables predictoras, que pueden ser tanto categóricas como continuas.

Matemáticamente, la regresión logística se basa en la función *logit*, que es el logaritmo del cociente entre la probabilidad de que el evento de interés ocurra y la probabilidad de que no ocurra. La ecuación general del modelo es:

$$\log\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = \beta_0 + \beta_1 \mathrm{x}_1 + \beta_2 \mathrm{x}_2 + \ldots + \beta_\mathrm{k} \mathrm{x}_\mathrm{k}$$

Donde:

- p es la probabilidad de que el evento ocurra, y toma valores entre 0 y 1 (incluyendo ambos)
- β_0 es el intercepto.
- $\beta_1, \beta_2, \ldots, \beta_k$ son los coeficientes de las variables predictoras $x_1, x_2, \ldots, x_k$

La estimación de los coeficientes se realiza en base a un histórico de datos en la etapa de entrenamiento del modelo. Tras calcular estos coeficientes, es posible transformar la ecuación para obtener p, la probabilidad buscada:

$$p = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}$$

Para aplicar este método a nuestro caso debemos realizar un análisis descriptivo de los datos disponibles para lo cual nuestra estrategia inicial será la siguiente:

1. **Selección de Variables**: Una vez que tengamos acceso a los datos, identificaremos las variables que podrían influir en la variable objetivo. Esta elección se basará tanto en el conocimiento del dominio como en la correlación observada entre las variables.
2. **Preparación de Datos**: Antes de aplicar el modelo, aseguraremos que los datos estén limpios, gestionaremos los valores faltantes y posiblemente realicemos alguna transformación para garantizar que las variables predictoras sean adecuadas para la regresión logística.
3. **Modelado**: Utilizaremos un conjunto de datos de entrenamiento para entrenar el modelo de regresión logística y un conjunto de datos de prueba para evaluar su rendimiento, es decir para calcular los ceoficientes de las variables.
4. **Evaluación**: Una vez entrenado, evaluaremos el modelo usando métricas, como la precisión, para determinar su capacidad predictiva. Métricas típicas de los sistemas de clasificación binaria, son aquellas derivadas de la matriz de confusión, la cual se explica con más detalle en la 3.2. En concreto, en este trabajo se ha utilizado la precisión del modelo.

En resumen, la regresión logística será una herramienta esencial en nuestro estudio, permitiéndonos comprender y predecir la relación entre nuestras variables de interés.

3.1.3 Random Forest

Por otro lado, exploraremos la capacidad del método de ML Random Forest, para poder predecir el resultado final de las carreras. Este “Bosque Aleatorio", es un algoritmo de aprendizaje supervisado utilizado en el campo del Machine Learning (ML) para clasificación y regresión. Se caracteriza por su capacidad para manejar una gran cantidad de datos con una alta dimensión de características, ofreciendo resultados precisos y robustos. Su concepto y su funcionamiento es el siguiente:

- **Ensamble de Árboles de Decisión**: Random Forest es un algoritmo de ensamble. Esto significa que combina múltiples árboles de decisión para obtener un resultado más preciso y estable que un solo árbol de decisión.
- **Selección Aleatoria de Datos y Características:** Para cada árbol individual, el algoritmo selecciona aleatoriamente una muestra de los datos de entrenamiento (con reemplazo, conocido como bootstrap sampling). Además, en cada división de un árbol, selecciona aleatoriamente un subconjunto de las características en lugar de usar todas las características disponibles. Esta aleatoriedad ayuda a hacer que el modelo sea más robusto contra el sobreajuste.
- **Construcción de Árboles:** Cada árbol se construye de manera que crezca al máximo sin poda, lo cual es diferente de los árboles de decisión individuales donde la poda es común.

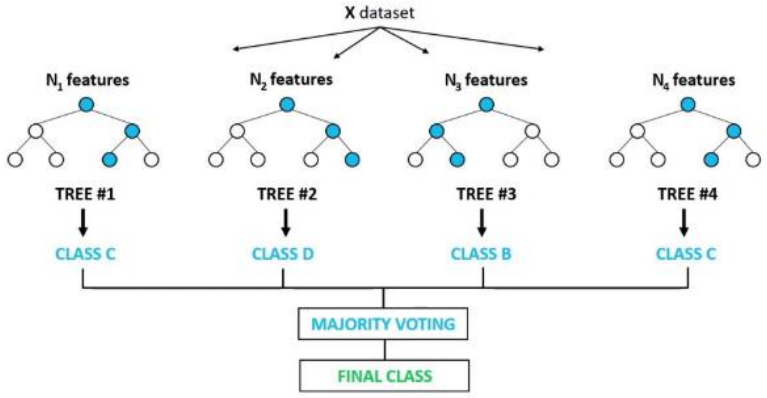

Figura 7: Esquema de la técnica Random Forest con sus arboles de decisión. Fuente: freecodecamp.org

En la fase de predicción, Random Forest agrega las predicciones de cada árbol individual. Para tareas de clasificación, este proceso suele involucrar un sistema de votación mayoritaria, donde la clase que recibe más votos de los árboles individuales es la predicción final del modelo. En tareas de regresión, se promedian los resultados de los árboles.

La eficacia de este método se debe a su capacidad para capturar complejidades en los datos al utilizar múltiples árboles y, al mismo tiempo, mantener un nivel controlado de varianza y sesgo, gracias a la aleatoriedad introducida en el proceso de construcción del modelo. Además, es útil para tratar con conjuntos de datos de gran

tamaño y con un alto número de características, y es relativamente insensible a datos atípicos y a la falta de escalado de las características.

En resumen, esta es una técnica potente y versátil en el aprendizaje automático que combina la simplicidad de los árboles de decisión con la robustez de los métodos de ensemble, siendo capaz de proporcionar predicciones precisas y fiables.

3.2 Métricas de calidad

Evaluar el rendimiento de cualquier modelo de predicción es fundamental para comprender su eficacia en tareas del mundo real. En este estudio, utilizaremos ciertas métricas para juzgar cuán precisas son nuestras predicciones en relación con los datos reales.

- **Matriz de Confusión**: una herramienta esencial para evaluar el rendimiento de un modelo de clasificación. La matriz de confusión nos brinda una visión detallada del rendimiento del modelo al mostrar cuántas predicciones reales y erróneas se hicieron para cada clase. Esto es especialmente útil para identificar áreas específicas donde el modelo podría necesitar mejoras.

  La matriz de confusión proporciona una representación visual de cómo se desempeñaron las predicciones del modelo en relación con los valores verdaderos. Se puede interpretar como:

  - Verdaderos Positivos (TP): Es el valor en la esquina inferior derecha. Representa las instancias donde el modelo predijo correctamente que un piloto estaría en el podio y efectivamente estuvo en el podio.
  - Verdaderos Negativos (TN): Es el valor en la esquina superior izquierda. Indica las instancias donde el modelo predijo correctamente que un piloto NO estaría en el podio y efectivamente no estuvo en el podio.
  - Falsos Positivos (FP): Es el valor en la esquina superior derecha. Se refiere a las instancias donde el modelo predijo que un piloto estaría en el podio, pero no lo estuvo.
  - Falsos Negativos (FN): Es el valor en la esquina inferior izquierda. Representa las instancias donde el modelo predijo que un piloto NO estaría en el podio, pero efectivamente estuvo en el podio.

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (TP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (TN)

Tabla 1:Matriz de Confusión. Elaboración propia

- **Accuracy (Precisión global):** Esta métrica proporciona una visión general de la eficacia del modelo al medir la proporción de predicciones correctas (tanto positivas como negativas) en relación con el total de casos evaluados. En el contexto de tu ejemplo de carreras, donde se predice quiénes llegarán al podio, la exactitud mide cuán frecuentemente el modelo acierta tanto en los pilotos que llegan al podio como en los que no. Se calcula con la fórmula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision (Precisión)**: esta métrica se emplea para determinar la relación entre las predicciones correctas y el total de predicciones hechas. La precisión es útil para tener una visión general de cómo se está desempeñando el modelo en su totalidad. Dado que en una carrera típica participan alrededor de 20 pilotos y solo 3 logran el podio, acertar en las predicciones sobre los 17 pilotos que no alcanzan el podio es relativamente sencillo. Por ello, para obtener una medida de precisión más representativa, nos centraremos en cuántas veces el modelo predice correctamente que un piloto alcanzará el podio y este efectivamente lo logra. Esta aproximación ofrece una visión más ajustada del verdadero desempeño del modelo en escenarios prácticos.

$$Precisión = \frac{TP}{TP + TN}$$

- **Recall (Sensibilidad)**: Esta métrica se refiere a la capacidad del modelo para identificar todos los casos relevantes dentro del conjunto de datos. En este contexto, se refiere a la proporción de pilotos que realmente alcanzaron el podio y que fueron correctamente identificados por el modelo. Es especialmente útil para comprender la efectividad del modelo en la identificación de verdaderos positivos. Una sensibilidad alta indica que el modelo tiene una baja tasa de falsos negativos.

$$Sensibilidad = \frac{TP}{TP + FN}$$

- **Specificity (Especificidad):** Esta métrica mide la capacidad del modelo para identificar correctamente las instancias negativas. En nuestro caso, representa la proporción de pilotos que no alcanzaron el podio

y que fueron correctamente identificados por el modelo. Una especificidad alta indica que el modelo tiene una baja tasa de falsos positivos. Es especialmente útil cuando se quiere entender la eficacia del modelo en la identificación de verdaderos negativos.

$$Especificidad = \frac{TN}{TN + FP}$$

El uso de métricas de calidad adecuadas es fundamental para asegurar que el modelo no solo funcione bien teóricamente, sino que también sea aplicable y útil en escenarios prácticos en el mundo real de la Fórmula 1. Estas métricas no solo proporcionan una evaluación teórica del rendimiento, sino que también ofrecen *insights* valiosos sobre su aplicabilidad en situaciones reales. Mediante la comprensión y análisis detallado de la matriz de confusión y sus derivados, estamos mejor preparados para afinar y optimizar nuestro modelo, garantizando predicciones más precisas y respaldando decisiones basadas en datos del mundial.

4 CASO DE ESTUDIO

En este capítulo, llevaremos a cabo un análisis detallado del conjunto de datos recolectado para este estudio de final de carrera, enfocado en las carreras de Fórmula 1 hasta el año 2022.

En el proceso de investigación y análisis de datos para nuestro trabajo, es esencial contar con fuentes confiables y detalladas. Para este propósito, hemos optado por recopilar datos de Ergast, una plataforma que ofrece un registro histórico de datos de carreras de motor. La *Ergast Developer API* es un servicio web experimental que proporciona la base de datos de toda la F1 para fines no comerciales. Esta API provee datos desde el inicio de los campeonatos mundiales en 1950 hasta la actualidad.

En la primera sección, proporcionaremos una revisión exhaustiva de todas las variables incluidas en el conjunto de datos. Se describen características que registran detalles específicos de las carreras de Fórmula 1, desde detalles personales del piloto, pasando por información relativa a su equipo, hasta condiciones de la carrera. Este análisis permitirá comprender mejor el contexto histórico y las diferentes facetas que han influido en los resultados de las carreras a lo largo del tiempo.

En la siguiente sección, visualizaremos y examinaremos de forma más profunda el impacto de las diversas variables sobre los resultados de las carreras. Mediante gráficas y estadísticas, se explorarán temas como la influencia de la posición de salida en el resultado final. Este análisis permitirá sacar conclusiones más tangibles y significativas sobre los factores que pueden determinar el éxito en una carrera de Fórmula 1.

4.1 Análisis descriptivo de los datos

La base de datos empleada en este proyecto consta de varias características que registran detalles específicos de las carreras de Fórmula 1 hasta 2022. Cada registro se centra en una carrera específica de un piloto individual. A continuación, presentamos un análisis detallado de cada una de las columnas o características de los datos numéricos que vamos a utilizar:

- **Temporada**: Esta es la primera característica y, simplemente, representa el año natural en el que tuvo lugar la temporada. Este detalle es crucial para entender el contexto histórico de los resultados de las carreras, ya que la Fórmula 1 ha experimentado muchos cambios a lo largo de los años en términos de tecnología, normativas, equipos y pilotos. Los datos van desde 1983 hasta 2022, abarcando casi cuatro décadas de carreras de Fórmula 1. Las normativas y tecnologías cambian a lo largo de los años. Una temporada reciente podría estar dominada por ciertos equipos debido a avances tecnológicos o ajustes a las reglas.

- **Ronda**: Representa el orden de la carrera en la temporada correspondiente. Algunos circuitos pueden favorecer a ciertos equipos o pilotos debido a las diferencias en el diseño del circuito. Además, los pilotos y los equipos también pueden mejorar a lo largo de la temporada, lo que hace que las carreras posteriores sean potencialmente más competitivas. Las primeras rondas pueden tener resultados más impredecibles ya que los equipos están adaptándose a las nuevas normas y mejoras de la temporada. Las rondas posteriores pueden reflejar una mayor estabilidad en los resultados, a medida que los equipos se familiarizan con el desafío de la temporada. Es interesante observar que la temporada con más carreras ha sido con 21.
- **Fecha de Nacimiento**: Esta es la fecha de nacimiento del piloto. Esta información podría utilizarse para calcular la edad del piloto en el momento de cada carrera, lo que podría influir en factores como la experiencia, la resistencia física y la capacidad de toma de decisiones. Los pilotos más jóvenes pueden tener una mayor resistencia física, pero los pilotos mayores pueden tener más experiencia y habilidades de toma de decisiones desarrolladas.
- **Posición de Salida**: La posición en la que el piloto comienza la carrera. En la Fórmula 1, la posición de salida es un factor muy importante, ya que las posiciones delanteras suelen dar una ventaja significativa. Una buena posición de salida puede ser el resultado de una clasificación exitosa y puede aumentar las posibilidades de un buen resultado en la carrera. Es más fácil mantener la posición que adelantar a otros competidores. Con una posición de salida media de 11,7, queda claro que la posición inicial no suele ser una ayuda para la mayoría de los pilotos.
- **Tiempo**: Este es el tiempo que tarda el piloto en terminar la carrera. Evidentemente, cuanto menor sea este valor, más rápido ha completado el piloto la carrera, lo que generalmente se traduce en una mejor posición final. No obstante, esta característica tiene muchos valores nulos, lo que sugiere que no todos los pilotos terminan cada carrera. Además, cuando los pilotos son doblados, tampoco registran tiempo final. También es un indicador de la eficiencia del vehículo y de la habilidad del piloto. Cabe destacar el Gran Premio de Adelaida de 1991 que registra el tiempo mínimo en nuestros datos con 24 minutos y 35 segundos. La carrera solo duró catorce vueltas antes de ser suspendida debido a las intensas lluvias. A los pilotos se les otorgó la mitad de los puntos, convirtiendo esta carrera en la más corta de la historia de la Fórmula 1 hasta que fue superada por el Gran Premio de Bélgica en 2021. En contraposición, el tiempo máximo corresponde al Gran Premio de Canadá de 2011. Esta carrera se prolongó hasta las 4 horas, 4 minutos y 39 segundos, ya que se contabilizó el tiempo entre la primera bandera roja y la reanudación de la carrera. No obstante, este parámetro no será seleccionado para análisis, ya que determina directamente el resultado de la carrera: los tres tiempos más rápidos corresponden, precisamente, a los tres pilotos en el podio
- **Puntos**: Los puntos que el piloto consigue en una carrera específica. En la Fórmula 1 actual, solo los primeros diez pilotos que terminan la carrera obtienen puntos, siendo mayor la cantidad de puntos

cuanto mejor es la posición final. Los pilotos que ganan más puntos son aquellos que terminan en mejores posiciones. Los pilotos que acumulan más puntos a lo largo de la temporada probablemente tendrán un mejor rendimiento global. La media de 2,4 puntos indica que una cantidad significativa de pilotos no logra anotar puntos en cada carrera, lo que destaca la alta competitividad del deporte. El máximo valor es 50 puntos, debido a que excepcionalmente, en la temporada 2014 la última carrera se concedían el doble de puntos. Sin embargo, este parámetro tampoco será seleccionado para el análisis, ya que directamente indica el éxito en una carrera y, por ende, su correlación con estar en el podio es directa.

- **Posición Final**: La posición en la que el piloto termina la carrera. Junto con los puntos, esta es una de las medidas más directas del éxito en una carrera. Este es el resultado final de todos los otros factores. Los pilotos que regularmente terminan en posiciones más altas probablemente sean más exitosos en general. Al igual que con la posición de salida, la media de 11,7 para la posición final resalta la gran variabilidad en los resultados de las carreras. Sin embargo, será esencial para etiquetar a los pilotos que alcanzan el podio, asignando un "1" a aquellos que lo logren, permitiendo así entrenar al modelo de forma adecuada.

- **Lat y Long**: Estas características representan la ubicación geográfica del circuito. La ubicación puede tener un impacto significativo en las condiciones de la carrera, incluyendo el clima, la altitud y la atmósfera local, todo lo cual puede afectar a los coches y a los pilotos de diferentes maneras. El clima y las condiciones de la pista pueden variar según la ubicación, lo que puede favorecer a ciertos equipos o pilotos.

- **Mes de la carrera**: La fecha en la que se celebra la carrera. Al igual que la ubicación, la fecha puede influir en las condiciones de la carrera, particularmente en términos de clima. Las condiciones climáticas pueden variar según la época del año, lo que puede tener un impacto en el rendimiento del piloto y del vehículo.

- **Puntos del Piloto:** Esta es la suma total de puntos que el piloto ha acumulado en la temporada antes de la carrera en cuestión. Una puntuación más alta aquí indica un rendimiento más fuerte y más consistente a lo largo de la temporada. Un piloto con más puntos puede tener una mayor confianza o impulso, lo que puede llevar a un mejor rendimiento. El recuento máximo de puntos es de 387, por Lewis Hamilton en 2019, lo que destaca la enorme superioridad de su rendimiento y del coche respecto a los demás competidores.

- **Victorias del Piloto:** Similar a los puntos del piloto, este es el total de victorias que el piloto ha conseguido en la temporada antes de la carrera en cuestión. Una mayor cantidad de victorias sugiere que el piloto es muy competitivo y capaz de ganar carreras. Además, un piloto que ha ganado muchas carreras puede tener una moral alta y una mayor confianza en su capacidad para ganar. Michael

Schumacher consiguió 13 victorias en la temporada en 2004, lo que también demuestra su enorme diferencia respecto a los rivales junto a su equipo.

- **Posición de Clasificación**: Esta es la posición en la clasificación de la temporada con la que el piloto llega al gran premio. Una posición más alta en la clasificación sugiere que el piloto ha tenido un buen rendimiento en la temporada hasta ese momento. Los pilotos con una mejor posición de clasificación suelen tener un mejor rendimiento, ya que demuestra su capacidad para competir al más alto nivel. El máximo es 30, lo que indica que el máximo de participantes en una misma temporada ha sido tal número.
- **Puntos del Constructor:** Este es el total de puntos que el equipo del piloto ha acumulado en la temporada antes de la carrera en cuestión. Al igual que con los puntos del piloto, un recuento de puntos más alto sugiere que el equipo ha tenido un rendimiento fuerte y consistente a lo largo de la temporada. Un equipo con más puntos probablemente tenga un mejor coche y/o un mejor equipo de apoyo, lo que puede resultar en un mejor rendimiento del piloto. El recuento máximo de puntos es de 722, por el equipo Mercedes en 2016, lo que destaca su superioridad respecto a los demás equipos.
- **Victorias del Constructor**: Similar a las victorias del piloto, este es el total de victorias que el equipo del piloto ha logrado en la temporada antes de la carrera en cuestión. Un número mayor de victorias sugiere que el equipo es muy competitivo y capaz de ganar carreras. Un equipo que ha ganado muchas carreras probablemente tenga una alta moral y una mayor confianza en su capacidad para ganar. Al igual que los puntos, en el año 2016, el equipo Mercedes alcanzaron 18 victorias, el récord hasta ahora.
- **Posición de Clasificación del Constructor:** Esta es la posición en la clasificación de la temporada con la que el equipo llega al gran premio. Una posición más alta en la clasificación sugiere que el equipo ha tenido un buen rendimiento en la temporada hasta ese momento. Los equipos con una mejor posición de clasificación suelen tener un mejor rendimiento, ya que demuestra su capacidad para competir al más alto nivel.
- **Tiempo de Clasificación:** Este es el mejor tiempo que el piloto ha registrado en la sesión de clasificación para la carrera. Un tiempo más rápido en la clasificación suele resultar en una mejor posición de salida. Los pilotos que logran tiempos más rápidos en la clasificación suelen tener un mejor rendimiento en la carrera, ya que demuestra su capacidad para manejar el coche y el circuito eficientemente. Los registros de la base de datos han sido escogido en base a esta variable ya que los tiempos se registran desde 1983, y siendo un indicador tan importante de velocidad en cada circuito y gran premio, se han eliminado los registros anteriores a este año para tener en cuenta siempre esta variable.
- **Porcentaje de Victorias**: Este es el porcentaje de carreras que el piloto ha ganado en la temporada hasta el momento. Un porcentaje más alto indica un mayor nivel de éxito y competitividad. Un piloto con un

alto porcentaje de victorias probablemente tenga una gran habilidad y confianza, lo que puede llevar a un mejor rendimiento.

- **Porcentaje de Victorias del Constructor:** Similar al porcentaje de victorias del piloto, este es el porcentaje de carreras que el equipo del piloto ha ganado en la temporada hasta el momento. Un porcentaje más alto indica un mayor nivel de éxito y competitividad. Un equipo con un alto porcentaje de victorias es probable que tenga una gran capacidad de mejora y preparación, lo que puede llevar a mejores resultados.

- **Edad**: La edad del piloto puede influir en varios factores, incluyendo la experiencia y la resistencia física. Los pilotos más jóvenes pueden tener mayor agilidad y resistencia, pero los pilotos mayores pueden tener más experiencia y habilidades de toma de decisiones. Esto puede influir su rendimiento en la carrera.

- **Diff_tiempo_clasificacion:** Esta variable representa la diferencia porcentual entre el tiempo que un piloto registró en la sesión de clasificación y el tiempo más rápido registrado en esa misma sesión de clasificación. Similar al anterior, pero en este caso, se considera el tiempo de clasificación. Para cada sesión de clasificación, se identifica el tiempo mínimo registrado. Luego, para cada piloto, se calcula la diferencia entre su tiempo de clasificación y este tiempo mínimo de clasificación. Esta diferencia se convierte a un porcentaje. Un valor de 0% indica que el piloto tuvo el tiempo más rápido en esa sesión de clasificación. Un valor positivo, por ejemplo, 3%, indica que el piloto fue un 3% más lento que el tiempo más rápido de la clasificación. La fórmula para calcular esta variable es:

$$\text{diff\_tiempo\_clasificacion} = \frac{\text{tiempo\_clasificacion} - \textit{min\_tiempo\_clasificacion}}{\text{min\_tiempo\_clasificacion}} \times 100$$

Donde “min_tiempo_clasificacion” es el tiempo de clasificación más rápido de cada ronda y temporada. Esta representación normalizada permite una comparación más equitativa del rendimiento de los pilotos a lo largo de diferentes temporadas y circuitos, teniendo en cuenta la evolución de la tecnología y las variaciones en los circuitos.

Columna	*Significado*	*Unidades*	*Min.*	*Máx.*	*Media*	*Recuento*
temporada	Año natural de la temporada	Año	**1983**	**2022**	**2002.16**	**16207**
ronda	Orden de la carrera en la temporada	Número	**1**	**22**	**9,3**	**16207**
posicion_salida	Posición en la que sale el piloto	Número	**1**	**30**	**12.3**	**16207**
tiempo	Tiempo en terminar la carrera	Milisegundos	**207071**	**1.47e7**	**5.81e06**	**5329**
puntos	Puntos que consigue en la carrera	Número	**0**	**50**	**2,4**	**16207**
posicion_final	Posición en la que termina el piloto	Número	**1**	**30**	**12.4**	**16207**
lat	Latitud de la posición geográfica del circuito	Número	**-37,8**	**52,8**	**33,2**	**16207**
long	Longitud de la posición geográfica del circuito	Número	**-118,2**	**145**	**16,8**	**16207**
puntos_piloto	Total de puntos en la temporada con los que llega el piloto al gran premio	Número	**0**	**429**	**22**	**15502**
victorias_piloto	Total de victorias en la temporada con los que llega el piloto al gran premio	Número	**0**	**14**	**0,4**	**15502**
posicion_clasificacion	Posición en la temporada con los que llega el piloto al gran premio	Número	**1**	**30**	**12,6**	**15502**
puntos_constructor	Total de puntos en la temporada con los que llega el equipo al gran premio	Número	**0**	**722**	**43**	**16026**
victorias_constructor	Total de victorias en la temporada con los que llega el equipo al gran premio	Número	**0**	**18**	**0,7**	**16026**
posicion_clasificacion_constructor	Posición en la temporada con los que llega el equipo al gran premio	Número	**1**	**20**	**6,6**	**16026**
tiempo_clasificacion	Mejor tiempo anotado en la sesión de clasificación	Segundos	**53.377**	**1002.640**	**88.8670**	**15317**
porcentaje_victorias	Porcentaje de la temporada con los que llega el piloto al gran premio	Porcentaje	**0**	**85,7**	**3,6**	**15502**
porcentaje_victorias_constructor	Porcentaje en la temporada con los que llega el equipo al gran premio	Porcentaje	**0**	**91,7**	**7**	**16026**
mes_carrera	Mes en el que se disputa la carrera	Número	**3**	**12**	**7**	**16207**
edad	Edad del piloto	Número	**17**	**43**	**29**	**16207**
diff_tiempo_clasificacion	Diferencia de tiempo porcentualmente del tiempo de clasificación del más rápido	Número	**0**	**922.9**	**3.5**	**15317**

Tabla 2: Tabla de variables numéricas

En resumen, esta base de datos proporciona una visión detallada de las carreras individuales en la Fórmula 1, con énfasis en el rendimiento de los pilotos y de sus equipos. Cada característica ofrece una perspectiva única sobre los factores que pueden influir en el resultado de una carrera, desde detalles personales del piloto hasta detalles de rendimiento del equipo y condiciones de la carrera. Al combinar todas estas características, es posible realizar análisis y predicciones muy sofisticados sobre los resultados de las carreras de Fórmula 1.

Las variables de la tabla categórica también proporcionan información crucial sobre las carreras de Fórmula 1. Describamos estas variables en detalle:

- **Circuito**: Este identificador único representa el circuito donde se disputa cada carrera. Hay 49 circuitos únicos en estos datos, lo que refleja la diversidad de ubicaciones en las que se corre la Fórmula 1. El circuito de Monza es el más repetido en este conjunto de datos, con 850 carreras disputadas allí, que no

quiere decir Gran Premios. Monza, situado en Italia, es un circuito muy icónico y uno de los más antiguos en el calendario de la Fórmula 1. Su presencia frecuente en estos datos muestra su relevancia en la historia de este deporte.

- **Piloto**: Este campo representa el nombre del piloto que participa en la carrera. Hay 218 pilotos únicos en estos datos, reflejando la amplia gama de talentos que han participado en la Fórmula 1. Kimi Raikkonen es el piloto con más apariciones con 326 entradas, que hasta el 2020 era el piloto con más experiencia. En la actualidad, es ampliamente superado por el español Fernando Alonso con 352 apariciones. Esto no es sorprendente, ya que su talento y habilidad son ampliamente reconocidos.
- **Nacionalidad**: Esta variable indica el país de origen del piloto. Hay 34 nacionalidades diferentes representadas en estos datos. Los pilotos británicos son los más comunes en estos datos, con 2110 entradas. Esto puede reflejar la fuerte tradición del Reino Unido en las carreras de automóviles y su papel clave en la Fórmula 1, tanto en términos de pilotos como de equipos.
- **Constructor**: Esta variable identifica el equipo o constructor para el que compite el piloto. Hay 63 equipos únicos en estos datos. Ferrari es el equipo más repetido en estos datos, con 1298 entradas. Esto refleja la larga historia de Ferrari en la Fórmula 1 y su reputación como uno de los equipos más exitosos del deporte.
- **Carrera**: Este identificador define cómo ha finalizado el piloto la carrera. Puede indicar si el piloto ha terminado la carrera o no. Hay 106 estados únicos en estos datos. El estado más común es "Acabada", con 10027 entradas.
- **País**: Esta variable indica el país donde se disputa la carrera. Hay 29 países únicos representados en estos datos. Italia es el país más común en estos datos, con 1429 carreras. Esto se debe a la presencia de circuitos icónicos como Monza y Imola en dicho país, así como su importancia e impacto en el deporte debido a la presencia de equipos de Fórmula 1 de alto perfil como Ferrari y AlphaTauri.

Columna	*Significado*	*Recuento*	*N.º valores únicos*	*Mas repetido*	*Veces repetidos*
id_circuito	Identificador del circuito	16207	52	monza	915
piloto	Identificador Piloto	16207	241	alonso	352
nacionalidad	País de origen del piloto	16207	35	British	2340
constructor_x	Equipo o constructor con el que compite el piloto	16207	66	williams	1400
Carrera	Estado de finalización de la carrera	16207	2	Acabada	10026
pais	País donde se disputa	16207	31	Italy	1576

Tabla 3: Tabla de variables categóricas

Cada una de estas variables juega un papel crucial en la comprensión de las carreras de Fórmula 1 y tiene un impacto potencial en los resultados de las carreras. Por ejemplo, ciertos pilotos pueden tener un rendimiento mejor en ciertos circuitos o con ciertos equipos. Además, la ubicación de la carrera puede influir en las condiciones de carrera, como el equipo y la configuración del circuito, que pueden tener un impacto significativo en los resultados.

4.2 Análisis a través de gráficas de los datos

En este apartado, se realiza un análisis exhaustivo de diversos aspectos relacionados con la Fórmula 1 desde 1983 hasta 2020. Mediante el uso de gráficas y datos estadísticos, se examina la influencia de la posición de salida en las carreras, la capacidad de los pilotos para superar a sus compañeros de equipo y la frecuencia de los podios alcanzados. Además, se exploran aspectos demográficos como las nacionalidades de los pilotos y la edad en la que se obtienen victorias. A través de estos análisis, se busca comprender mejor los factores que influyen en el rendimiento y el éxito en este deporte de motor.

En primer lugar, el histograma de la Figura 8 representa el porcentaje de victorias obtenidas desde la *pole position* para los 25 circuitos donde más se ha corrido en la historia de la Fórmula 1. La *pole position* se refiere al lugar privilegiado en la parrilla de salida que se otorga al piloto más rápido en la calificación (*poleman*[1]) ha ganado también la carrera. A menudo se considera una ventaja significativa ya que proporciona una pista despejada para el piloto al comienzo de la carrera:

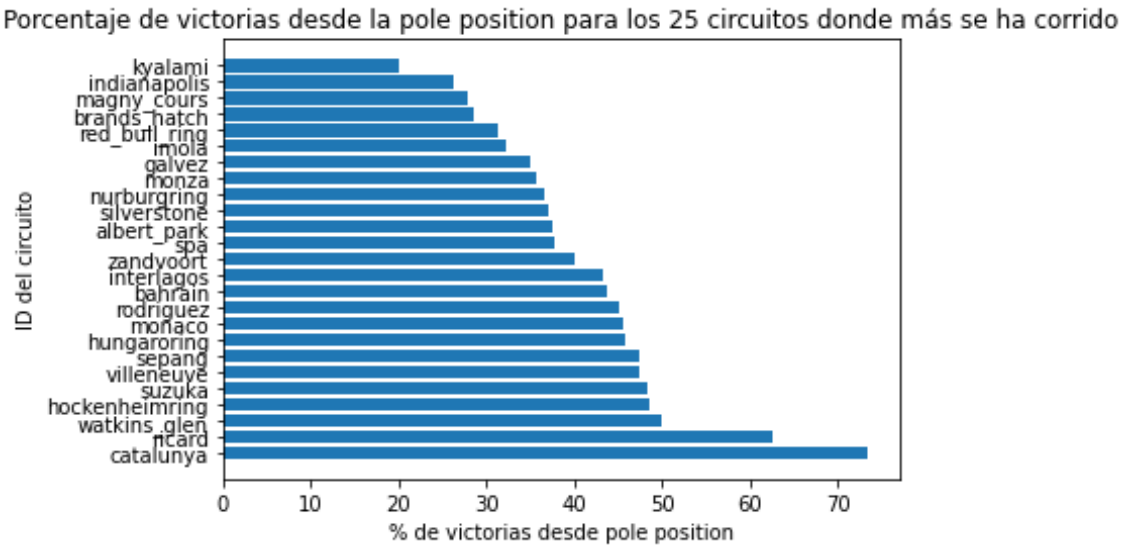

Figura 8: Histograma del porcentaje de victorias desde la primera posición

[1] Piloto que consigue salir de la primera posición en la carrera

Lo que se analiza en esta figura es cuán ventajosa es realmente la primera posición en términos de posibilidades de ganar la carrera en cada uno de estos 25 circuitos. Esto se hace calculando el porcentaje de veces que un piloto que comenzó en la *pole position* terminó ganando la carrera.

De los datos, se puede observar que el circuito de Catalunya encabeza la lista con un impresionante 73.33% de las carreras ganadas por el piloto que comenzó en la pole. Esto podría sugerir que este circuito puede favorecer a los pilotos que empiezan en la *pole position*, quizás debido a las características específicas de la pista o a la dificultad de adelantar en este circuito.

El circuito de Paul Ricard (Francia) sigue en segundo lugar con un porcentaje del 62.5%, mientras que Watkins Glen (E.E.U.U.) muestra un 50% de victorias desde la pole. Algunos circuitos históricos y populares como Suzuka, Villeneuve, Monaco y Sepang también muestran porcentajes cercanos al 50%, lo que indica que salir de la primera posición sigue siendo una ventaja considerable en estos lugares.

Por otro lado, hay circuitos como Kyalami e Indianapolis, donde el porcentaje de victorias desde la *pole* es significativamente más bajo, a 20% y 26.3% respectivamente. Esto podría indicar que estos circuitos pueden permitir más oportunidades para adelantar o que las condiciones de la carrera en estos circuitos pueden ser más impredecibles, reduciendo la ventaja de comenzar en primer lugar.

En general, estos resultados sugieren que, aunque la *pole position* a menudo proporciona una ventaja, la magnitud de esa ventaja puede variar considerablemente dependiendo del circuito. Por lo tanto, aunque obtener la primera posición de salida es un logro importante, no garantiza una victoria y los pilotos y equipos deben considerar una variedad de estrategias y factores en cada circuito para maximizar sus posibilidades de éxito.

La variabilidad en los tiempos de vuelta entre los diferentes circuitos de la Fórmula 1 es una manifestación clara de las características únicas y desafíos que cada pista presenta. En el análisis, centraremos nuestra atención en los tiempos por vuelta y cómo estos varían entre los distintos circuitos. La Figura 9 que presentaremos a continuación muestra la variabilidad de los tiempos en completar una vuelta en diferentes circuitos de la Fórmula 1.

Los diagramas de caja, es una herramienta gráfica que representa la distribución de un conjunto de datos a través de cinco medidas estadísticas: el mínimo, el primer cuartil (o percentil 25), la mediana (o segundo cuartil o percentil 50), el tercer cuartil (o percentil 75) y el máximo. Estas medidas dividen el conjunto de datos en cuatro intervalos que contienen, aproximadamente, el mismo número de datos. Por lo tanto, cada barra del gráfico representa un circuito específico, con la altura de la barra indicando el tiempo medio de vuelta en ese circuito. Los "bigotes" representan los tiempos mínimo y máximo, la "caja" abarca desde el primer hasta el tercer cuartil, y la línea dentro de la caja indica la mediana.

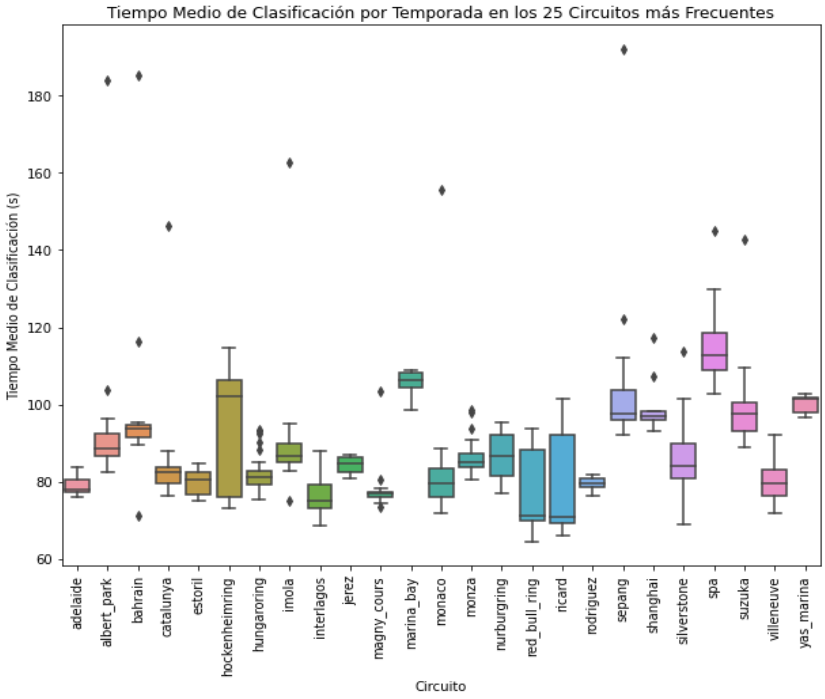

Figura 9: Tiempo medio de clasificación históricamente en cada circuito.

Como se puede observar, el circuito más rápido en promedio es el Red Bull Ring de Austria. Este circuito es un trazado relativamente más corto, rápido y simple en comparación con otros circuitos y esto puede explicar la media de tiempo más reducida. En el extremo opuesto, el circuito más lento en promedio es Spa-Francorchamps (Bélgica). Este circuito es conocido por ser uno de los más largos en el calendario de la Fórmula 1 con más de 7 kilómetros, lo que se refleja en sus tiempos de clasificación más extensos en promedio.

La gráfica ilustra la variabilidad de los tiempos de clasificación por circuito. Es notable que Hockenheimring (Alemania) muestra la mayor variabilidad en los tiempos de clasificación, lo que sugiere que los tiempos en este circuito han experimentado fluctuaciones significativas a lo largo de los años. Estas variaciones pueden estar influenciadas por múltiples factores, como cambios en las condiciones climáticas, evolución tecnológica de los monoplazas o posibles modificaciones en la configuración del trazado.

Por otro lado, Magny-Cours (Francia) presenta la menor variabilidad en tiempos de clasificación. Esto podría estar vinculado al hecho de que este circuito ha sido sede de un Gran Premio en 18 temporadas, desde 1991 hasta 2018, de forma consecutiva. En períodos más cortos y consecutivos, la evolución técnica de los vehículos tiende a ser menos drástica, lo que podría explicar esta consistencia en los tiempos. Adicionalmente, la relativa poca frecuencia con la que se ha corrido en Magny-Cours, junto con las temporadas seguidas en las que ha sido incluido, puede haber contribuido a que las condiciones y características del circuito se mantuvieran consistentes, reflejándose en tiempos de clasificación más estables.

El rendimiento de los pilotos en la Fórmula 1 es un factor esencial en su evaluación. Para ello, es fundamental compararlos con una referencia que comparta condiciones similares, y qué mejor referencia que el compañero de equipo. Ambos pilotos disponen del mismo coche y recursos técnicos, lo que ofrece una base de comparación

objetiva y relevante. La Figura 10 ilustra la proporción de veces que ciertos pilotos han superado a sus compañeros de equipo en las carreras. Se consideraron únicamente aquellos pilotos que han participado en más de 35 carreras para garantizar una muestra significativa de su desempeño a lo largo de su carrera.

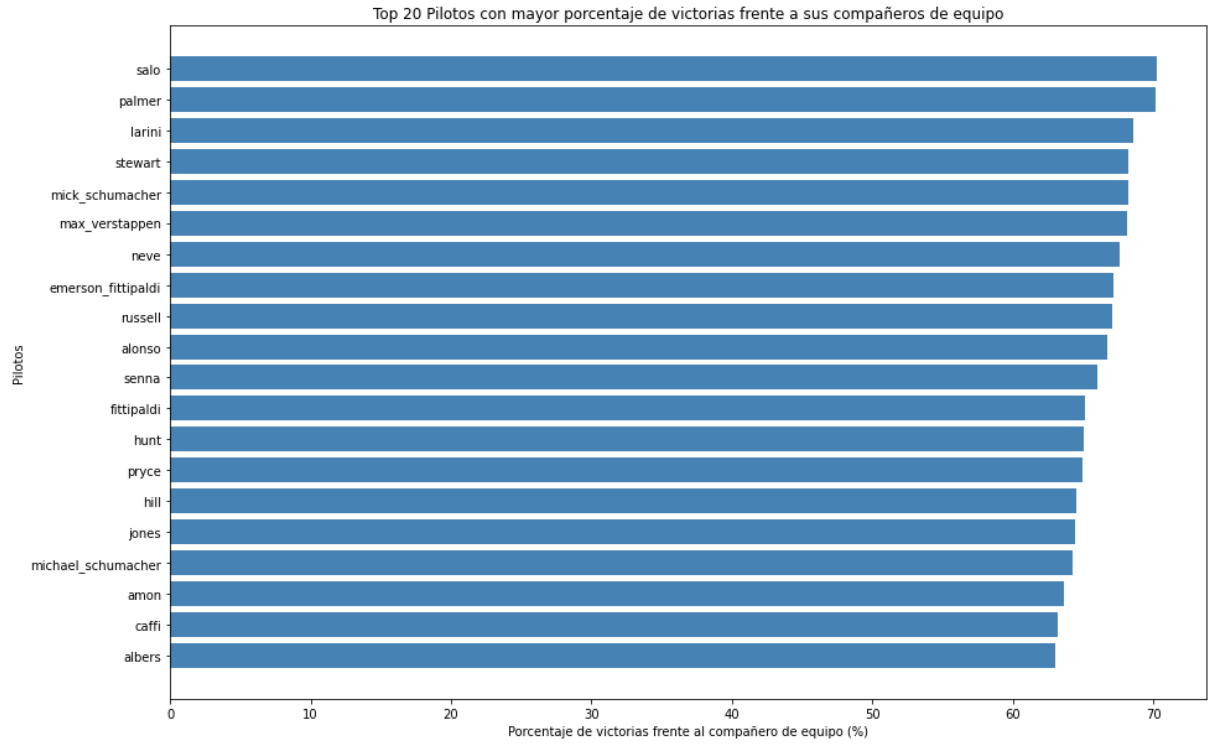

Figura 10: Histograma con los pilotos que más veces han terminado la carrera por delante de su compañero de equipo

Este es un indicador relevante de la habilidad y rendimiento de un piloto, ya que los compañeros de equipo generalmente tienen acceso a los mismos recursos y tecnología dentro de un equipo, lo que proporciona un punto de comparación útil. Mika Salo se destaca en la lista, superando a su compañero en el 70,27% de las carreras entre 1994 y 2002. Esta cifra sugiere una habilidad excepcional y un rendimiento consistentemente superior dentro de sus equipos. Pilotos como Russell, Palmer y Alonso también se distinguen, superando a sus compañeros en más del 69% de las ocasiones. Es relevante mencionar la presencia de distinguidos campeones del mundo como Senna y Michael Schumacher, con un rendimiento que supera el 64%.

Es interesante ver que incluso los pilotos más exitosos y conocidos no siempre superan a sus compañeros de equipo, lo que refuerza la naturaleza competitiva de la Fórmula 1 y el hecho de que un buen rendimiento depende de una variedad de factores.

Por otro lado, estos datos pueden ser un indicativo para los equipos de Fórmula 1 al evaluar el rendimiento de sus pilotos. Un piloto que regularmente supera a su compañero de equipo puede ser visto como un activo valioso para el equipo.

El siguiente histograma revela, en la carrera de un piloto, su proporción de podios en comparación con sus compañeros de equipo a lo largo de sus carreras. Aquí, el "podio" se refiere a los tres primeros puestos en una carrera. La Figura 11 nos proporciona una perspectiva detallada sobre cómo se desempeñan los pilotos en relación con sus rivales más directos.

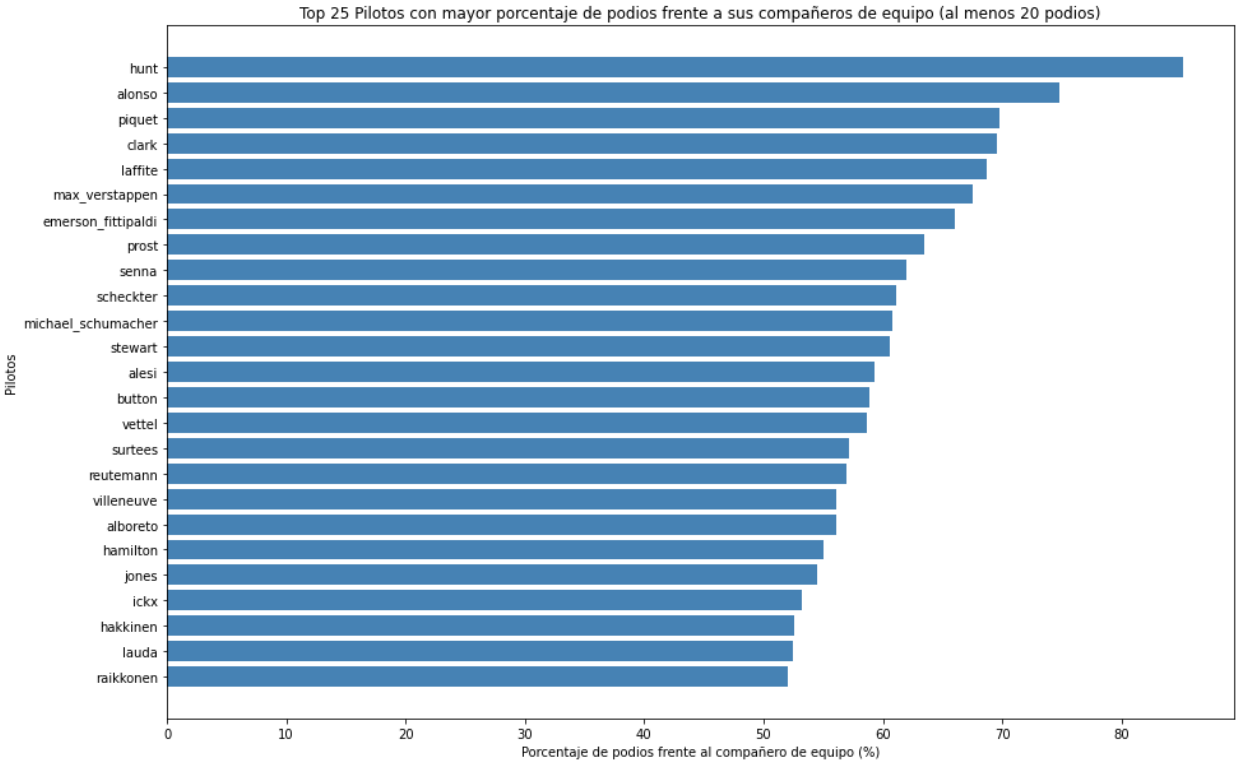

Figura 11: Histograma con el porcentaje de podios con respecto a sus compañeros de equipo de cada temporada

Dado que los pilotos dentro de un mismo equipo tienen acceso a recursos tecnológicos y de ingeniería similares, las diferencias en rendimiento suelen ser una representación directa de habilidades, decisiones estratégicas y, en ocasiones, la gestión de la carrera.

Fernando Alonso destaca en este análisis, en segundo lugar de la lista con un notable porcentaje. Esto significa que, a lo largo de su carrera, ha tenido una frecuencia alta de apariciones en el podio en comparación con sus compañeros de equipo. Otras figuras sobresalientes en el gráfico son James Hunt, Nelson Piquet, Jeremy Clark y Max Verstappen. Estos pilotos han demostrado consistentemente su capacidad para obtener resultados superiores en comparación con sus compañeros de equipo.

Es esencial en la Fórmula 1 lograr posiciones en el podio. Estos lugares no sólo otorgan puntos cruciales para los campeonatos de pilotos y constructores, sino que también garantizan una amplia visibilidad en los medios de comunicación, lo cual es altamente valorado por los equipos y patrocinadores.

Al contrastar estos hallazgos con otros análisis, como la frecuencia con la que los pilotos terminan carreras por delante de sus compañeros de equipo, podemos obtener una percepción más profunda sobre qué pilotos no sólo tienen un rendimiento superior, sino que también son capaces de convertir ese rendimiento en posiciones de podio, que son de suma importancia en el mundo de la Fórmula 1.

En los gráficos que analizamos en la Figura 12, se muestran dos aspectos clave de los pilotos de Fórmula 1, el número de carreras en las que han participado y el número de pilotos por nacionalidad:

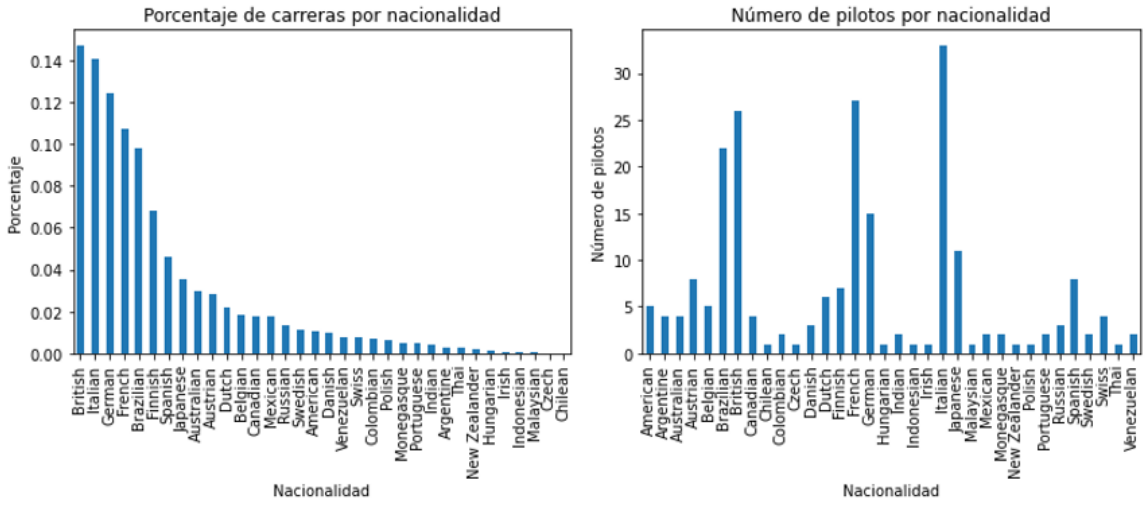

Figura 12: Histogramas con el porcentaje de carreras disputadas por nacionalidad y el número de pilotos que han competido el mundial por país

En primer lugar, destaca la prevalencia de pilotos británicos en el deporte, con un total de 2110 carreras disputadas por 26 pilotos diferentes. Esto supone un promedio de aproximadamente 81 carreras por piloto, lo que evidencia la larga tradición y la continuidad de los pilotos británicos en la Fórmula 1.

En contraposición, los pilotos alemanes, aunque menos en número (15 pilotos), han participado en un total de 1784 carreras, lo que representa un impresionante promedio de casi 119 carreras por piloto. Este dato refleja la duración de las carreras profesionales de muchos pilotos alemanes

Además, se puede apreciar una fuerte presencia de pilotos brasileños, italianos y franceses, con 1401, 2016 y 1533 carreras respectivamente. Sin embargo, el número de pilotos de cada una de estas nacionalidades es bastante diferente, lo que resulta en un promedio de carreras por piloto diverso: aproximadamente 64 para los brasileños, 61 para los italianos y 57 para los franceses.

En resumen, estos dos gráficos proporcionan una visión interesante de la distribución de las nacionalidades en la Fórmula 1, así como la intensidad de la participación de estas nacionalidades a lo largo del tiempo. Muestran claramente cómo algunas nacionalidades tienen una presencia significativa en el deporte, ya sea en términos del número total de carreras disputadas o del número de pilotos que representan a cada país.

El siguiente análisis se centra en los constructores y su impacto en las carreras de Fórmula 1, tanto en términos de longevidad como de éxito. Las siguientes figuras proporcionan una visión detallada de este impacto. La Figura 13 presenta el número de temporadas que cada uno de los 35 principales constructores ha competido en la historia de la Fórmula 1.

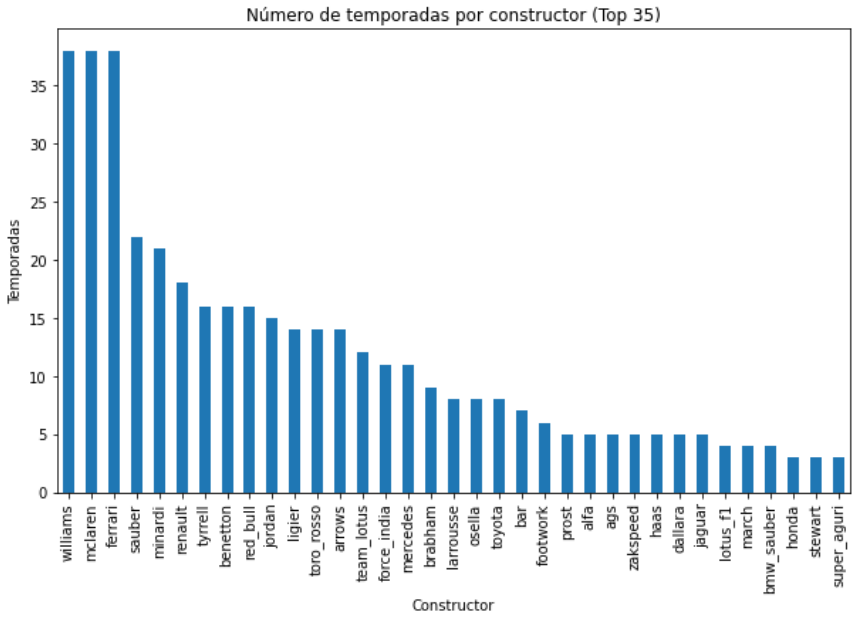

Figura 13: Histograma con el número de temporadas disputadas por constructor

Se puede ver claramente que Williams, McLaren y Ferrari encabezan la lista con 38 temporadas cada uno. Sin embargo, este dato en sí mismo no es suficiente para evaluar la competitividad o el éxito de un equipo. Sauber y Minardi, por ejemplo, han estado presentes en la Fórmula 1 durante varias décadas sin lograr ninguna victoria.

En nuestro próximo análisis, hemos dirigido nuestra atención hacia el éxito en las pistas de carreras. En la Figura 14 medimos el éxito de los constructores de la Fórmula 1 a través del número de victorias que han logrado a lo largo de su historia en el deporte. Este enfoque nos proporciona una visión más profunda de las habilidades técnicas y estratégicas que cada equipo ha demostrado en la competencia.

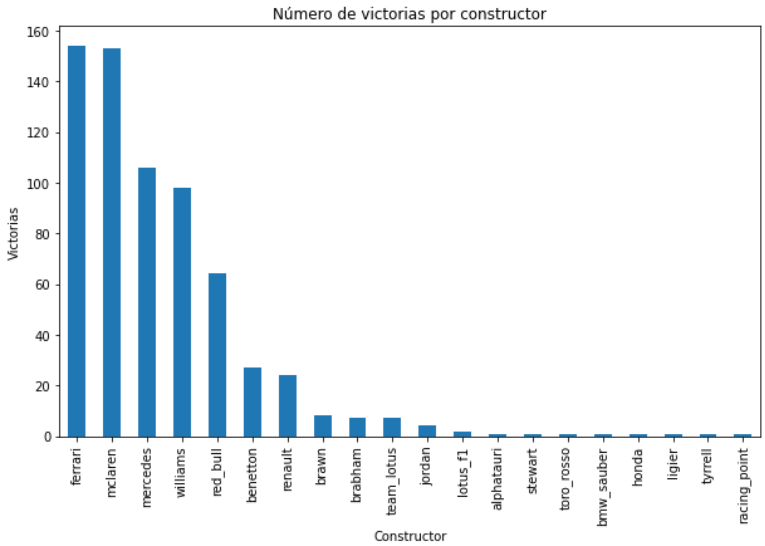

Figura 14: Histograma con el número de victorias por constructor

Para tener una perspectiva más completa, Ferrari encabeza la lista de victorias con 154, seguido de cerca por McLaren con 153. Esto nos dice que tener una larga historia en el deporte no garantiza necesariamente un gran número de victorias. Un claro ejemplo de esto es Mercedes, que con sólo 11 temporadas ha logrado un impresionante total de 106 victorias, lo que indica una alta eficiencia en términos de rendimiento.

En el análisis de los pilotos, la edad es un factor clave. La Figura 15 muestra una visión interesante de cómo la edad de los pilotos influye en su capacidad para obtener victorias. Observamos la distribución de las victorias en relación con la edad de los pilotos, proporcionando una idea del arco de la carrera de un piloto de F1 y cómo su rendimiento puede variar con la edad.

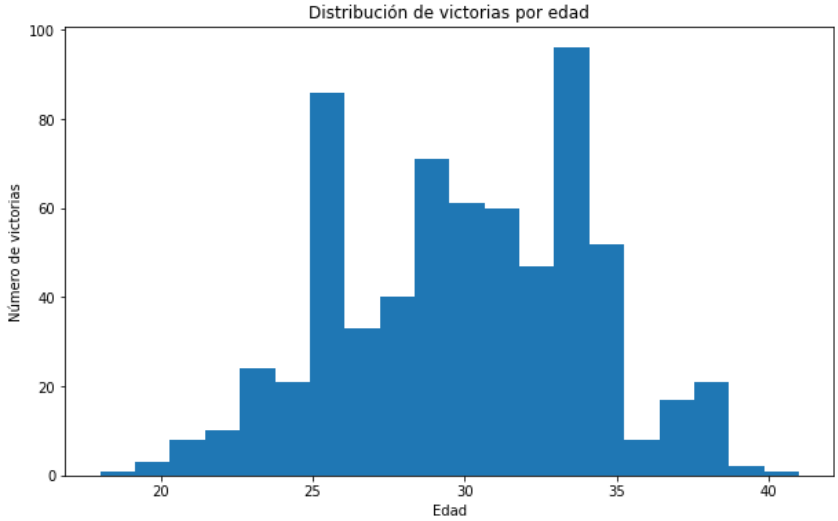

Figura 15: Histograma con la distribución de victorias por edad

Como se muestra en el histograma, la mayoría de las victorias se obtienen cuando los pilotos tienen entre 29 y 35 años. Este período puede considerarse el pico de la carrera de un piloto, donde la habilidad y la experiencia

se combinan de manera óptima. Sin embargo, este gráfico también nos muestra que las victorias son posibles fuera de este rango de edad, habiendo una clara desviación en los 25 años. Esto puede ser debido a la atracción que tienes los equipos ganadores por apostar por pilotos jóvenes que han demostrado en las temporadas anteriores un buen rendimiento.

Por último, nos centramos en la longevidad de las carreras de los pilotos de Fórmula 1. En la Figura 16, analizamos el número de carreras que han competido los pilotos a diferentes edades. Esta evaluación nos ofrece una visión del rango de edades en las que los pilotos están más activos en la F1, y podría arrojar luz sobre el equilibrio entre la experiencia y la forma física en este deporte de alta intensidad.

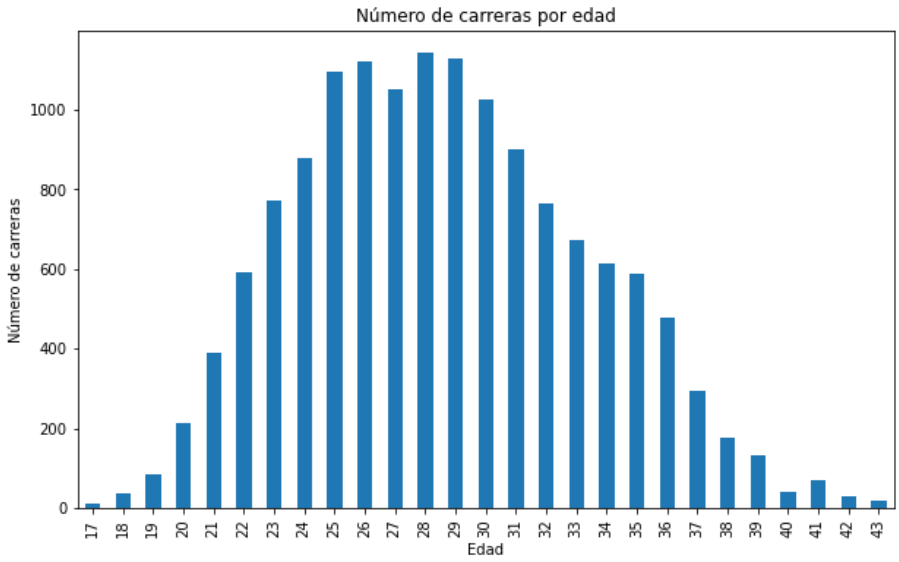

Figura 16: Histograma con el número de carreras disputadas en cada franja de edad

Aquí vemos que la mayoría de las carreras se han corrido entre los 25 y 33 años. Esto sugiere que esta es la franja de edad en la que los pilotos están en su mejor momento físico y mental, y también tienen la experiencia suficiente para competir en la F1. Sin embargo, también vemos que hay un número significativo de carreras que se han corrido con pilotos más jóvenes y veteranos, lo que indica que hay oportunidades para los pilotos en todas las etapas de su carrera.

La Fórmula 1 es un deporte que ha evolucionado drásticamente a lo largo de los años, no solo en términos de velocidad y tecnología, sino también en cuanto a la fiabilidad de los coches y la seguridad. Una manera de evaluar estos cambios es observando el porcentaje de coches que terminan una carrera en comparación con los que empiezan. Un porcentaje más alto de finalización puede indicar una mayor fiabilidad de los vehículos y una mejor seguridad en la pista. A continuación, la Figura 17 presenta la muestra el porcentaje de carreras terminadas por temporada.

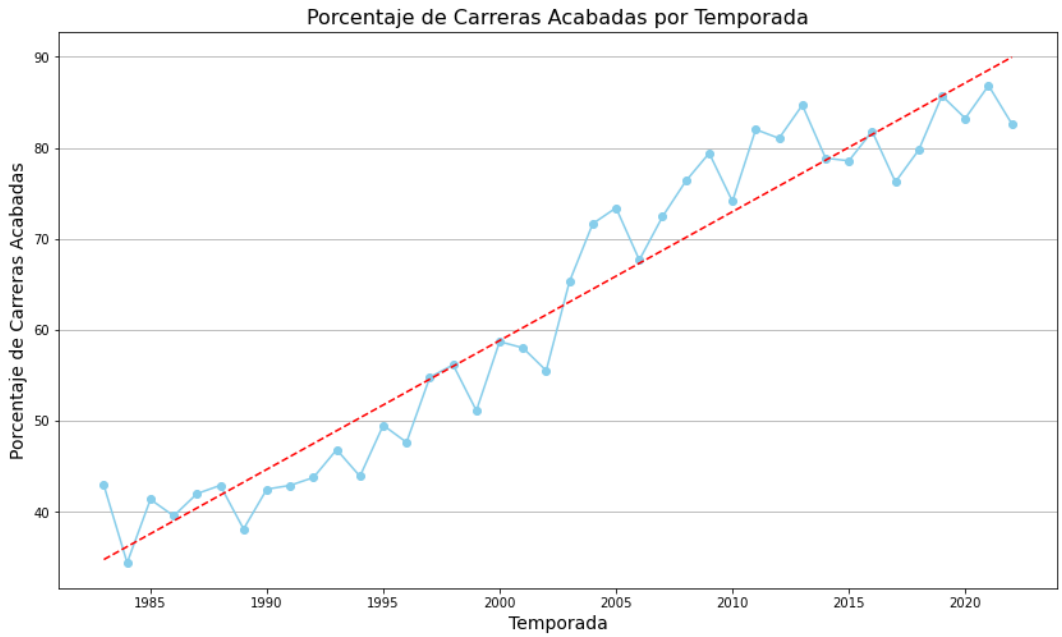

Figura 17: Porcentaje de carreras acabadas por temporada

El gráfico muestra la evolución del porcentaje de coches que terminaron carreras de Fórmula 1 desde 1983 hasta 2022. Es evidente que ha habido un incremento sustancial en la fiabilidad y posiblemente en la seguridad de los coches a lo largo de las décadas. A continuación, se destacan algunos puntos clave basados en estos datos:

- Década de 1980 a principios de 1990: Durante este período, el porcentaje de finalización fluctuó en torno al 40%. Esto indica que aproximadamente 6 de cada 10 coches no terminaban las carreras. Es probable que la combinación de la tecnología menos avanzada y las medidas de seguridad menos rigurosas de la época contribuyeran a estos números.
- Mediados y finales de la década de 1990: En este período, se observa un aumento gradual en el porcentaje de finalización, alcanzando más del 50% hacia finales de la década. Este incremento puede deberse a mejoras en la tecnología de los coches y a medidas de seguridad más estrictas implementadas en el deporte.
- Década de 2000: Esta década vio un aumento aún más significativo en el porcentaje de finalización, comenzando con alrededor del 56% en 1998 y alcanzando un impresionante 76% en 2008. La fiabilidad de los coches definitivamente mejoró durante este período, lo que indica una combinación de avances tecnológicos y posiblemente una mayor atención a la seguridad en la pista.
- Década de 2010 a 2020: El porcentaje de finalización se mantuvo generalmente por encima del 70%, y en algunos años, como 2011, 2013 y 2019, superó el 80%. Esto muestra una notable mejora en la fiabilidad y posiblemente en la seguridad en comparación con las décadas anteriores.

- Años recientes (2020-2022): El porcentaje de finalización sigue siendo alto, oscilando alrededor del 80%. Esto sugiere que la Fórmula 1 ha logrado mantener una alta fiabilidad y seguridad en las carreras recientes.

En resumen, la figura proporciona una visión clara de cómo ha mejorado la fiabilidad de los coches de Fórmula 1 a lo largo de los años. Es evidente que, con el paso del tiempo, un mayor porcentaje de coches ha sido capaz de completar las carreras, lo que es un testimonio de los avances en tecnología y seguridad en el deporte. Estos datos subrayan la importancia de la innovación y el desarrollo en la Fórmula 1, y cómo ha contribuido a hacer del deporte lo que es hoy en día.

A continuación, la Figura 18 representa los tiempos de clasificación por temporada en Monza (Italia). Esta nos ofrece una visión muy interesante sobre la evolución de la Fórmula 1 en este emblemático circuito. Monza, conocido como el 'Templo de la Velocidad', ha sido un escenario clave para mostrar los avances tecnológicos en el deporte.

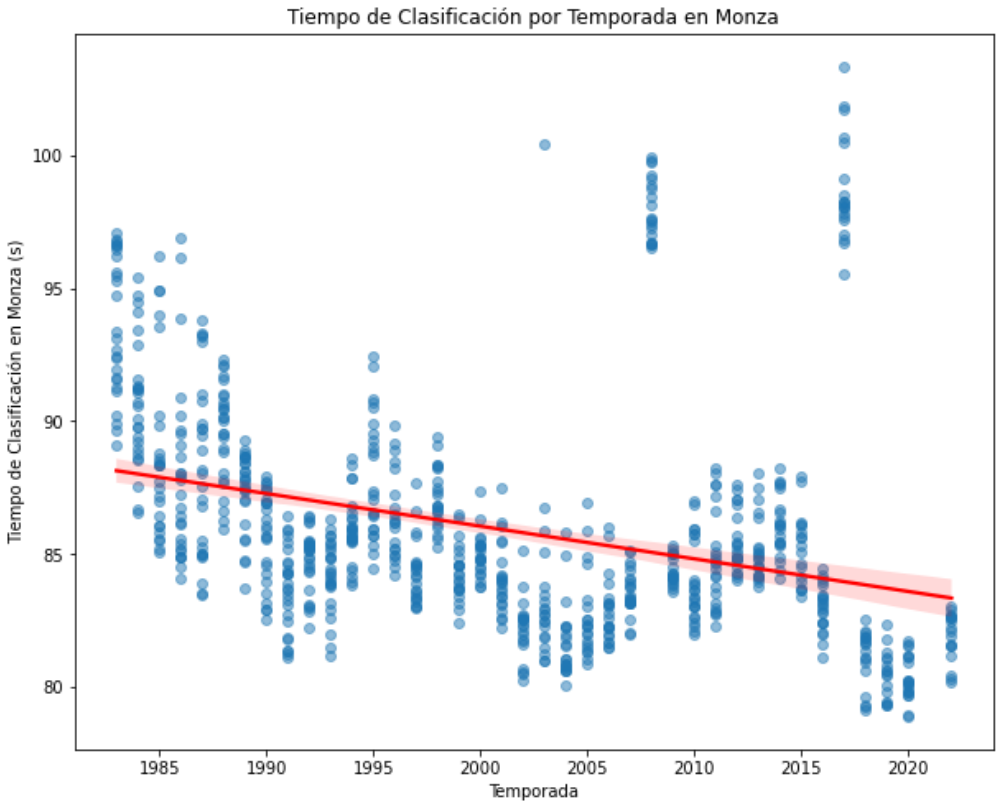

Figura 18: Tiempos de clasificación en el circuito de Monza (Italia)

Como podemos observar, existe una clara tendencia de descenso de los tiempos por temporada. Aunque pueden existir ciertas desviaciones debido a cambios de regulación o factores climáticos. Como punto clave, se debe destacar, en la tendencia decreciente de tiempos, el avance continuo en tecnología, aerodinámica, diseño de coches, y técnicas de conducción. Un tiempo de vuelta más rápido indica que los coches se han vuelto más rápidos y eficientes en términos de rendimiento. La tendencia general hacia tiempos más rápidos subraya la

importancia de la innovación tecnológica en la Fórmula 1. Con cada temporada, los equipos invierten recursos significativos en I+D para mejorar el rendimiento de sus coches. Estos esfuerzos se reflejan claramente en los tiempos de vuelta.

Por último, la Figura 19 es una matriz de correlación, que proporciona el coeficiente de correlación entre cada par de variables en tus datos. El coeficiente de correlación puede variar de -1 a 1.

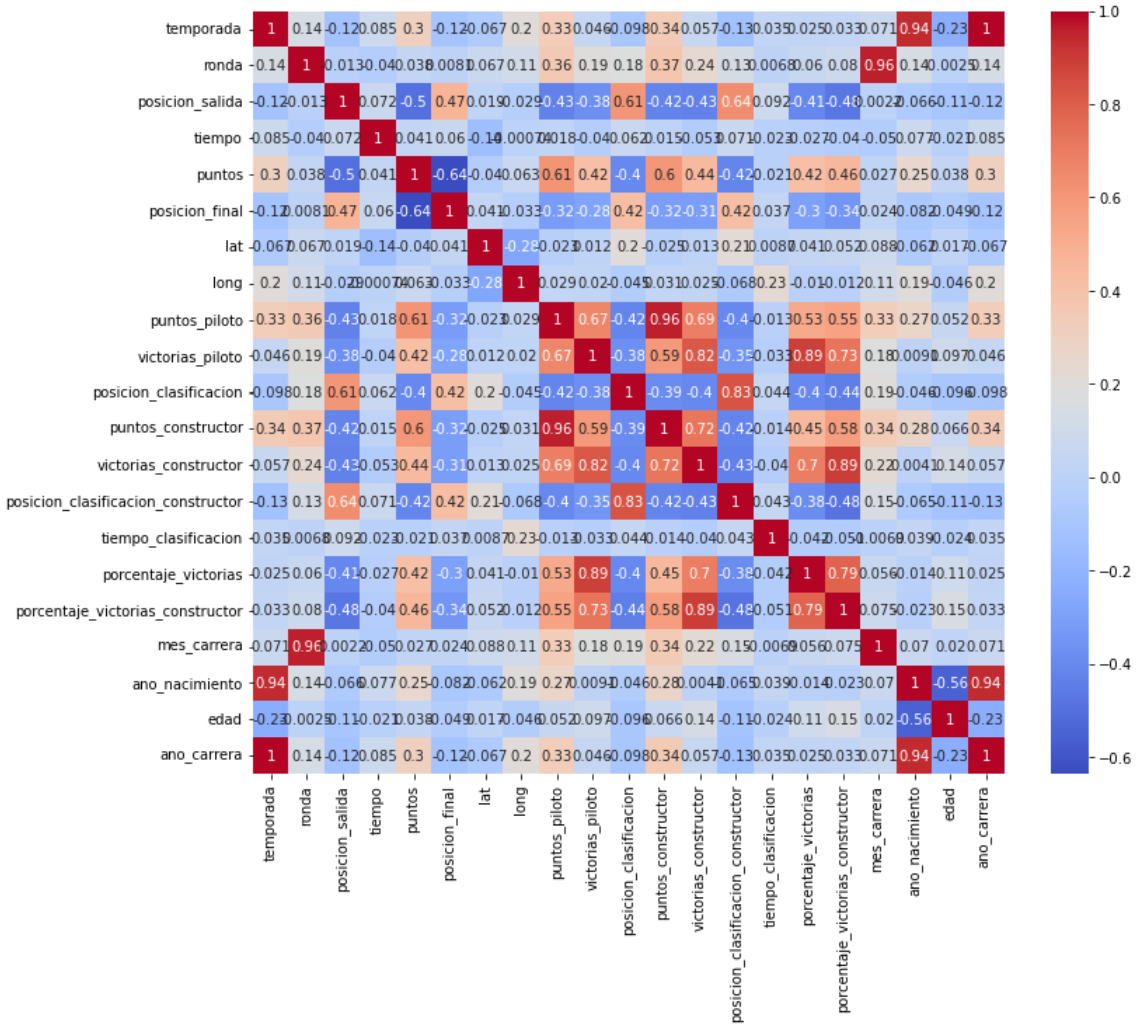

Figura 19: Matriz de correlación entre los datos

Un coeficiente de correlación cercano a 1 indica una fuerte correlación positiva, cuando una variable aumenta, la otra variable tiende a aumentar. Un coeficiente cercano a -1 indica una fuerte correlación negativa, cuando una variable aumenta, la otra variable tiende a disminuir. Un coeficiente cercano a 0 indica que no hay una correlación lineal significativa entre las dos variables.

Es importante destacar que la matriz de correlación solo mide las relaciones lineales entre variables. Esto significa que el coeficiente de correlación proporcionado indica la fuerza y la dirección de una relación lineal entre dos variables. Sin embargo, si las relaciones entre las variables son de otro tipo, como exponenciales, logarítmicas, sigmoidales, u otras formas no lineales, estas no se reflejarán adecuadamente en el coeficiente de

correlación. En estos casos, el coeficiente podría ser cercano a cero, lo que sugeriría una falta de relación lineal, aun cuando podría existir una relación significativa pero no lineal entre las variables.

Podemos destacar algunos resultados interesantes. La posición de salida y los puntos ganados tienen una correlación negativa fuerte (-0.5). Esto puede sugerir que a medida que la posición de salida en una carrera aumenta (por ejemplo, comenzando desde más atrás en la parrilla), los puntos que un piloto obtiene tienden a disminuir. Por otro lado, los puntos del piloto y los puntos del constructor tienen, evidentemente, una correlación positiva muy fuerte (0.96). Esto muestra que los puntos del piloto contribuyen a los puntos del constructor.

De la misma forma también encontramos interesante que la posición de salida y la posición final tienen una correlación positiva fuerte (0.47). Esto indica que a medida que la posición de salida aumenta, es decir, comenzar desde más atrás en la parrilla, la posición final después de la carrera también tiende a ser más alta y por tanto más lejos de los primeros puestos.

4.2.1 Análisis a través de gráficas de dispersión

El gráfico de dispersión es una herramienta visual esencial que muestra la relación entre dos variables numéricas. Cada punto en el gráfico representa una observación y su posición en los ejes X e Y indica los valores de esas dos variables para esa observación específica. A continuación, en la Figura 20 vamos a analizar las variables posición de salida y posición al final de la carrera.

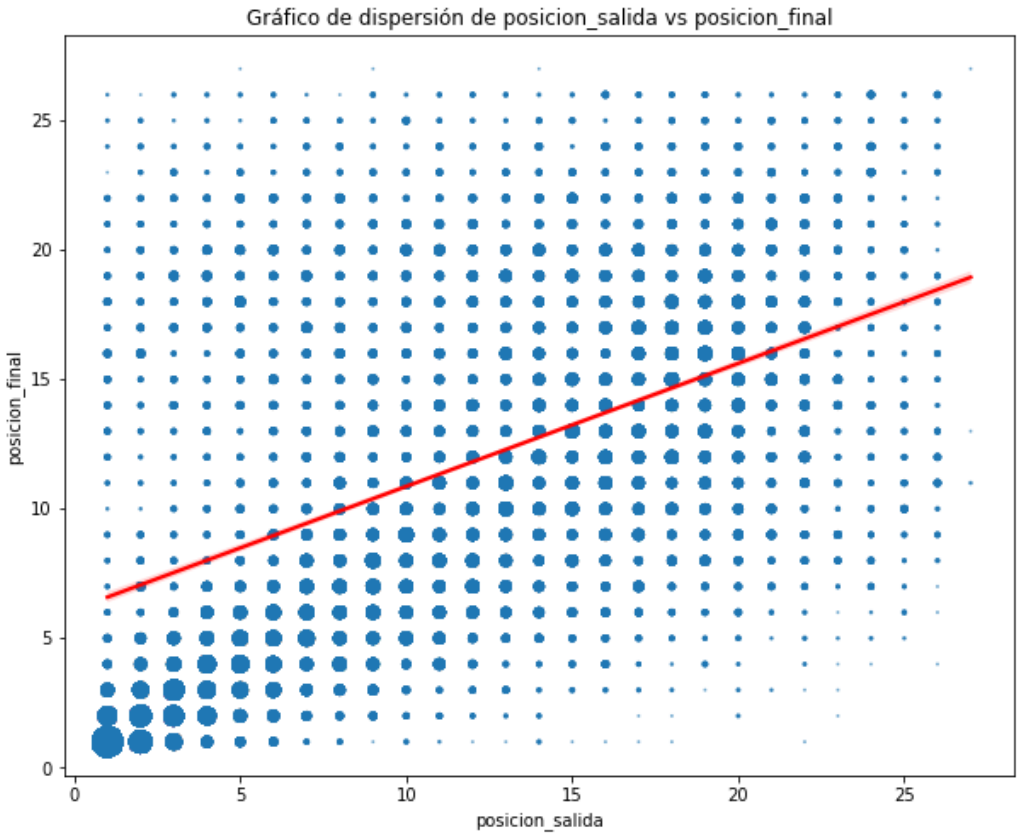

Figura 20: Gráfico de dispersión de posición de salida vs posición final

A simple vista, el gráfico proporciona información sobre cómo la posición de salida de un piloto puede influir en su posición al final de la carrera. Si las posiciones de salida y final fueran completamente independientes, esperaríamos ver puntos distribuidos de manera uniforme por todo el gráfico. Sin embargo, si hay una relación (o correlación) entre las dos, como existe, los puntos se agrupan en un patrón reconocible.

A partir de la información proporcionada, podemos observar que hay una concentración de puntos cerca de una línea diagonal, lo que indica que la posición de salida tiene una fuerte correlación con la posición final. Es decir, un piloto que comienza la carrera desde una posición adelantada tiene más probabilidades de terminar también en una posición adelantada.

Esta correlación tiene sentido desde una perspectiva lógica y estratégica en las carreras. Los pilotos que califican en las posiciones delanteras generalmente tienen tiempos de vuelta más rápidos, lo que indica que tienen una combinación de habilidad, experiencia y un coche bien ajustado. Además, al iniciar la carrera desde una posición delantera, un piloto puede evitar gran parte del tráfico y las complicaciones que suelen ocurrir en el medio o en la parte trasera del grupo.

En conclusión, el gráfico de dispersión revela una relación significativa entre la posición de salida y la posición final en las carreras. Esto subraya la importancia de la sesión de clasificación en las carreras, ya que una buena posición al inicio puede ser un fuerte indicador de éxito en la carrera. Estos hallazgos también pueden ser útiles para los equipos y pilotos al planificar estrategias para las carreras y al establecer expectativas para las carreras basadas en los resultados de la clasificación.

En contraposición, la Figura 21 muestra la relación de la edad con la posición final.

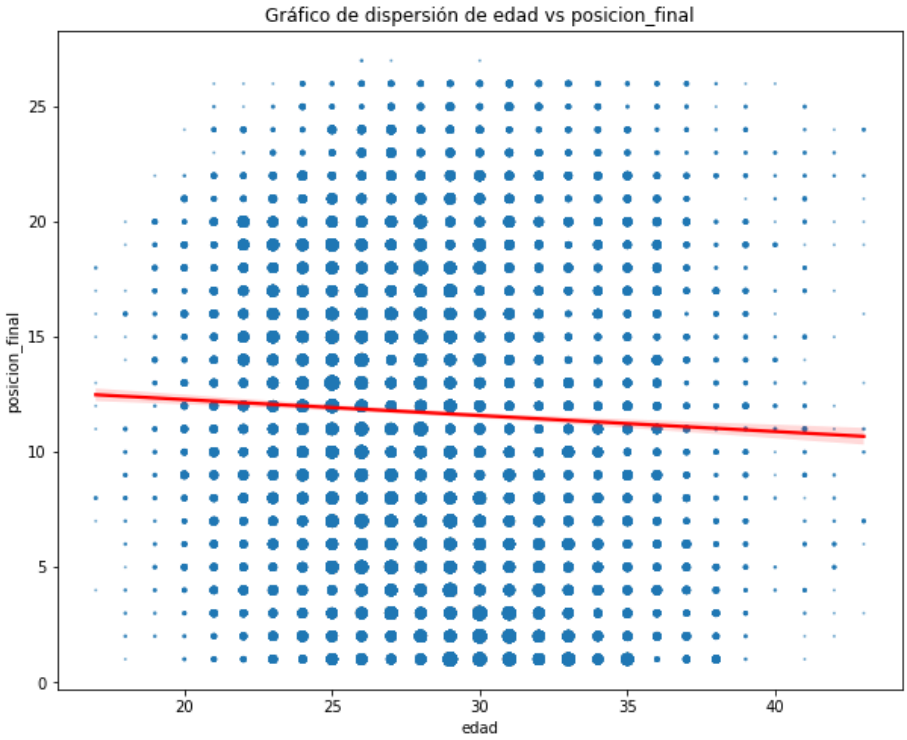

Figura 21: Gráfico de dispersión de edad vs posición final

En este caso, es evidente que no hay una relación clara o tendencia entre estas dos variables. Los puntos están distribuidos de manera bastante uniforme en todo el gráfico, lo que indica que la edad de un piloto no tiene un impacto significativo en su posición final en la carrera. Por lo tanto, podemos concluir que, al menos en este conjunto de datos, la edad no es un factor determinante para el éxito en las carreras.

A continuación, comparando en la Figura 22 la posición de un piloto en la clasificación del mundial con su posición final en una carrera específica, se muestra una clara tendencia. Aquellos pilotos que tienen una posición más alta en la clasificación del mundial tienden a terminar en posiciones más avanzadas en las carreras. Este patrón sugiere que el rendimiento general de un piloto a lo largo de la temporada (reflejado en la clasificación del mundial) es un buen indicador de cómo puede desempeñarse en una carrera individual.

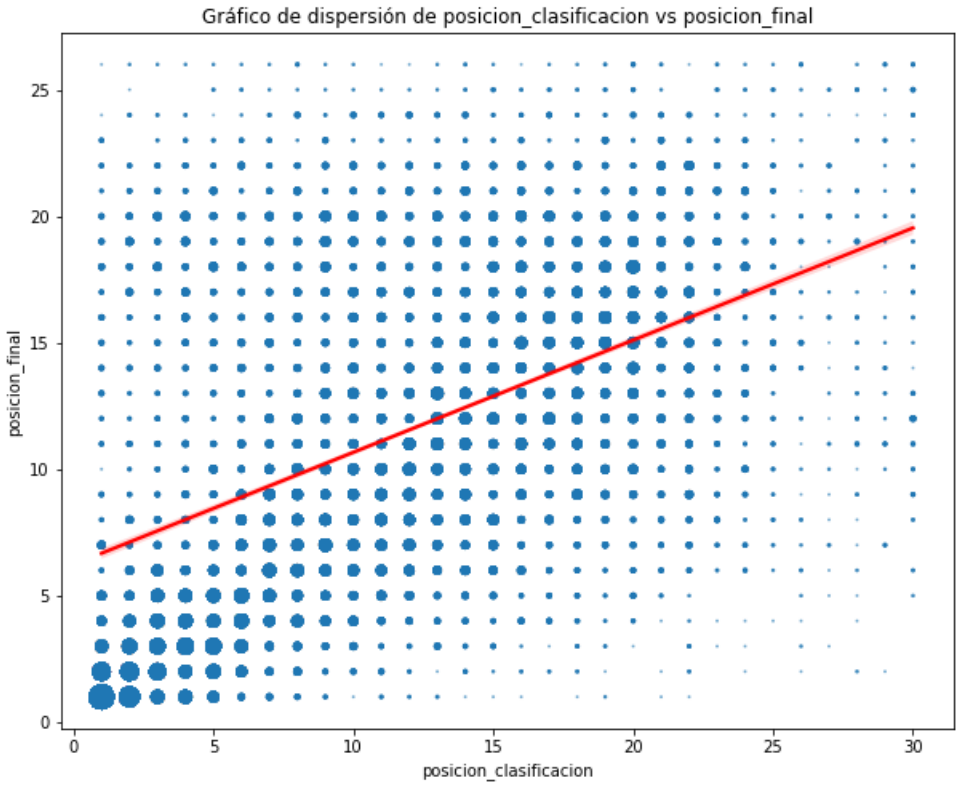

Figura 22: Gráfico de dispersión de posición de clasificación vs posición final

Similarmente, al analizar la Figura 23, que relaciona la posición de un constructor en la clasificación del mundial con la posición final de sus coches en una carrera, también se observa una tendencia definida. Aquellos equipos con una posición más alta en la clasificación tienden a tener mejores resultados en las carreras. Esto indica que el rendimiento global de un equipo en términos de diseño, estrategia y ejecución tiene un fuerte impacto en los resultados individuales de las carreras.

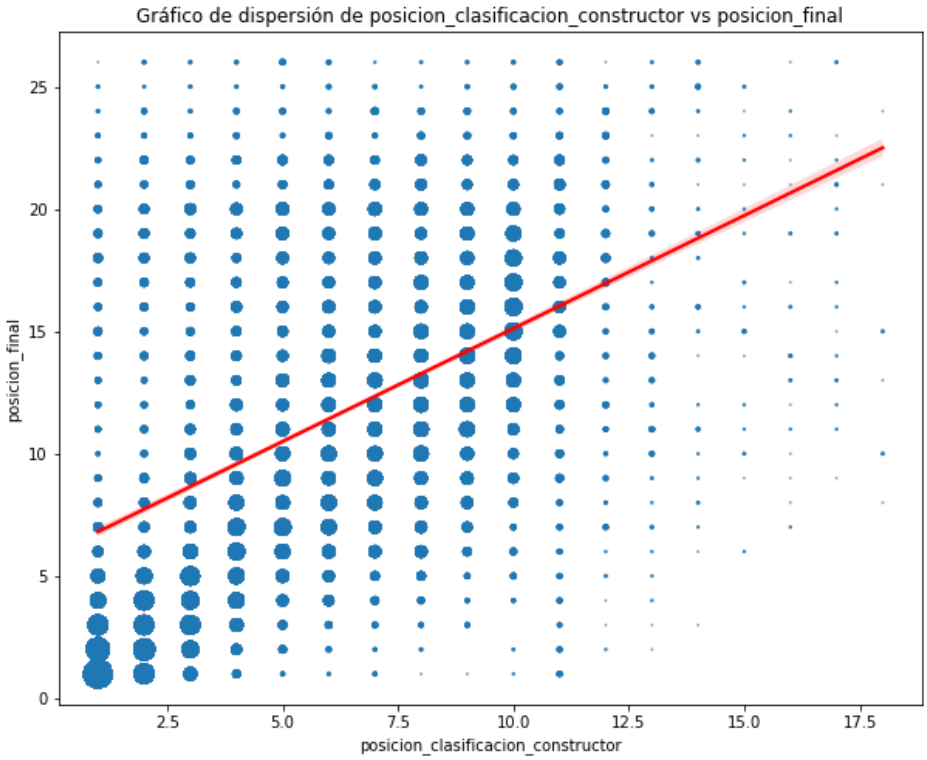

Figura 23: Gráfico de dispersión de posición en el mundial de constructores vs posición final

Estas correlaciones no son sorprendentes. En la Fórmula 1, la consistencia es clave. Tanto pilotos como equipos que demuestran un alto rendimiento a lo largo de la temporada suelen estar mejor preparados, tener acceso a mejor tecnología y estrategias más refinadas. Estas ventajas se traducen en mejores posiciones en las carreras individuales.

En conclusión, existe una fuerte correlación entre la posición en la clasificación del mundial (tanto para pilotos como para constructores) y la posición final en las carreras. Estos gráficos refuerzan la idea de que la Fórmula 1 no solo se trata de habilidades individuales o rendimiento en una sola carrera, sino de consistencia y rendimiento a lo largo de toda la temporada.

La Figura 24 es un diagrama de caja, también conocido como diagrama de caja o *boxplot*. Esta gráfica es una forma estandarizada de representar la distribución de datos basada en un resumen de cinco números ("mínimo", primer cuartil (Q1), mediana, tercer cuartil (Q3) y "máximo").

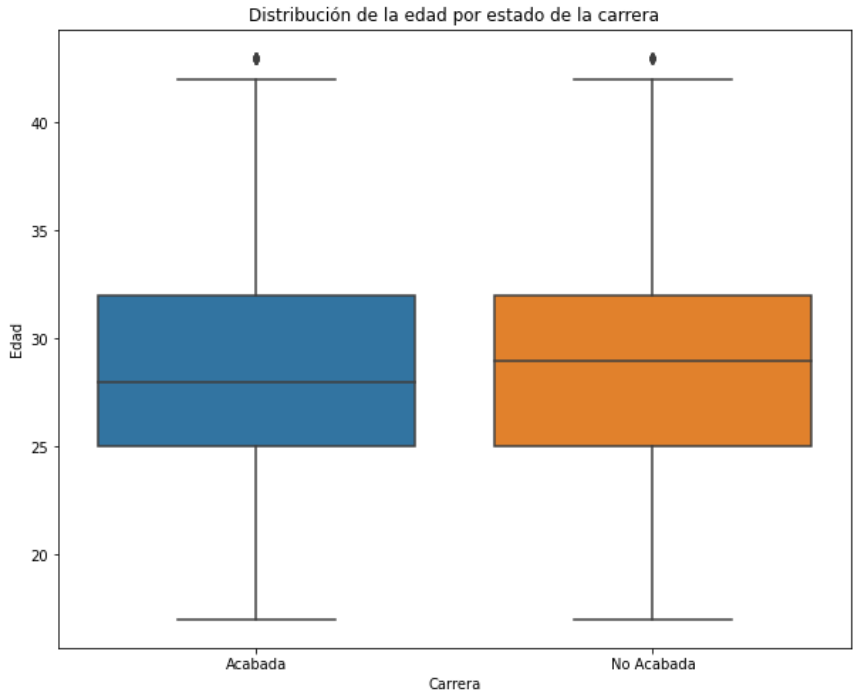

Figura 24:Diagrama de caja de la relación entre la edad y las carreras finalizadas

En este *boxplot*, tienes dos cajas, una para cada categoría de la variable 'Carrera': 'Acabada' y 'No Acabada'. Cada caja muestra cómo se distribuyen las edades de los conductores que acabaron o no acabaron la carrera.

La línea inferior de la caja indica el primer cuartil (Q1), que es el valor por debajo del cual se encuentra el 25% de los datos. La línea del medio de la caja indica la mediana (el segundo cuartil Q2), que es el valor que separa la mitad inferior de la mitad superior de los datos. La línea superior de la caja indica el tercer cuartil (Q3), que es el valor por debajo del cual se encuentra el 75% de los datos.

En este caso, parece que la mediana de la edad de los conductores que no acabaron la carrera es ligeramente más alta que la de los que sí acabaron la carrera. Esto podría sugerir que los conductores mayores tienen una ligera tendencia a no terminar las carreras, aunque la diferencia es mínima. Además, hay algunos valores atípicos en ambos grupos.

5 TRATAMIENTO DE DATOS

El mundo actual y más concretamente el deporte que nos concierna está impregnado de datos. Desde sistemas financieros hasta redes sociales, pasando por la medicina y la ingeniería, cada sector está generando y utilizando vastas cantidades de información. En el apogeo de esta revolución de datos, el Machine Learning (ML) ha surgido como una herramienta poderosa para extraer conocimiento y patrones útiles de estos conjuntos de datos, a menudo vastos y complejos. Sin embargo, la eficacia de estos modelos de ML depende en gran medida de la calidad del material con el que se alimentan: los datos.

En capítulos anteriores, hemos presentado las variables que componen nuestra base de datos y hemos mostrado la cantidad de valores que estas contienen. Como pudimos observar, no siempre estas cifras son consistentes. Esta inconsistencia en la cantidad de datos, así como la presencia de valores faltantes o erróneos, puede parecer a primera vista un pequeño obstáculo. Sin embargo, es un desafío fundamental que puede comprometer la integridad y utilidad de cualquier modelo de ML.

A continuación, explicaremos algunas razones claves de porqué es esencial el tratamiento de datos:

- **Integridad del Modelo:** Los modelos de ML, en su esencia, son algoritmos matemáticos que aprenden patrones a partir de datos. Si estos datos están incompletos, sesgados o contienen errores, el modelo resultante reflejará estos mismos problemas, lo que puede llevar a predicciones y análisis inexactos.
- **Evitar Errores y Fallos:** Algunos algoritmos no pueden manejar valores nulos o faltantes y fallarán al intentar procesarlos. El tratamiento de datos garantiza que el conjunto de datos esté en un formato que cualquier algoritmo pueda manejar sin errores.
- **Mejorar la Precisión:** Incluso si un modelo de ML puede manejar valores faltantes o erróneos, estos pueden reducir la precisión del modelo. Al tratar adecuadamente los datos, se maximiza la precisión y eficacia del modelo.
- **Consistencia y Comparabilidad:** Para comparar modelos o resultados a lo largo del tiempo o entre diferentes conjuntos de datos, es esencial que los datos estén en un formato consistente. Esto garantiza que cualquier variación observada se deba a diferencias reales y no a inconsistencias en los datos.
- **Aumentar la Confiabilidad:** Cuando se toman decisiones basadas en predicciones o análisis de ML, es crucial que se pueda confiar en los resultados. Un conjunto de datos bien tratado y limpio garantiza que los resultados sean confiables y representativos de la realidad subyacente.

El tratamiento de datos no es simplemente una tarea preliminar que se realiza antes del análisis real. Es un proceso integral que requiere una comprensión profunda tanto de los datos como del dominio del problema. A menudo, las decisiones tomadas durante esta fase pueden tener un impacto significativo en los resultados finales y en las conclusiones extraídas.

En este capítulo, exploraremos detalladamente cómo abordar y gestionar los desafíos asociados con datos faltantes, inconsistentes o erróneos. A medida que avanzamos, es vital recordar que cada decisión tomada en esta fase tiene el potencial de influir en el resultado final. Por lo tanto, es esencial abordar esta etapa con cuidado, consideración y, sobre todo, con un profundo conocimiento de los datos y el contexto en el que se utilizan.

Junto con el tratamiento de valores nulos, existe otra consideración crucial en la preparación de datos para Machine Learning, especialmente en el contexto de la predicción: la eliminación o manejo de datos que pueden revelar inadvertidamente la respuesta o resultado que intentamos predecir. Esta consideración es especialmente vital en escenarios donde el objetivo es predecir un resultado futuro basado en datos históricos y actuales.

En nuestro caso específico del deporte motor, hay ciertas variables que, aunque son parte integral de nuestros registros, pueden dar pistas directas sobre el resultado de una carrera. Por ejemplo, las variables como posicion final, tiempo y puntos son directamente indicativas del desempeño de un piloto en una carrera. Si un piloto tiene un tiempo registrado, sugiere que completó la carrera, y si ese tiempo es uno de los tres más rápidos, es probable que haya conseguido un lugar en el podio. Del mismo modo, las puntuaciones más altas generalmente indican un buen desempeño, y las tres puntuaciones más altas generalmente corresponden a los pilotos que alcanzaron el podio.

Incluir estas variables en nuestro modelo sería contraproducente. No solo haría que nuestras predicciones fueran inexactas, sino que también podría llevarnos a una falsa sensación de precisión, ya que estas variables actuarían como "spoilers" de los resultados que estamos tratando de predecir.

Por lo tanto, es esencial identificar y manejar adecuadamente este tipo de variables antes de entrenar nuestros modelos. La eliminación de estas variables "reveladoras" garantiza que nuestro modelo se base en patrones genuinos y características significativas, y no simplemente en datos que revelen directamente el resultado.

Este proceso de identificación y eliminación no es simplemente un acto de censura, sino un paso esencial para garantizar que nuestro modelo de Machine Learning funcione como una herramienta predictiva genuina y no simplemente como un reflejo directo de resultados ya conocidos.

Además de las consideraciones anteriores, es esencial mencionar que no todos los datos recopilados o disponibles son necesariamente útiles o relevantes para el análisis que estamos realizando. En algunos casos, ciertas variables pueden no tener un impacto directo en el resultado que intentamos predecir, o pueden ser redundantes en presencia de otras variables más informativas.

Por ejemplo, en nuestro conjunto de datos, hemos decidido no utilizar ciertas variables que, aunque presentes, no consideramos cruciales para determinar el resultado de una carrera. Estas variables son:

- **Latitud y longitud geográfica**: Estas coordenadas, aunque proporcionan una ubicación precisa, no ofrecen una perspectiva directa sobre el rendimiento de un piloto o un equipo. Además, está completamente correlacionada con la variable “id_circuito”

- **Estado de finalización de carrera**: Si bien esta variable puede indicar si un piloto completó una carrera, puede falsear las predicciones de una carrera ya que es más fácil deducir si un piloto es favorito de terminar en el podio si ha terminado la carrera y no es posible que terminado la carrera si no ha terminado la carrera.
- **Edad del piloto**: La edad, como hemos visto en los análisis anteriores no revela el rendimiento real de los pilotos a priori. Aunque puede influir en la experiencia, no se muestra como un indicador directo del rendimiento en una carrera específica.
- **País de la carrera:** Dado que ya contamos con el nombre del circuito, la inclusión del país también se vuelve redundante.
- **Nacionalidad del piloto**: Similarmente, la nacionalidad del piloto no tiene un impacto directo en su rendimiento en una carrera específica. La valía depende del piloto y dos pilotos con el mismo origen pueden tener resultados muy dispares

En nuestra estrategia de tratamiento de datos, también hemos decidido eliminar la variable 'tiempo_clasificacion'. Aunque esta variable es esencial para entender el rendimiento de un piloto en las sesiones de clasificación, hemos optado por una representación más normalizada y útil de la misma. Hemos introducido la variable '**diff_tiempo_clasificacion'**, que calcula la diferencia porcentual del tiempo de clasificación de un piloto con respecto al tiempo más rápido en esa ronda y circuito.

En resumen, la selección y transformación adecuada de las variables es tan crucial como el tratamiento de valores nulos o erróneos. Cada paso en el proceso de preparación de datos tiene el potencial de impactar en la calidad y precisión de los modelos de Machine Learning que eventualmente entrenaremos.

Las variables con valores nulos en el *dataframe* que finalmente utilizaremos ("df_final") abarcan desde estadísticas de tiempo de carrera, puntos y victorias de pilotos y constructores, hasta clasificaciones en el campeonato mundial. Es importante destacar que la naturaleza de estos valores nulos puede variar. Pueden ser resultado de carreras en las que un piloto o equipo no participó o podrían deberse a fallos en la recopilación de datos o en la transmisión de datos.

La estrategia para rellenar estos valores nulos se basa en la naturaleza y significado de cada variable:

Estrategia	Valores Nulos en Puntos y Victorias	'posicion_salida'	'posicion_clasificacion' y 'posicion_clasificacion_constructor'	'diff_tiempo_clasificacion'
1	Eliminar filas con valores nulos	Eliminar filas con valor 0	Eliminar filas con valores nulos o 0	Eliminar filas con valores nulos
2	Rellenar con 0	Eliminar filas con valor 0	Eliminar filas con valores nulos o 0	Eliminar filas con valores nulos
3	Rellenar con 0	Reemplazar nulos o 0 con máximo valor de la ronda +1	Reemplazar nulos o 0 con máximo valor de la ronda +1	Eliminar filas con valores nulos
4	Rellenar con 0	Reemplazar nulos o 0 con máximo valor de la ronda +1	Reemplazar nulos o 0 con máximo valor de la ronda +1	Eliminar la columna
5	Rellenar con 0	Reemplazar nulos o 0 con máximo valor de la ronda +1	Reemplazar nulos o 0 con máximo valor de la ronda +1	Rellenar con el valor máximo de la ronda y temporada
6	Rellenar con 0	Reemplazar nulos o 0 con máximo valor de la ronda +1	Reemplazar nulos o 0 con máximo valor de la ronda +1	Categorización en rangos
7	Rellenar con 0	Reemplazar nulos o 0 con máximo valor de la ronda +1	Se asigna la última posición registrada históricamente del piloto o constructor. Si no se encuentra se reemplazan los nulos o 0 con máximo valor de la ronda +1	Rellenar con tiempo máximo +1. Se eliminan los registros con puntos muy degenerados, partir del percentil 95.
8	Rellenar con 0	Reemplazar nulos o 0 con máximo valor de la ronda +1	Se asigna la última posición registrada históricamente del piloto o constructor. Si no se encuentra se reemplazan los nulos o 0 con máximo valor de la ronda +1	Rellenar con el valor máximo de la ronda y temporada

Tabla 4: Comparación de las estrategias

5.1 Primera estrategia

La primera estrategia de tratamiento de datos aplicada al *dataframe*, implica una serie de pasos específicos para manejar valores problemáticos en distintas columnas. A continuación, se detalla cada paso de esta estrategia:

- **Tratamiento de Valores Nulos en Puntos y Victorias**: Para las variables de puntos del piloto, victorias del piloto, puntos del constructor, números de victorias del constructor, porcentaje de victorias y porcentaje de victorias del constructor, se asume que la ausencia de información (valores nulos) en estas variables indica la inexistencia de puntos o victorias hasta ese momento. Por ello, se eliminan las filas

con valores nulos en estas columnas, ya que representan registros incompletos o no relevantes para el análisis.

- **Eliminación de Filas con 'posicion_salida' igual a 0**: La posición de salida igual a 0 no tiene sentido práctico en las carreras, ya que no existe tal posición.
- **Eliminación de Filas con 'posicion_clasificacion' o 'posicion_clasificacion_constructor' Nulos o Cero**: Estos valores en estas columnas pueden indicar datos faltantes o incorrectos. No tener una posición de clasificación válida desvirtúa la integridad del registro. Por tanto, se eliminan las filas donde estas columnas son nulas o tienen un valor de cero, asegurando que solo se mantengan registros con información de clasificación válida y completa.
- **Eliminación de Valores Nulos en 'diff_tiempo_clasificacion'**: Los valores nulos en la diferencia de tiempo de clasificación pueden indicar falta de información o registros incompletos. Se eliminan las filas con valores nulos en 'diff_tiempo_clasificacion' para mantener la coherencia y precisión en los datos relacionados con los tiempos de clasificación.

5.2 Segunda estrategia

La segunda estrategia de tratamiento de datos aplicada implica un enfoque ligeramente diferente en comparación con la primera estrategia. A continuación, se detalla cada paso de esta segunda estrategia:

- **Rellenar Valores Nulos en Puntos y Victorias con 0:** Para las variables puntos del piloto, victorias del piloto, puntos del constructor, victorias del constructor, porcentaje de victorias y porcentaje de victorias del constructor. En lugar de eliminar las filas con valores nulos en estas columnas (como se hizo en la primera estrategia), se opta por rellenar los valores nulos con cero. Esto permite mantener los registros, asumiendo que la falta de puntos o victorias se debe a que no se han obtenido hasta el momento.
- **Eliminación de Filas con 'posicion_salida' Igual a 0**
- **Eliminación de Filas con 'posicion_clasificacion' o 'posicion_clasificacion_constructor' Nulos o Cero**
- **Eliminación de Valores Nulos en 'diff_tiempo_clasificacion'**

5.3 Tercera estrategia

La tercera estrategia para el tratamiento de datos introduce un enfoque novedoso para abordar valores nulos o cero en ciertas columnas clave, combinando métodos de las dos estrategias previas. A continuación, se detallan los pasos de esta estrategia:

- **Rellenar Valores Nulos en Puntos y Victorias con 0**:
- **Reemplazo de Valores Nulos o Cero en 'posicion_salida', 'posicion_clasificacion' y 'posicion_clasificacion_constructor'**: Para estas columnas, la estrategia se aparta de la simple eliminación o asignación de ceros. En lugar de ello, se implementa una lógica más sofisticada: los valores nulos o cero se reemplazan buscando el máximo valor en esa columna específica para la misma ronda y temporada, y sumando una unidad a este valor. Este enfoque asume que, si un valor es nulo o cero, el piloto o el constructor correspondiente debe haber ocupado la última posición en esa ronda y temporada. Esto garantiza que los registros mantengan una coherencia lógica y temporal.
- **Eliminación de Valores Nulos en 'diff_tiempo_clasificacion'**: Como en las estrategias anteriores, los valores nulos en la columna 'diff_tiempo_clasificacion' se eliminan. Esto se debe a que la falta de información en esta variable podría desvirtuar los análisis relacionados con los tiempos de clasificación.

5.4 Cuarta estrategia

La cuarta estrategia sigue la lógica de la tercera estrategia, pero introduce un cambio significativo: la eliminación de la columna 'diff_tiempo_clasificacion'. Esta decisión se basa en la idea de que los tiempos de clasificación, aunque informativos, pueden no ser esenciales para el análisis de predicción de podios en carreras. Los detalles de esta estrategia son los siguientes:

- **Rellenar Valores Nulos en Puntos y Victorias con 0**
- **Reemplazo de Valores Nulos o Cero en 'posicion_salida', 'posicion_clasificacion' y 'posicion_clasificacion_constructor'**: Utilizando la misma lógica que en la tercera estrategia, los valores nulos o cero en estas columnas se reemplazan con el máximo valor más uno para la misma ronda y temporada. Esto asegura que los registros mantengan una consistencia lógica y temporal, asumiendo que los valores nulos o cero indican la última posición.
- **Eliminación de la Columna 'diff_tiempo_clasificacion':** A diferencia de las estrategias anteriores, en la cuarta estrategia se decide eliminar completamente la columna 'diff_tiempo_clasificacion'. Esta decisión se basa en varias razones, como la complejidad añadida que esta variable aporta al modelo, la posible falta de relevancia de esta variable para la predicción del podio, o la presencia de demasiados valores faltantes que podrían sesgar el análisis.

5.5 Quinta estrategia

La quinta estrategia para el tratamiento de datos presenta un enfoque diferenciado en el manejo de los valores nulos en la columna 'diff_tiempo_clasificacion', combinando elementos de las estrategias anteriores. Los pasos de esta estrategia son los siguientes:

- **Rellenar Valores Nulos en Puntos y Victorias con 0**
- **Reemplazo de Valores Nulos o Cero en 'posicion_salida', 'posicion_clasificacion' y 'posicion_clasificacion_constructor'**: Esta parte de la estrategia sigue el enfoque de la tercera y cuarta estrategia.
- **Rellenar Valores Nulos en 'diff_tiempo_clasificacion' con el Máximo Valor para esa Ronda y Temporada**: A diferencia de las estrategias anteriores, donde los valores nulos en 'diff_tiempo_clasificacion' se eliminaban, esta estrategia adopta un enfoque de imputación. Los valores nulos en esta columna se rellenan con el máximo valor de 'diff_tiempo_clasificacion' encontrado en la misma ronda y temporada. Esta técnica permite retener más datos, asumiendo que los valores nulos pueden ser aproximados por el peor rendimiento en esa carrera específica.

5.6 Sexta estrategia

La sexta estrategia introduce un enfoque innovador en el manejo de la variable 'diff_tiempo_clasificacion', categorizándola en rangos. A continuación, se detallan los pasos de esta estrategia:

- **Rellenar Valores Nulos en Puntos y Victorias con 0**
- **Reemplazo de Valores Nulos o Cero en 'posicion_salida', 'posicion_clasificacion' y 'posicion_clasificacion_constructor':** Siguiendo la lógica de estrategias previas, se reemplazan los valores nulos o cero en estas columnas con el máximo valor más uno para la misma ronda y temporada, asumiendo que representan la última posición.
- **Categorización de 'diff_tiempo_clasificacion' en Rangos:** En lugar de eliminar o rellenar los valores nulos en 'diff_tiempo_clasificacion', esta estrategia introduce una categorización basada en rangos predefinidos. Se utiliza una función para asignar categorías como 'Muy bajo', 'Bajo', 'Medio', 'Alto' y 'Muy alto' a los valores de esta columna, incluyendo una categoría 'Desconocido' para los valores nulos. Esta técnica permite transformar una variable numérica continua en una variable categórica, lo que puede ser útil en ciertos análisis o modelos de predicción.
- **Eliminación de la Columna Original 'diff_tiempo_clasificacion':** Tras aplicar la categorización, la columna original 'diff_tiempo_clasificacion' se elimina del dataframe, dejando solo la nueva columna categórica. Esto simplifica el dataframe y se enfoca en la información categorizada que podría ser más relevante para ciertos análisis.

5.7 Séptima estrategia

La séptima estrategia de tratamiento de datos se centra en un enfoque detallado para manejar valores nulos o cero en las columnas 'posicion_clasificacion' y 'posicion_clasificacion_constructor'. Los pasos de esta estrategia son los siguientes:

- **Rellenar Valores Nulos en Puntos y Victorias con 0**
- **Reemplazo de Valores Nulos o Cero en 'posicion_salida', 'posicion_clasificacion' y 'posicion_clasificacion_constructor':** Idénticamente a las últimas estrategias, se asume que tienen la última posición.
- **Rellenar 'diff_tiempo_clasificacion' con el Tiempo Máximo más Uno**: Para los valores nulos en 'diff_tiempo_clasificacion', se reemplazan con el tiempo máximo de clasificación en esa ronda y temporada, más uno. Este enfoque busca mantener la coherencia dentro de los tiempos de clasificación, suponiendo que un valor nulo indica un desempeño significativamente peor.
- **Eliminación de Valores Extremos en 'diff_tiempo_clasificacion'**: Se calcula el percentil 95 para 'diff_tiempo_clasificacion' y se eliminan las filas donde este valor es superado, con el objetivo de remover los outliers o valores extremadamente altos que podrían distorsionar el análisis.

5.8 Octava estrategia

La última estrategia implementada en el dataframe sigue un enfoque similar al de la séptima estrategia, pero introduce variaciones en el manejo de los valores nulos o cero en 'posicion_clasificacion' y 'posicion_clasificacion_constructor':

- **Rellenar Valores Nulos en Puntos y Victorias con 0**
- **Reemplazo de Valores Nulos o Cero en 'posicion_clasificacion' y 'posicion_clasificacion_constructor'**: En esta estrategia, se busca primero el máximo valor en la ronda y temporada actuales para reemplazar el valor nulo o cero. Si no se encuentra un valor máximo actual, se busca en la temporada anterior. Si aún así no se encuentra un valor adecuado, se asigna la última posición posible en esa ronda y temporada.
- **Rellenar 'diff_tiempo_clasificacion' con el Tiempo Máximo de la Ronda y Temporada**: Similar a la séptima estrategia, se rellenan los valores nulos en 'diff_tiempo_clasificacion' con el máximo valor encontrado para esa ronda y temporada.Normalización de Datos

El proceso de predicción, a pesar de ser la culminación de todo el trabajo previo, no es simplemente el acto de aplicar un modelo a un conjunto de datos. La preparación adecuada del dataset y la interpretación de los resultados son vitales para garantizar que las predicciones sean significativas y útiles. En esta sección,

abordaremos el último paso antes de la predicción: la normalización, así como el proceso de predicción en sí y cómo interpretar los resultados.

Antes de alimentar nuestros datos al modelo, es esencial que estén en una escala uniforme. Muchos algoritmos de ML, incluida la regresión logística, son sensibles a las escalas de las características. Si las características tienen escalas muy diferentes entre sí, pueden dominar unas sobre otras, lo que puede llevar a un rendimiento subóptimo del modelo.

Para abordar este problema, utilizamos StandardScaler de scikit-learn, que es una técnica de normalización que transforma cada característica para que tenga una media de 0 y una desviación estándar de 1. Esta transformación garantiza que todas las características tengan el mismo peso en el modelo.

El análisis y modelado de datos en cualquier dominio exige una meticulosa preparación y tratamiento de los datos. A través de este capítulo hemos revelado no solo la riqueza y profundidad de la información disponible, sino también la complejidad y los desafíos asociados con su manejo y transformación para la extracción de conocimientos valiosos.

De esta forma, hemos destacado una verdad fundamental del Machine Learning: los modelos son tan buenos como los datos con los que se alimentan. Desde la recopilación inicial hasta las etapas finales de preparación, cada paso, cada decisión y cada transformación juega un papel vital en la calidad de las predicciones finales. En resumen, el tratamiento de datos no es simplemente un preludio al análisis real; es el fundamento sobre el cual se construye todo el edificio del Machine Learning.

6 RESULTADOS

Tras todo el trabajo de preparación, análisis y modelado de datos, finalmente llegamos a la fase que es la culminación de todo el esfuerzo: la obtención y análisis de resultados. En este capítulo, analizaremos los resultados generados por nuestro modelo de Machine Learning, analizando la precisión y eficacia con la que ha predicho los desenlaces de las carreras de Fórmula 1 a lo largo de las temporadas.

Para obtener estos resultados, hemos empleado un proceso estructurado y sistemático. Basándonos en el conocimiento acumulado a lo largo de los capítulos anteriores, hemos definido un protocolo específico para evaluar el desempeño de nuestro modelo. Comenzamos por dividir nuestro conjunto de datos en segmentos de entrenamiento y prueba basados en temporadas. Esta estrategia de segmentación temporal es esencial para simular un escenario realista de predicción, donde buscamos prever el resultado de una temporada futura basándonos en los datos históricos disponibles hasta ese momento. La Figura 25 muestra barras apiladas que representan la cantidad de temporadas utilizadas para el entrenamiento del modelo año tras año.

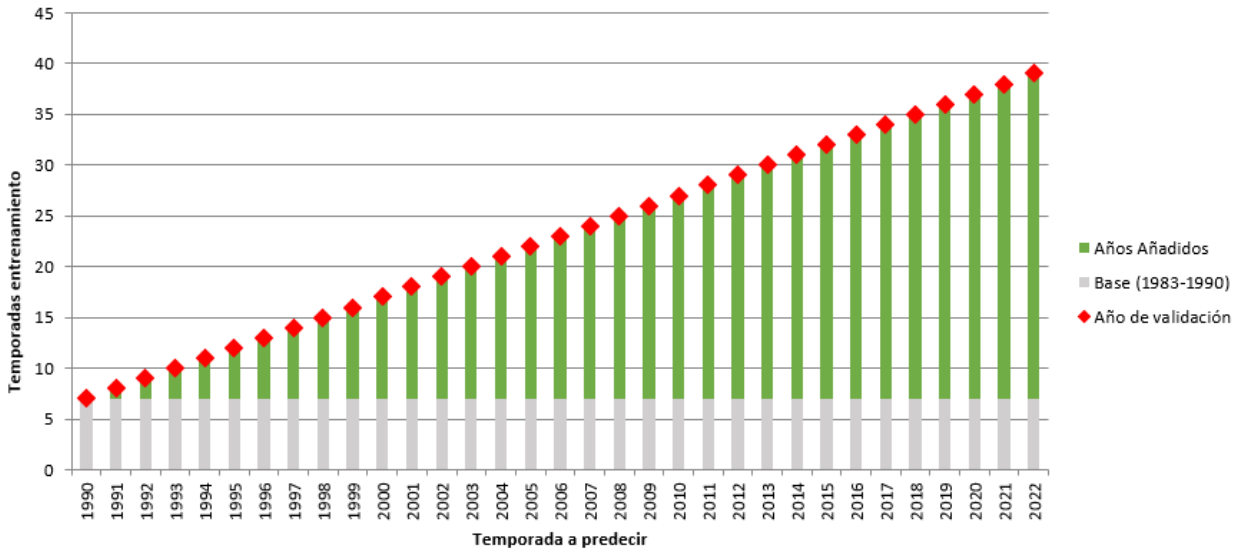

Figura 25: Esquema de representación de los años empleados para entrenar y validar el modelo de predicción

En este esquema, las barras grises representan una base de datos de entrenamiento inicial que abarca desde 1983 hasta 1990. Estos datos se usaron para entrenar el primer modelo, que se validó con la temporada de 1991.

A medida que avanzamos cronológicamente, cada nueva temporada se añade al conjunto de entrenamiento, representado por las barras verdes adicionales. Por ejemplo, para validar el modelo para 1992, se incluyen los datos de 1991 en el entrenamiento. Este proceso se repite, añadiendo secuencialmente cada nueva temporada al conjunto de entrenamiento y utilizando la temporada siguiente como el conjunto de prueba para la validación.

Los cuadrados rojos indican el año específico que se utiliza para la validación del modelo entrenado hasta el año anterior.

En el código que hemos implementado hemos realizado un bucle que itera desde el año 1990 hasta el 2022, lo que implica que el modelo se ha entrenado y validado 33 veces, cada vez incorporando datos históricos acumulativos para entrenar el modelo y utilizando los datos del año en curso para la validación.

Posteriormente, procedimos a codificar las variables categóricas, como el piloto, el constructor y el circuito, utilizando codificación one-hot. Esta técnica transforma cada categoría en una columna binaria, permitiendo que los modelos matemáticos procesen y utilicen esta información de manera efectiva.

Sin embargo, al codificar estas variables, nos encontramos con un reto. No todos los pilotos, constructores o circuitos están presentes en todas las temporadas. Esto significa que, después de codificar, los conjuntos de entrenamiento y prueba podrían no tener el mismo conjunto de columnas. Para solucionar este problema, aseguramos que ambos conjuntos tuvieran las mismas columnas, agregando columnas faltantes con valor cero cuando fuera necesario.

Durante este proceso, se toman medidas para asegurar que las variables categóricas se codifiquen correctamente y que los conjuntos de datos se normalicen antes de proceder con el entrenamiento del modelo de regresión logística. Una vez normalizados los datos, hemos entrenado nuestro modelo de regresión logística. Elegimos este modelo debido a su capacidad para manejar problemas de clasificación binaria, como el que estamos abordando: predecir si un piloto alcanzará el podio o no. Los tres pilotos con más probabilidad de alcanzar el podio en cada ronda serán evaluados con un 1 y los demás con un 0.

Tras entrenar el modelo, procederemos a evaluar el desempeño de los pilotos. Utilizaremos una métrica clave para medir la precisión de nuestras predicciones: la proporción de aciertos basada en coincidencias de 1's. En otras palabras, nos preguntamos con qué frecuencia nuestro modelo predecirá correctamente que un piloto alcanzará el podio. Hemos decidido que esta sea la metodología ya que en cada carrera compiten, como mínimo, 20 pilotos. Dado que solo tres de ellos conseguirán un lugar en el podio, predecir y acertar a los que no lo lograrán resulta más sencillo. Por lo tanto, nuestro foco está en predecir correctamente a aquellos que sí lo lograrán.

6.1 Resultados de la primera estrategia mediante Regresión Logística

En primer lugar, hemos probado la primera estrategia de tratamiento de datos y hemos obtenido los primeros resultados. Esta estrategia es la más sencilla y se centra en garantizar la integridad y coherencia de los datos, eliminando registros que contienen valores que no tienen sentido práctico (como posición de salida 0), así como aquellos con información incompleta o nula en variables clave. Al hacerlo, se mejora la calidad del *dataset* para análisis posteriores, asegurando que se basen en datos precisos y completos. La Figura 26 representa la precisión alcanzada en la predicción que se ha realizado para cada año.

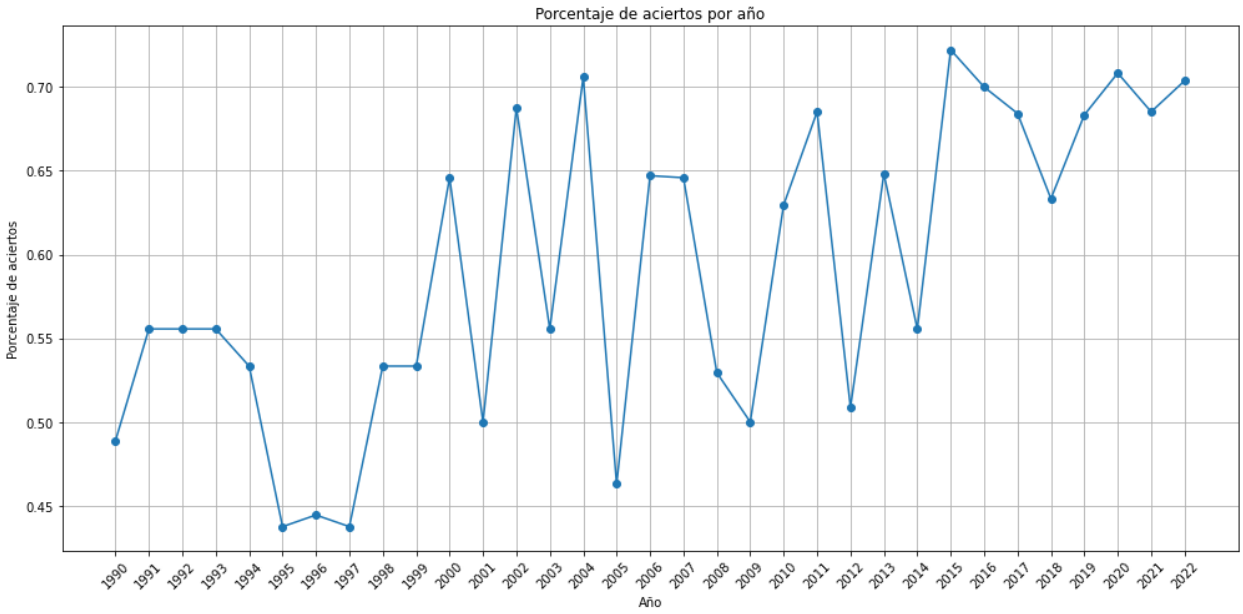

Figura 26: Precisión de las predicciones anualmente con la primera estrategia

La tendencia general observada a lo largo de los años muestra un incremento sostenido en el porcentaje de aciertos del modelo, particularmente a partir del año 2000. Este aumento sugiere que la estrategia de tratamiento de datos ha sido efectiva en mejorar la capacidad predictiva del modelo con el paso del tiempo. Sin embargo, la variabilidad en los porcentajes de acierto entre 1990 y 2000 es notable, fluctuando significativamente entre aproximadamente el 43.75% y el 64.58%. Esta fluctuación podría estar relacionada con cambios en el deporte, como modificaciones en las reglas de la Fórmula 1, la evolución en la composición de los equipos o la calidad de los datos disponibles en esos años.

A partir de 2010, los porcentajes de acierto se han estabilizado y muestran una mayor consistencia, manteniéndose generalmente por encima del 60%. Esto indica que tanto el modelo como la estrategia de tratamiento de datos han sido refinados, resultando en predicciones más precisas y confiables. Además, se han observado picos de alto rendimiento en años específicos, como 2004, 2015, 2016 y 2022, donde el porcentaje de acierto superó el 70%. Estos picos pueden reflejar temporadas en las que la dinámica de la Fórmula 1 fue más predecible, posiblemente debido al dominio de ciertos pilotos o equipos. Específicamente en 2022, el modelo alcanzó un impresionante 70.37% de aciertos, destacando la eficacia de la estrategia implementada.

La matriz de confusión es una herramienta esencial en estadísticas y aprendizaje automático para evaluar la precisión de un algoritmo de clasificación. Nos permite desglosar las predicciones del modelo en cuatro categorías: verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. En nuestro caso, las clases que estamos tratando de predecir son "Podio" (si un piloto alcanza el podio) y "No Podio" (si un piloto no lo logra). Dicha matriz se presenta de la siguiente manera en la Figura 27:

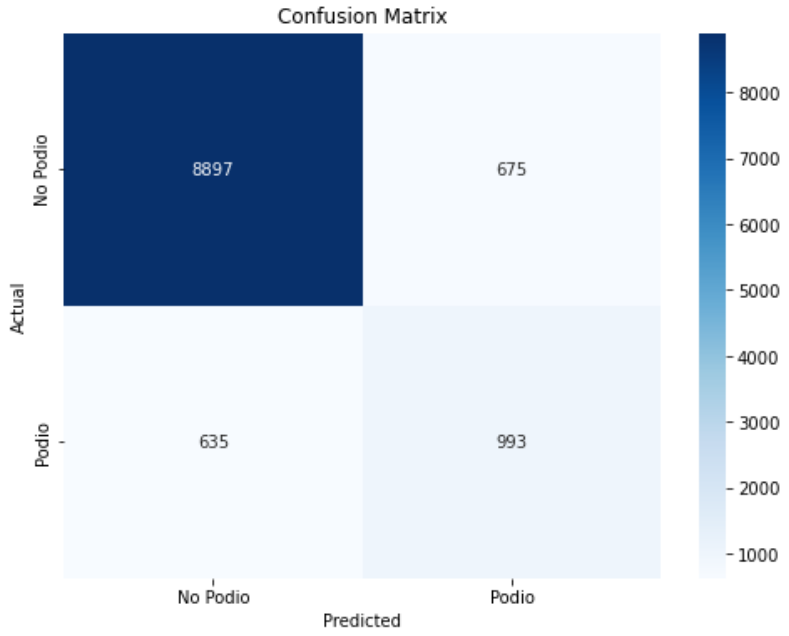

Figura 27: Matriz de confusión con los resultados de la primera estrategia

Para interpretarla:

- El valor 8897 (esquina superior izquierda) nos indica los verdaderos negativos (TN). Representa las veces que el modelo acertadamente predijo que un piloto no llegaría al podio y efectivamente no lo hizo.
- El valor 675 (esquina superior derecha) nos muestra los falsos positivos (FP). Son las ocasiones en las que el modelo predijo que un piloto llegaría al podio, pero no fue así.
- El 635 (esquina inferior izquierda) representa los falsos negativos (FN). Aquí, el modelo anticipó que un piloto no lograría el podio, pero en la realidad sí lo hizo.
- Finalmente, el 993 (esquina inferior derecha) son los verdaderos positivos (TP). Indica las veces que el modelo correctamente predijo que un piloto alcanzaría el podio y efectivamente fue así.

Un detalle que destacar es la elevada cantidad de verdaderos negativos. Dado que en cada carrera sólo tres pilotos de más de 20 pueden llegar al podio, es estadísticamente más probable acertar al anticipar que un piloto no llegará al podio que lo contrario. Esta naturaleza del deporte hace que las predicciones negativas (No Podio) sean más sencillas de acertar.

Sin embargo, lo realmente relevante es la cantidad de verdaderos positivos (933). A pesar de las dificultades para predecir quiénes estarán en el podio, el modelo ha demostrado ser efectivo en esta tarea. Esto sugiere que ha logrado identificar y usar correctamente la información relevante para hacer predicciones informadas.

A pesar de estos aciertos, la presencia de falsos positivos (675) y falsos negativos (635) indica áreas en las que el modelo podría mejorar. Los falsos positivos señalan pilotos que, según el modelo, tenían potencial de estar en

el podio, pero no lo lograron, mientras que los falsos negativos indican pilotos que superaron las expectativas del modelo.

El rendimiento del modelo no solo puede evaluarse a través de la matriz de confusión, sino que también se puede analizar más detalladamente mediante métricas específicas que se derivan de ella. A continuación, se revisan tres métricas cruciales:

- **Accuracy (Precisión global):** Esta métrica proporciona una visión general de la proporción de predicciones correctas en relación con todas las predicciones realizadas. En nuestro contexto:

$$Accuracy = \frac{993 + 8897}{993 + 8897 + 675 + 635} \approx 0.883$$

Una Precisión de aproximadamente 0.899 significa que el modelo acertó en el 88.3% de sus predicciones totales.

- **Precision (Precisión):** Esta medida se refiere a la proporción de predicciones positivas correctas en relación con todas las predicciones positivas realizadas por el modelo. El cálculo sería el siguiente:

$$Precisión = \frac{993}{993 + 675} \approx 0.5953$$

El porcentaje de acierto global de 59.53% es significativamente superior al azar, considerando que en cada carrera solo tres pilotos alcanzan el podio de entre al menos 20 competidores. Este nivel de acierto subraya la eficacia del modelo en un contexto desafiante, teniendo en cuenta la complejidad y la naturaleza competitiva de la Fórmula 1.

- **Recall (Sensibilidad):** Esta métrica indica la proporción de verdaderos positivos en relación con la suma de verdaderos positivos y falsos negativos. Es especialmente útil para comprender cuán bien el modelo identifica los eventos positivos reales. Matemáticamente, utilizando los valores de nuestra matriz, se calcula como:

$$Recall = \frac{993}{993 + 635} \approx 0.6099$$

Un Recall de aproximadamente 0.61 indica que el modelo pudo identificar correctamente el 60.99% de los pilotos que realmente alcanzaron el podio. Es un valor bastante destacable considerando la complejidad y la variabilidad inherente a las carreras de Fórmula 1.

- **Specificity (Especificidad):** Esta métrica evalúa la proporción de verdaderos negativos en relación con la suma de verdaderos negativos y falsos positivos. Ayuda a entender cuán bien el modelo identifica los eventos negativos reales. Para nuestro caso se formula como:

$$Specificity = \frac{8897}{8897 + 675} \approx 0.9295$$

Una Specificity de aproximadamente 0.9295 sugiere que el modelo es muy eficiente (92,95% de precisión) al predecir correctamente aquellos pilotos que no alcanzarían el podio. Dada la naturaleza de las carreras de

Fórmula 1, donde en una carrera típica participan más de 20 pilotos y solo 3 logran el podio, es lógico que la especificidad sea alta. Hay una mayor probabilidad de acertar al anticipar que un piloto no llegará al podio que lo contrario. En otras palabras, es estadísticamente más sencillo y probable predecir correctamente a aquellos pilotos que no alcanzarán posiciones de podio, lo que explica la elevada especificidad observada en nuestro modelo.

6.2 Resultados de la segunda estrategia mediante Regresión Logística

La principal diferencia entre la primera y la segunda estrategia radica en el tratamiento de los valores nulos en las variables relacionadas con puntos y victorias. Mientras que en la primera estrategia se optó por eliminar dichas filas, en la segunda estrategia se decide rellenar los valores nulos con cero. Esto nos hacía pensar que nos permitía conservar un mayor número de registros en el análisis, bajo la suposición de que la ausencia de puntos o victorias es equivalente a no haberlos obtenido hasta el momento. Sin embargo, no ha sido así y los resultados eran prácticamente los mismos. Deducimos que es debido a que el registro era el mismo y que aunque se rellenaban los puntos y victorias con 0, se eliminaban la mayoría por las demás condiciones.

6.3 Resultados de la tercera estrategia mediante Regresión Logística

En resumen, como recordatorio, esta tercera estrategia equilibra la necesidad de preservar tantos registros como sea posible (rellenando algunos valores nulos con ceros) con la necesidad de mantener la coherencia lógica y temporal de los datos (reemplazando otros valores nulos o cero con el máximo valor más uno para cada caso específico).

Así, la estrategia de de análisis de datos para las predicciones en la Fórmula 1 muestra resultados interesantes como podemos observar en la Figura 28:

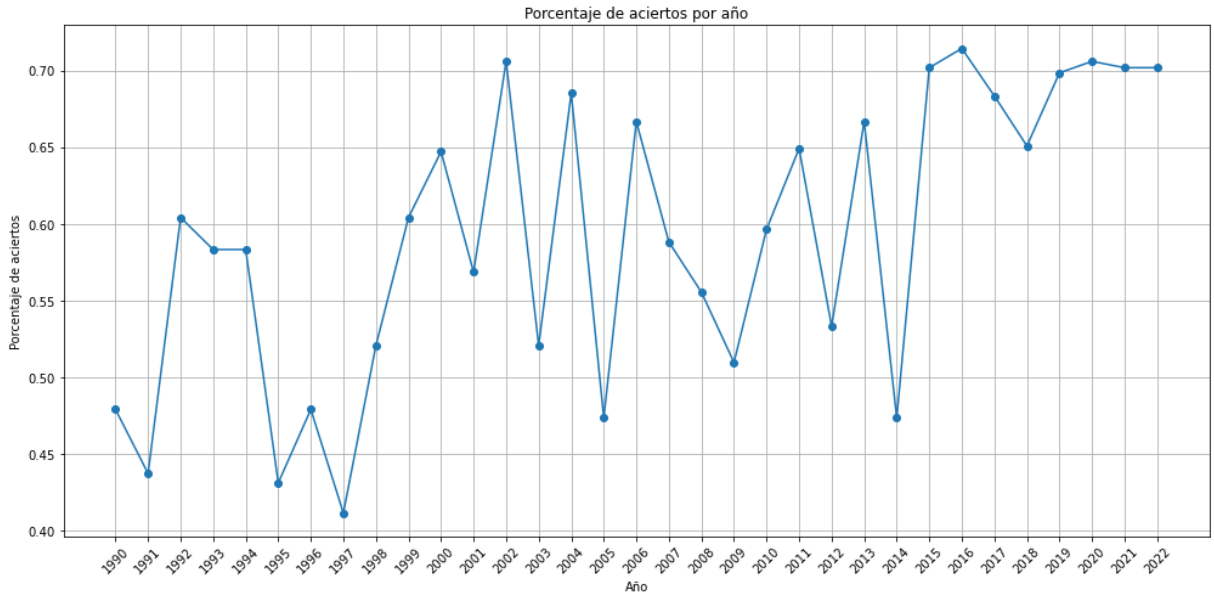

Figura 28: Precisión de las predicciones anualmente con la tercera estrategia

Con un porcentaje de precisión global del 59.61%, esta estrategia se sitúa ligeramente por encima de la primera en términos de precisión. Aunque la precisión global es ligeramente superior a la de la primera estrategia, la diferencia puede no ser muy llamativa, pero no menos importante. Cuando analizamos la matriz de confusión de la Figura 29, podemos observar un aumento significativo de los verdaderos positivos y negativos.

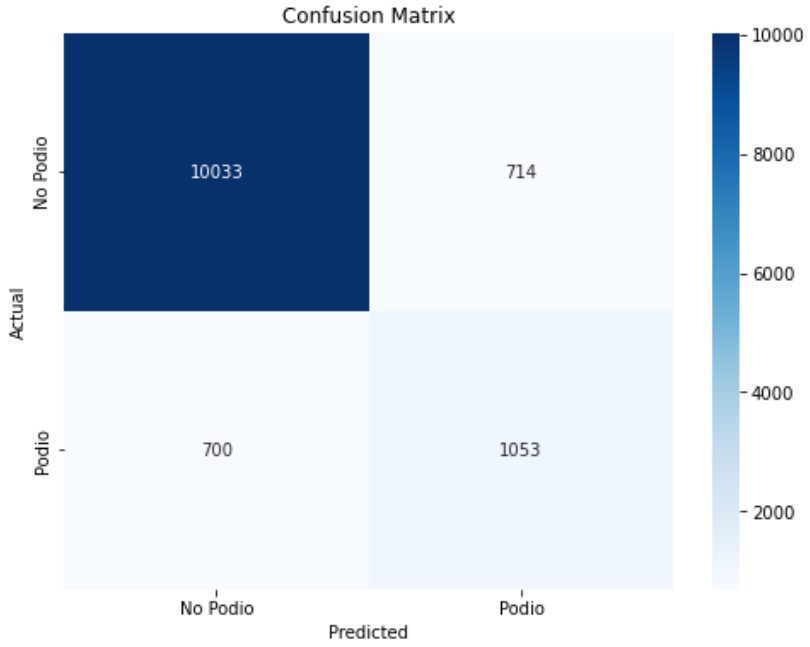

Figura 29: Matriz de confusión con los resultados de la tercera estrategia

La matriz muestra 10,033 verdaderos negativos y 1,053 verdaderos positivos. Esto indica una buena capacidad del modelo para predecir correctamente tanto los pilotos que no alcanzarán el podio como aquellos que sí lo harán, aunque sigue existiendo margen de mejora en la reducción de los falsos positivos y negativos. A partir de esta matriz, las métricas calculadas tienen los siguientes valores:

Accuracy = 0.8869 **Precision** = 0.5961 **Recall** = 0.6007 **Specificity** = 0.9336

La precisión global de esta estrategia es ligeramente inferior a la de la primera estrategia. Esto sugiere que, aunque la tercera estrategia puede haber abordado algunos aspectos específicos de los datos, no mejoró la capacidad general del modelo para realizar predicciones correctas. La precisión de esta estrategia (59.61%) es ligeramente superior a la de la primera estrategia (59.53%), por lo que la calidad de las predicciones positivas (es decir, predecir correctamente los podios) es ligeramente superior. El recall o sensibilidad es similar entre las dos estrategias, con un 60.13% en la tercera frente al 60.99% en la primera. Esto implica que la capacidad del modelo para identificar correctamente los podios reales es comparable en ambas estrategias. La especificidad es donde vemos una diferencia notable. La tercera estrategia tiene una especificidad ligeramente superior (93.36% frente a 92.95%), lo que sugiere una mejora marginal en la capacidad de predecir correctamente a los pilotos que no alcanzan el podio.

6.4 Resultados de la cuarta estrategia mediante Regresión Logística

La cuarta estrategia se enfoca en simplificar el modelo y eliminar la variable diff_tiempo_clasificacion. Esta nueva configuración produce resultados que merecen un análisis detallado.

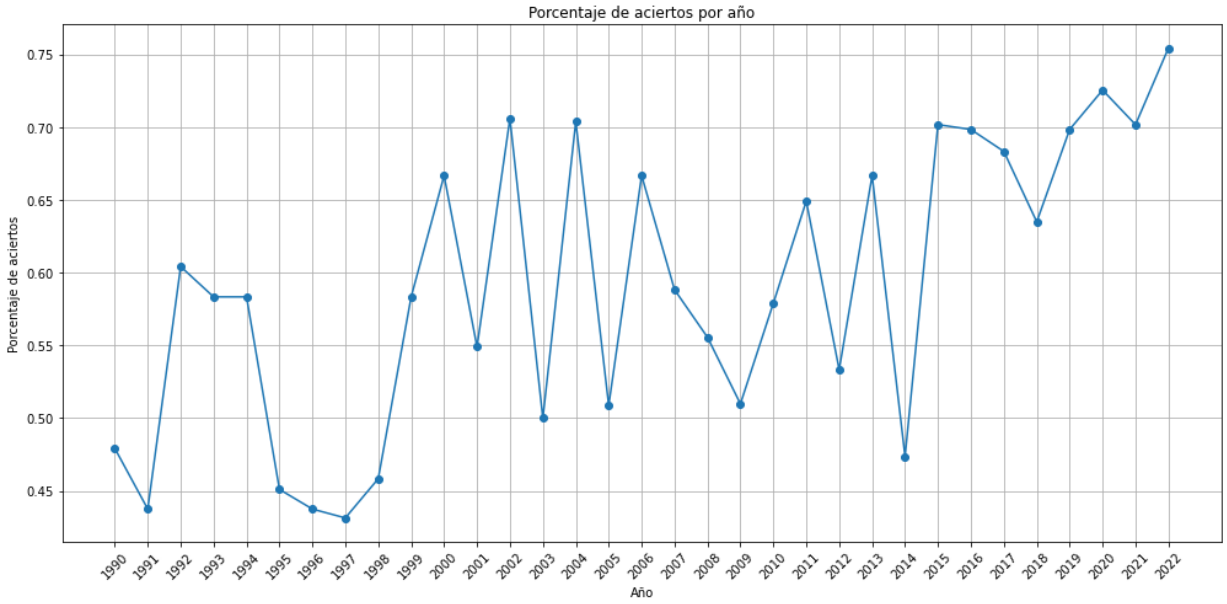

Figura 30: Precisión de las predicciones anualmente con la cuarta estrategia

Con un porcentaje de acierto global del 59.54%, esta estrategia muestra una leve mejora en comparación con la estrategia anterior. Aunque el incremento en la precisión es modesto, es un paso adelante en la dirección correcta, especialmente considerando la importancia de la métrica de precisión en nuestro análisis.

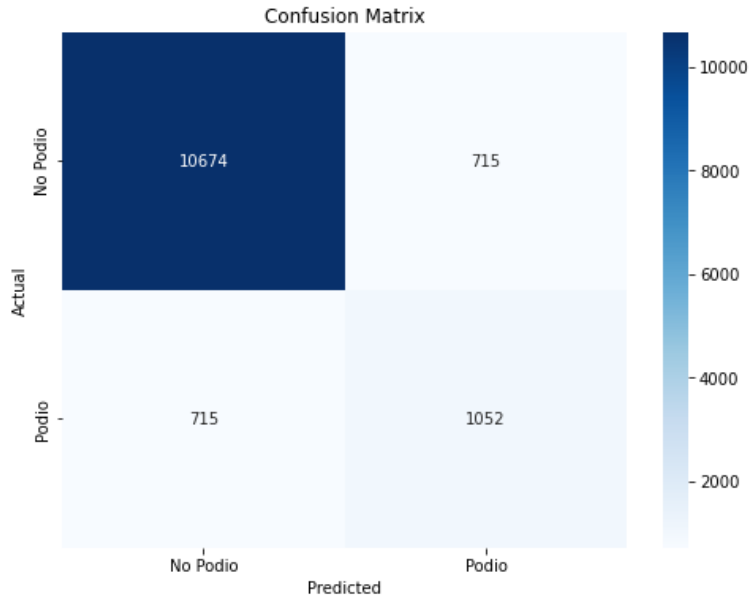

Figura 31: Matriz de confusión con los resultados de la cuarta estrategia

La matriz de confusión muestra 10,674 verdaderos negativos y 1,052 verdaderos positivos. Estos números sugieren una eficiencia comparable en la predicción de los pilotos que no lograrán el podio y aquellos que sí lo harán. Basándonos en esta matriz, calculamos las siguientes métricas:

Accuracy = 0.8913 **Precision** = 0.5954 **Recall** = 0.5954 **Specificity** = 0.9372

La precisión global de la cuarta estrategia es similar a las estrategias anteriores, lo que indica una consistencia en la capacidad del modelo para realizar predicciones correctas. La especificidad es ligeramente superior a las estrategias previas, lo que puede reflejar una mejora en identificar correctamente a los pilotos que no alcanzan el podio.

6.5 Resultados de la quinta estrategia mediante Regresión Logística

La quinta estrategia combina el reemplazo de valores nulos o cero en ciertas columnas clave con una técnica de imputación específica para 'diff_tiempo_clasificacion', buscando un equilibrio entre la retención de registros y la coherencia lógica de los datos. Los resultados anuales muestran la siguiente tendencia:

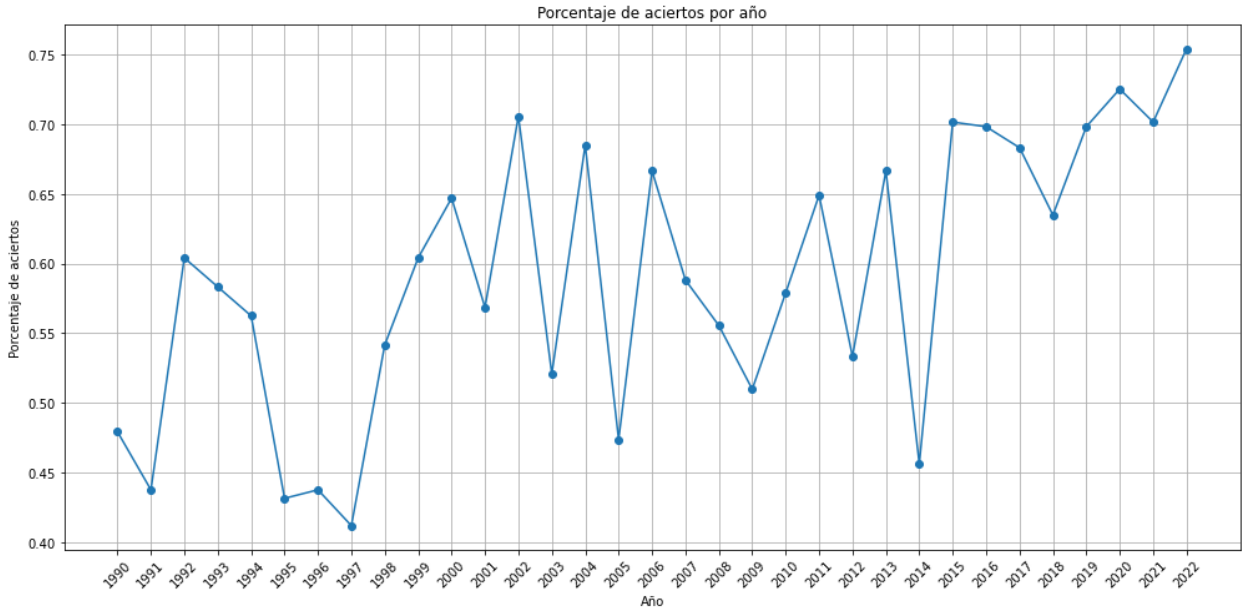

Figura 32: Precisión de las predicciones anualmente con la quinta estrategia

El porcentaje de acierto global es de 59.48%, lo que representa una ligera disminución en comparación con las estrategias anteriores. Esta reducción, aunque pequeña, es significativa dado que nuestro objetivo principal es mejorar la precisión.

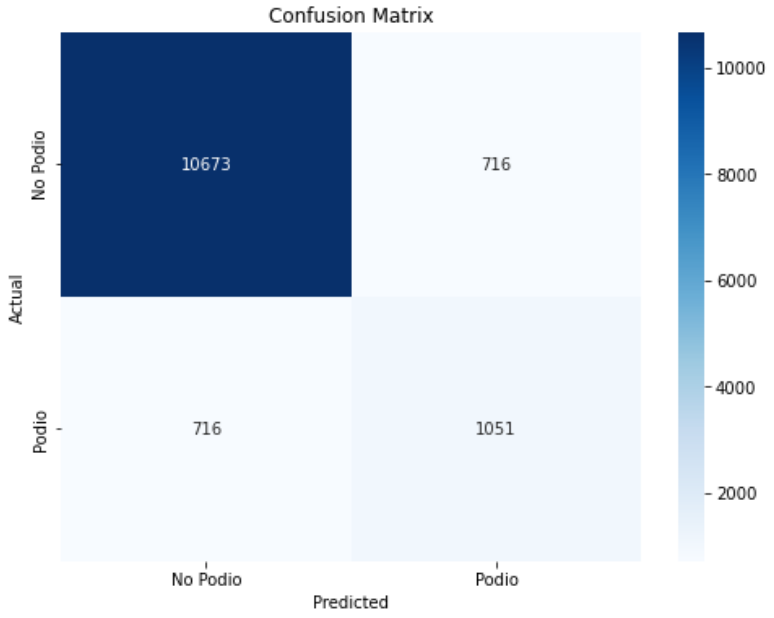

Figura 33:Matriz de confusión con los resultados de la quinta estrategia

La matriz de confusión muestra 10,673 verdaderos negativos y 1,051 verdaderos positivos, indicando una capacidad constante del modelo para predecir resultados. Las métricas de calidad resultan:

Accuracy = 0.8912 **Precision** = 0.5948 **Recall** = 0.5948 **Specificity** = 0.9371

6.6 Resultados de la sexta estrategia mediante Regresión Logística

La sexta estrategia se centra en la transformación de 'diff_tiempo_clasificacion' de una variable numérica a una categórica, buscando un enfoque más analítico y posiblemente más adecuado para modelos de clasificación o análisis que se beneficien de variables categóricas.

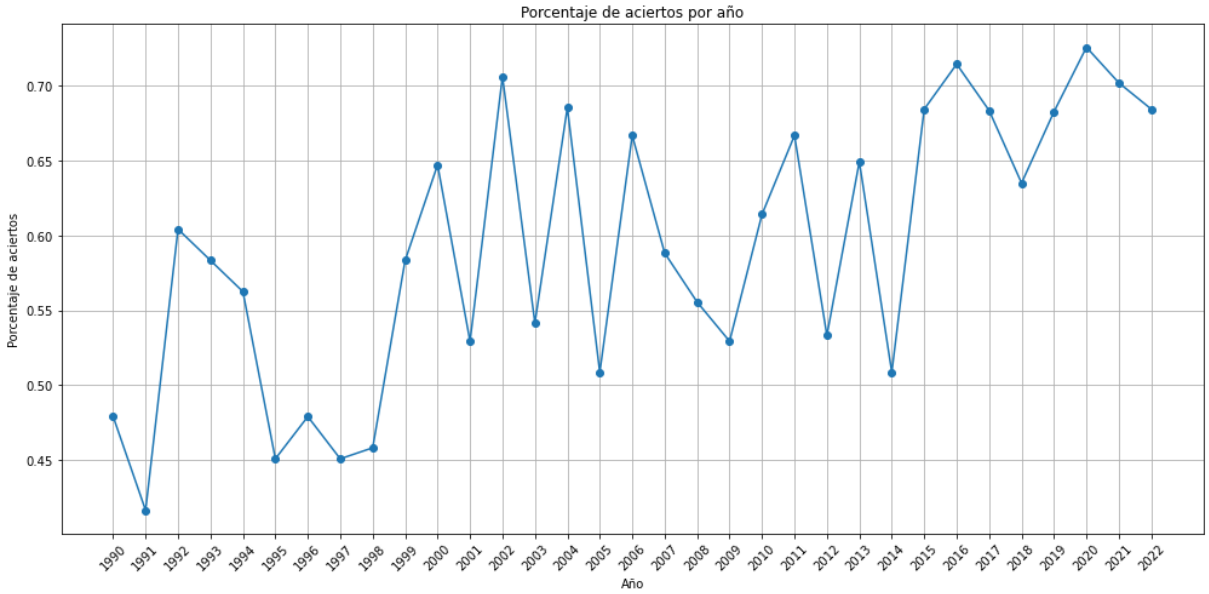

Figura 34: Precisión de las predicciones anualmente con la sexta estrategia

El porcentaje de acierto global es de 59.51%, idéntico al de la cuarta estrategia. Esto sugiere que los cambios implementados en esta estrategia no han tenido un impacto significativo en la mejora de la precisión.

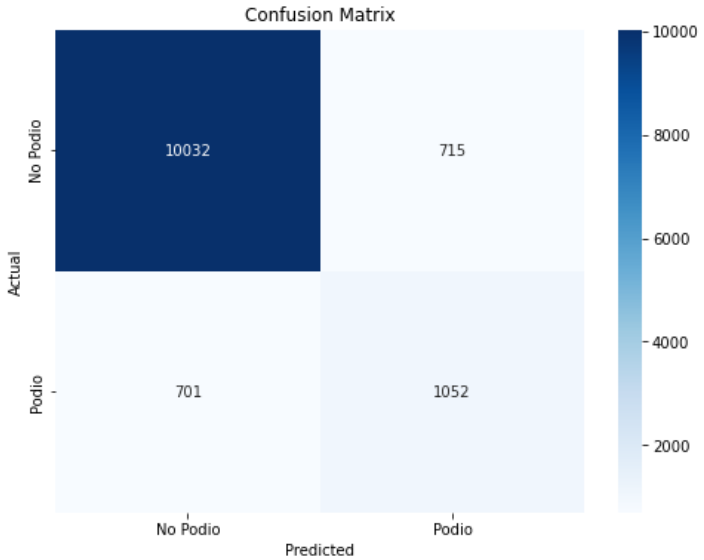

Figura 35: Matriz de confusión con los resultados de la sexta estrategia

La matriz muestra 10,032 verdaderos negativos y 1,052 verdaderos positivos, lo que demuestra una capacidad consistente del modelo para realizar predicciones precisas. Las métricas de calidad son:

Accuracy = 0.8922 **Precision** = 0.5951 **Recall** = 0.5999 **Specificity** = 0.9372

6.7 Resultados de la séptima estrategia mediante Regresión Logística

La séptima estrategia adopta un enfoque cuidadoso y contextual en el tratamiento de valores nulos o cero en 'posicion_clasificacion', 'posicion_clasificacion_constructor' y 'diff_tiempo_clasificacion'. Esta metodología se centra en preservar la integridad de los datos al tiempo que garantiza su relevancia y aplicabilidad para análisis detallados y modelización predictiva.

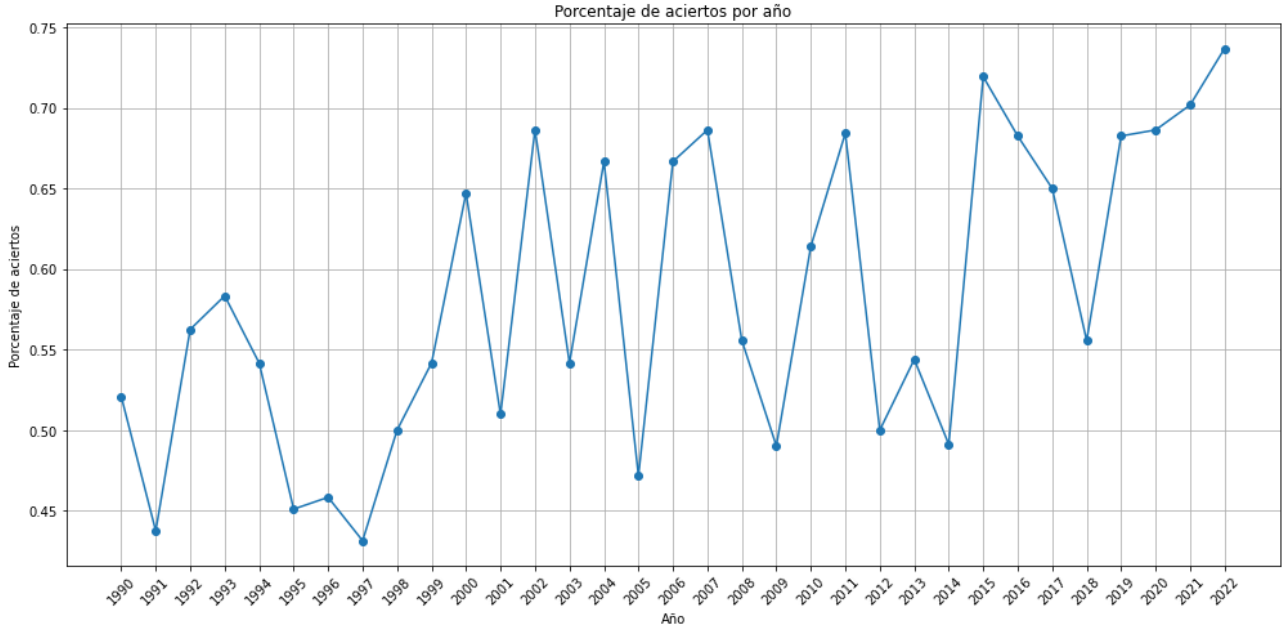

Figura 36: Precisión de las predicciones anualmente con la séptima estrategia

Con un porcentaje de acierto global del 58.55%, esta estrategia muestra una disminución en la precisión en comparación con las estrategias anteriores. Este descenso, aunque preocupante, ofrece valiosas lecciones para futuras estrategias.

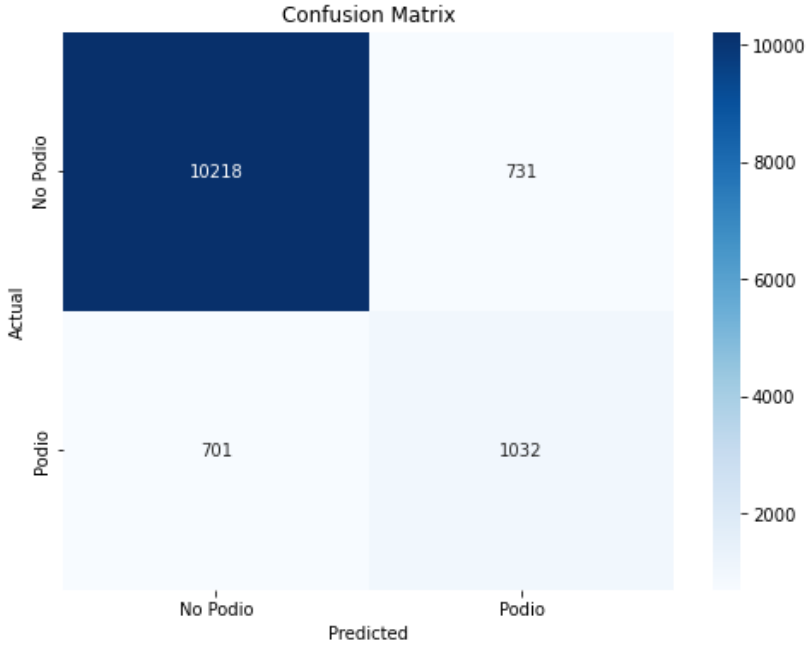

Figura 37: Matriz de confusión con los resultados de la séptima estrategia

La matriz de confusión revela 10,218 verdaderos negativos y 1,032 verdaderos positivos. Las métricas:

Accuracy = 0.8871 **Precision**= 0.5854 **Recall** = 0.5955 **Specificity** = 0.9332

6.8 Resultados de la octava estrategia mediante Regresión Logística

La octava estrategia se caracteriza por su enfoque exhaustivo y minucioso en el tratamiento de valores nulos o cero, especialmente en variables clave como 'posicion_clasificacion', 'posicion_clasificacion_constructor' y 'diff_tiempo_clasificacion'. Esta estrategia se destaca por su intento de maximizar la retención de datos útiles, evitando al mismo tiempo la distorsión de la información esencial.

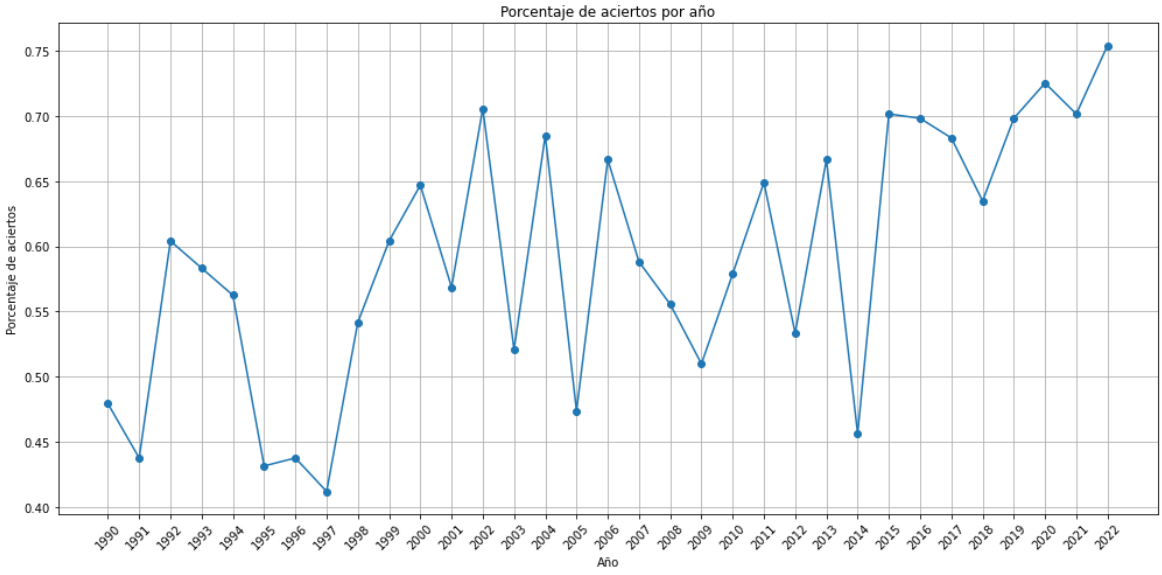

Figura 38: Precisión de las predicciones anualmente con la octava estrategia

El porcentaje de acierto global es de 59.48%, lo que indica una ligera ventaja en la precisión en comparación con las estrategias anteriores. Este resultado subraya la dificultad de mejorar la precisión en un campo tan complejo y variable como la Fórmula 1.

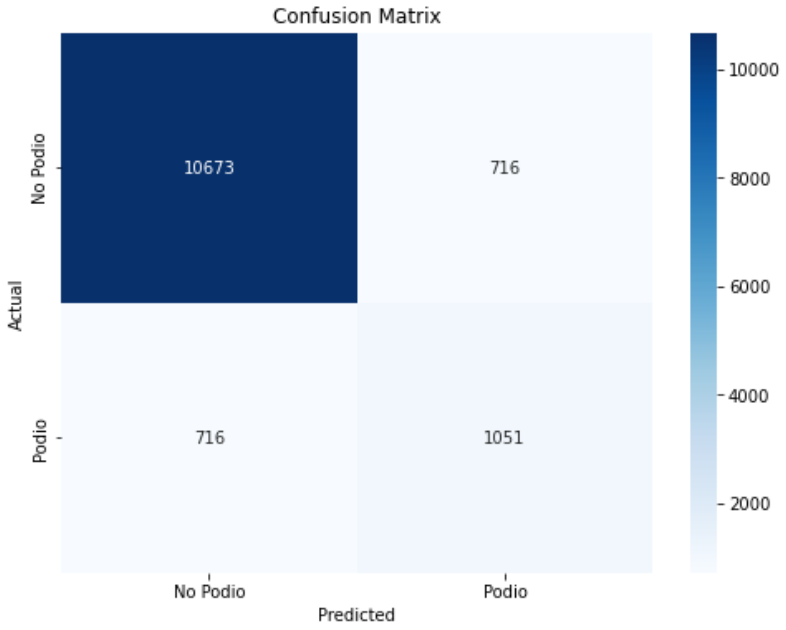

Figura 39: Matriz de confusión con los resultados de la octava estrategia

La matriz muestra 10,673 verdaderos negativos y 1,051 verdaderos positivos. Las métricas resultan:

Accuracy = 0.8912 **Precision**= 0.5948 **Recall** = 0.5948 **Specificity** = 0.9371

6.9 Discusión de la mejor estrategia

Para determinar la mejor estrategia, consideraremos tres factores principales:

- **Media de precisión de los últimos 10 años**: Nos resulta muy interesante medir esta variable ya que la predicción de estos últimos años cuenta con una base de datos más amplia y completa. Por tanto y como podemos observar en las gráficas de cada estrategia, la media precisión aumenta considerablemente estos años.
- **Número de filas en el dataframe**: Indica la cantidad de datos que se han utilizado para entrenar el modelo, lo cual es importante para la robustez del modelo.
- **Verdaderos positivos**: Como hemos explicado anteriormente, son únicamente 3 pilotos los que pueden conseguir el podio siendo, normalmente, más de 20 pilotos en la parrilla. Por ello, tener una gran cantidad de verdaderos positivos indica la habilidad del modelo para identificar correctamente el resultado final que buscamos.

	1ª estr.	*2ª estr.*	*3ª estr.*	*4ª estr.*	*5ª estr.*	*6ª estr.*	*7ª estr.*	*8ª estr.*
Prec. 10	67.24%	67.24%	66.99%	**67.39%**	67.21%	66.69%	65.19%	67.21%
Num. Fil.	13723	13723	15317	**16207**	**16207**	15317	15396	**16207**
VP	993	993	**1053**	1052	1051	1052	1032	1051
Accuracy	88.3%	88.3%	88.69%	89.13%	89.12%	**89.22%**	88.71%	89.12%
Precision	59.53%	59.53%	**59.61%**	59.54%	59.48%	59.51%	58.54%	59.48%
Recall	**60.99%**	**60.99%**	60.07%	59.54%	59.48%	59.99%	59.55%	59.48%
Specifity	92.95%	92.95%	93.36%	**93.72%**	93.71%	**93.72%**	93.32%	93.71%

Tabla 5: Resultados críticos de las ocho estrategias de tratamiento de datos

Como hemos podido demostrar, la cuarta estrategia destaca como la más prometedora. Presenta la media ponderada más alta (67.39%), lo que indica una precisión consistente y superior en los años más recientes. Además, cuenta con el mayor número de filas (16207), lo que implica una base de datos más robusta para realizar predicciones. Aunque su número de verdaderos positivos (1052) es ligeramente inferior al de la tercera estrategia, la diferencia es mínima y se compensa con su mayor precisión y volumen de datos.

La quinta y octava estrategias también muestran resultados sólidos, con una media ponderada muy cercana a la cuarta estrategia y un número igual de filas. Sin embargo, tiene ligeramente un verdadero positivo menos y una precisión media menor, situándola en un segundo lugar cercano.

Basándonos en los criterios establecidos, la cuarta estrategia se presenta como la más efectiva para predecir los podios en la Fórmula 1. Su equilibrio entre una alta precisión media, un volumen considerable de datos y un número elevado de verdaderos positivos la convierte en la opción más robusta y fiable para nuestras predicciones. Esto subraya la importancia de una estrategia bien equilibrada que no solo se enfoque en un único aspecto de los datos, sino que considere los factores puramente necesarios para maximizar la predictibilidad.

Los resultados, año tras año, ofrecen una visión profunda de la capacidad de nuestro modelo para predecir con precisión. Sin embargo, es esencial destacar que la precisión no es el único indicador de un buen modelo. También es crucial considerar la interpretabilidad y la relevancia de las características que influencian las predicciones. Por ello, también analizaremos los coeficientes del modelo para entender qué características tienen mayor influencia en las predicciones con la cuarta estrategia de 2022, ya que es el año que cuenta con más datos a la hora de predecir los podios. A continuación, presentamos en la Figura 40, los coeficientes más relevantes en términos absolutos obtenidos de la cuarta estrategia.

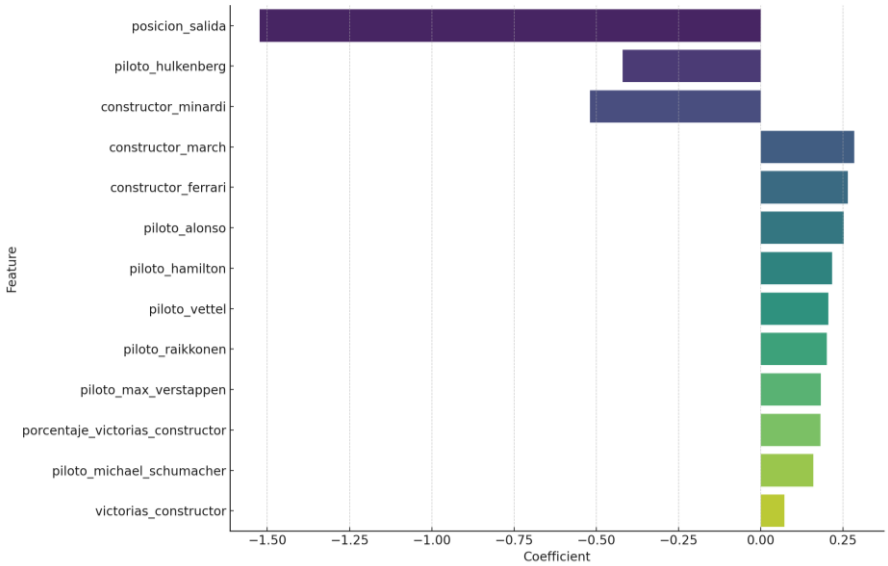

Figura 40: Coeficientes destacables de la regresión logística en la predicción de 2022

La posición de salida emerge como una de las características que más influyen en nuestro modelo, con un coeficiente de -1.52. Este valor negativo indica una relación inversa clara: cuanto más alta (o peor) es la posición de salida, menor es la probabilidad de que un piloto alcance el podio. Esta observación refuerza la importancia fundamental de las calificaciones en las carreras, ya que una posición de inicio favorable puede ser un precursor crucial para un buen rendimiento en la carrera.

Más allá de la posición inicial, la identidad del piloto y del constructor son factores significativos en la predicción del éxito en una carrera. Por ejemplo, tener a Fernando Alonso como piloto, con un coeficiente de 0.25, sugiere un aumento notable en las probabilidades de alcanzar el podio en comparación con un piloto promedio. Este valor resalta la habilidad y el historial exitoso de Alonso en la Fórmula 1. En contraste, pilotos como Nico Hulkenberg tienen un coeficiente negativo, en este caso -0.42, lo que implica que tiene menos probabilidades de alcanzar el podio en comparación con el piloto promedio. Esto puede ser debido a que este piloto ha tenido numerosas oportunidades de conseguir un podio cuando otros en su condición la han conseguido y no ha aprovechado la oportunidad.

Una vez hemos realizado este análisis sobre los pilotos, hemos querido conocer históricamente el coeficiente de los pilotos más exitosos y considerados los mejores, a lo largo de sus carreras. A continuación, se muestra en la Figura 41 la evolución de sus coeficientes a lo largo de la historia:

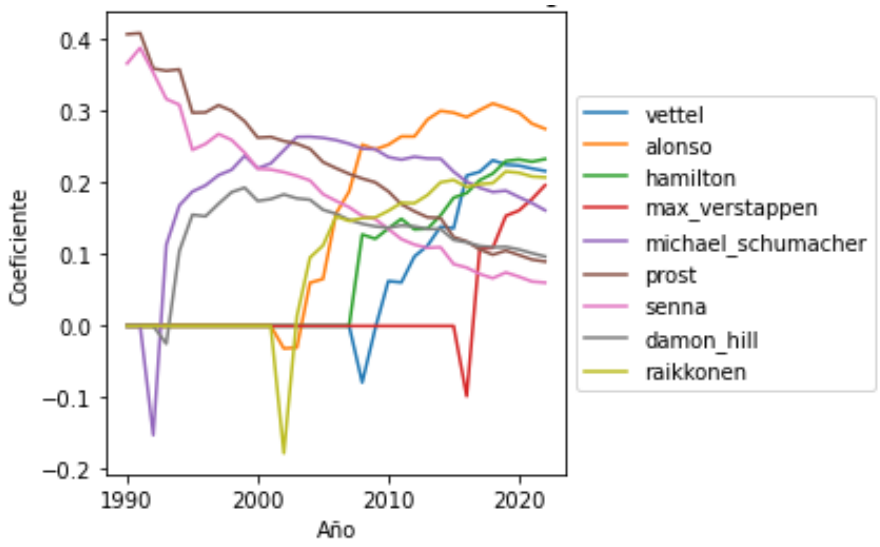

Figura 41: Evolución de los coeficientes de pilotos históricamente

El gráfico muestra la evolución temporal de los coeficientes Beta (β) asociados a destacados pilotos de Fórmula 1, reflejando la estimación cuantitativa de su influencia en la probabilidad de conseguir un lugar en el podio. Es el coeficiente que acompaña a cada variable independiente en la ecuación de la regresión logística. Indica cuánto contribuye esa variable al logaritmo de la razón de probabilidades del evento de interés. Cada coeficiente, calculado mediante el modelo de regresión logística, representa el impacto relativo de un piloto comparado con un piloto promedio. Estos valores son el resultado de un ajuste matemático que relaciona las características de los pilotos y las circunstancias de las carreras con sus resultados, permitiendo así una interpretación del efecto neto de cada piloto en su probabilidad de éxito. La tendencia de estos coeficientes a lo largo del tiempo nos ofrece una perspectiva sobre cómo ha cambiado la contribución de cada piloto a sus resultados a lo largo de su carrera.

La evolución histórica de los coeficientes de los pilotos en la Fórmula 1 revela tendencias muy interesantes y proporciona una perspectiva única sobre el impacto de cada piloto en las predicciones de podios. Estos coeficientes, derivados del modelo de regresión logística, indican la influencia de cada piloto en la probabilidad de alcanzar un podio en comparación con un piloto promedio. Un coeficiente más alto sugiere una mayor probabilidad de éxito, convirtiéndose en una métrica valiosa para valorar el desempeño relativo de los pilotos a lo largo del tiempo.

Examinando los datos históricos, observamos que ciertos pilotos han mantenido coeficientes consistentemente altos, reflejando su destreza y éxito en el deporte. Por ejemplo, Fernando Alonso ha mostrado un incremento progresivo en su coeficiente a lo largo de los años, alcanzando un máximo de 0.31 en 2018. Esto subraya su habilidad continua y su impacto en las carreras, incluso después de muchos años en el deporte.

La evolución de los coeficientes de otros pilotos muy destacados como Lewis Hamilton y Max Verstappen también es notable. Hamilton, particularmente, ha mantenido un coeficiente elevado, lo que indica su constante presencia en el podio y su competitividad en el deporte. Verstappen, por otro lado, muestra una tendencia ascendente, reflejando su creciente éxito y prominencia en la Fórmula 1.

En resumen, la Figura 41, proporciona una representación visual de cómo han cambiado las probabilidades de éxito de estos pilotos a lo largo de sus carreras. Estos datos no solo son relevantes para los aficionados y analistas del deporte, sino también para los equipos que buscan comprender mejor las dinámicas de los pilotos y su potencial para influir en los resultados de las carreras. Este análisis, ofrece una perspectiva valiosa sobre el impacto individual de los pilotos en las predicciones de podio y destaca la importancia de considerar el talento individual.

Esta tendencia de historial y habilidad también es evidente al considerar a los constructores. Ferrari, un equipo con un rico legado en la Fórmula 1, presenta un coeficiente positivo, reflejando una mayor probabilidad de éxito. Por otro lado, equipos como Minardi, con un historial menos destacado, tienen coeficientes negativos, lo que indica menor probabilidad de lograr un podio.

Finalmente, las variables que introdujimos, como el porcentaje de victorias del constructor (con un coeficiente de 0.18) y el número victorias del equipo (con un coeficiente de 0.07), aportan una dimensión adicional al análisis. Estas variables refuerzan la idea de que el desempeño de un equipo a lo largo de la temporada tiene un impacto en las predicciones.

En resumen, estas métricas ofrecen una imagen holística del rendimiento del modelo. Si bien el modelo ha demostrado ser altamente específico al predecir pilotos que no llegarían al podio, también ha mostrado una habilidad razonable para identificar a aquellos que sí lo lograrían, como lo indica el valor de Recall. La alta precisión global refuerza la confiabilidad y robustez del modelo en el contexto de las carreras de Fórmula 1. Sin embargo, como en cualquier modelo predictivo, existe margen para la mejora y optimización.

6.10 Resultados de la técnica Random Forest

En el capítulo 5, se presentan ocho estrategias de tratamiento de datos para mejorar la precisión en la predicción de resultados en carreras de Fórmula 1. Sin embargo, para este apartado hemos seleccionado específicamente solo dos estrategias para su aplicación con la técnica de Random Forest. Esta decisión se basa en la efectividad y la relevancia de estas estrategias, como se evidencia en los resultados obtenidos en los puntos anteriores. La cuarta estrategia, que implica la eliminación de la variable diff_tiempo_clasificacion, y la octava, destacan por su lógica sólida y resultados con la Regresión Logística, se consideran las más adecuadas para maximizar la precisión de las predicciones con Random Forest. La elección de estas estrategias se basa, por tanto, en sus respectivos enfoques hacia la variable 'diff_tiempo_clasificacion', ya que ha demostrado ser crucial para los resultados de las predicciones.

Al centrarnos en estas estrategias, este estudio busca optimizar los recursos y esfuerzos, enfocándose en las metodologías que han demostrado ser más efectivas y pertinentes para el contexto de la Fórmula 1. La cuarta estrategia, que excluye la variable 'diff_tiempo_clasificacion', nos permitió explorar la eficacia de un modelo más simplificado. Los resultados obtenidos son:

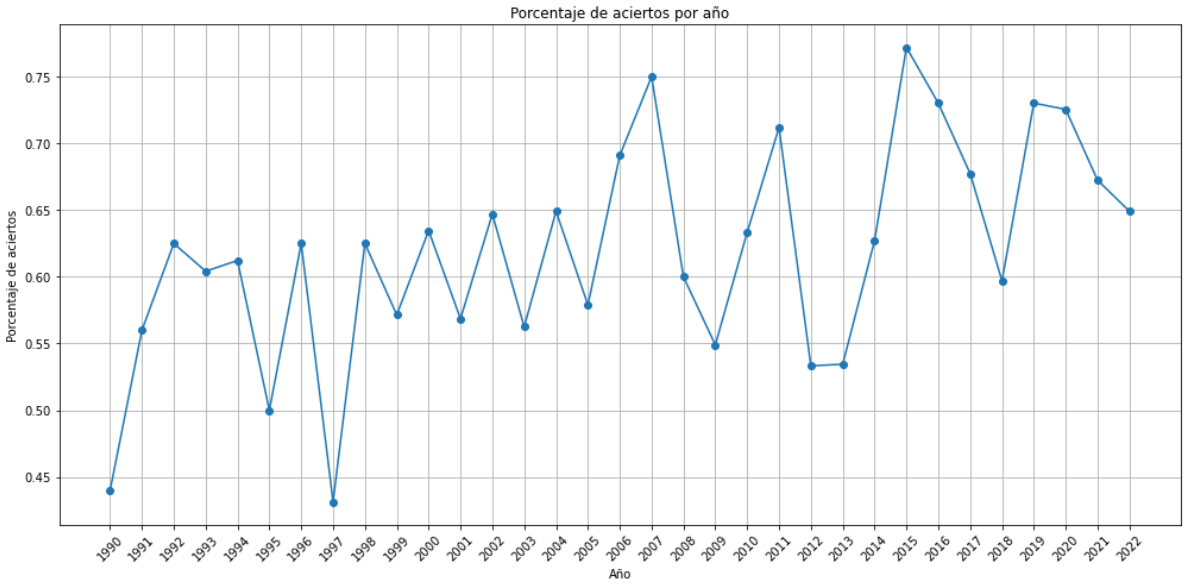

Figura 42: Precisión de las predicciones anuales con la cuarta estrategia y Random Forest

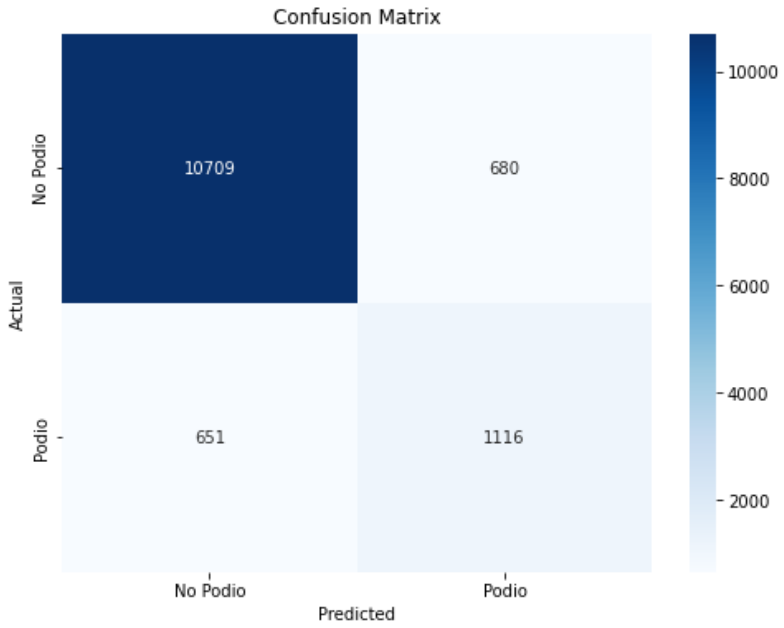

Figura 43: Matriz de confusión con los resultados de la cuarta estrategia y Random Forest

Accuracy = 0.8988 **Precision**= 0.6214 **Recall** = 0.6316 **Specificity** = 0.9403

Las métricas de calidad indican una mejora notable en comparación con los análisis anteriores, destacando un incremento en los verdaderos positivos. Sin embargo, la ausencia de 'diff_tiempo_clasificacion' plantea la cuestión de si podríamos lograr una mayor precisión incluyéndola.

La octava estrategia, que rellena los valores nulos de 'diff_tiempo_clasificacion' con el tiempo máximo de la ronda y la temporada y maneja los valores extremos de manera conservadora, nos brinda una perspectiva más completa. Los resultados son:

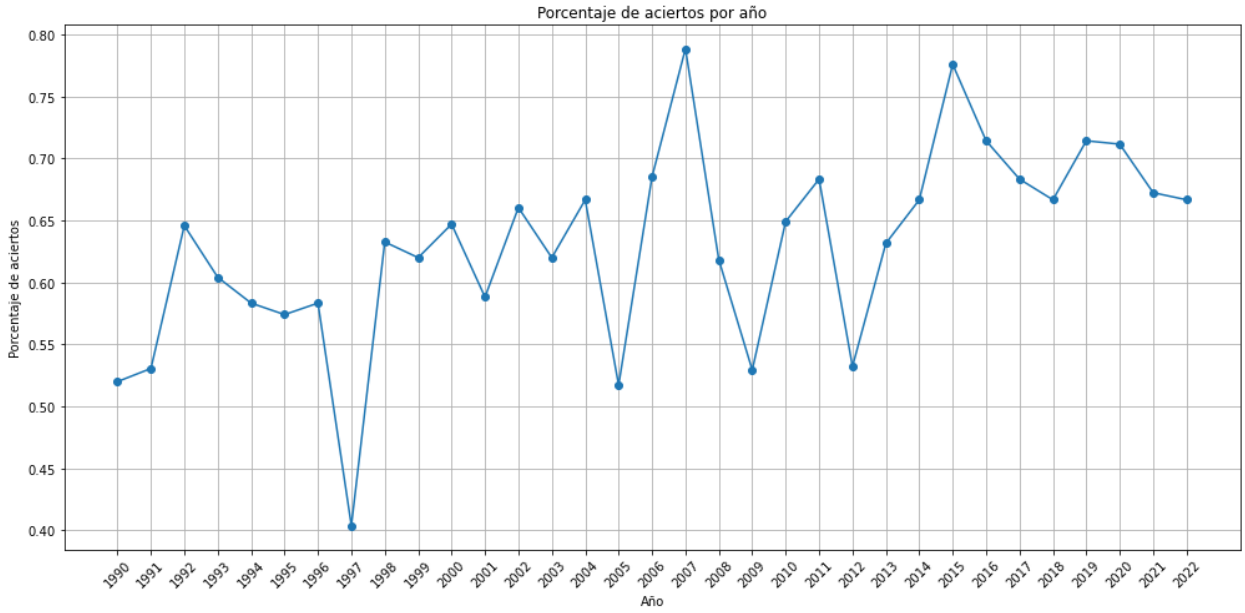

Figura 44: Precisión de las predicciones anuales con la octava estrategia y Random Forest

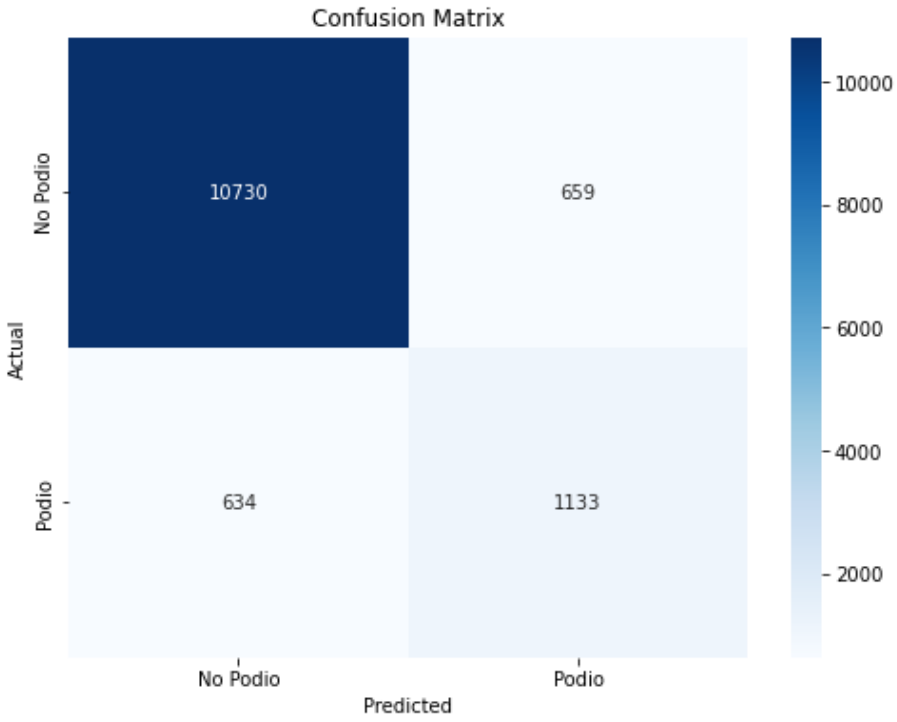

Figura 45: Matriz de confusión con los resultados de la octava estrategia y Random Forest

Accuracy = 0.9017 **Precision**= 0.6323 **Recall** = 0.6412 **Specificity** = 0.9421

Estos resultados muestran una mejora significativa en cuanto a todas las métricas y en el número de verdaderos positivos. La inclusión y el tratamiento específico de 'diff_tiempo_clasificacion' parece haber contribuido positivamente a la eficacia del modelo. Los resultados indican que la inclusión y el manejo cuidadoso de esta variable, como en la octava estrategia, son cruciales para mejorar las predicciones. La técnica de Random Forest,

conocida por su capacidad para manejar grandes conjuntos de datos y variables complejas, parece beneficiarse de un enfoque más detallado y matizado hacia esta variable.

	3ª estr. *Regr. Log*	***4ª estr.*** *Regr. Log*	***8ª estr.*** *Regr. Log*	***4ª estr.*** *Random Forest*	***8ª estr.*** *Random Forest*
Prec. 10	66.99%	67.39%	67.21%	67.18%	**69.09%**
Num. Fil.	15317	**16207**	**16207**	**16207**	**16207**
VP	1053	1052	1051	1116	**1133**
Accuracy	88.69%	89.13%	89.12%	89.88	**90.17%**
Precision	59.61%	59.54%	59.48%	62.14%	**63.23%**
Recall	60.07%	59.54%	59.48%	63.16%	**64.12%**
Specifity	93.36%	93.72%	93.71%	94.03%	**94.21%**

Tabla 6: Resultados críticos de las distintas de tratamiento de datos

Los resultados de la aplicación de las estrategias cuarta y octava con la técnica Random Forest reflejan una mejora significativa en la precisión de las predicciones en carreras de F1. La cuarta estrategia, al eliminar 'diff_tiempo_clasificacion', simplifica el modelo y muestra un aumento en los verdaderos positivos, con un Accuracy del 89.88% y un Precision del 62.14%. La octava estrategia, que maneja con ingenio y cuidado la variable 'diff_tiempo_clasificacion', ofrece un enfoque más completo y mejora todas las métricas, con un Accuracy del 90.17% y un Precision del 63.23%, lo que subraya la importancia de esta variable en las predicciones. Estos resultados apoyan la elección de estas estrategias para el uso en Random Forest, destacando su eficacia en el manejo de conjuntos de datos complejos en el contexto de la Fórmula 1.

En conclusión, la mejora en la precisión y en los verdaderos positivos con la última estrategia sugiere que, aunque su eliminación simplifica el modelo, la inclusión y el manejo adecuado de esta variable aportan una ventaja significativa en la precisión predictiva. La octava estrategia, aplicada en conjunto con la técnica de Random Forest, se presenta como un enfoque más efectivo para predecir los resultados de las carreras de Fórmula 1.

7 CONCLUSIONES

En este capítulo, recapitularemos y evaluaremos la trayectoria de nuestro proyecto desde su concepción hasta los resultados alcanzados, reflexionando sobre cómo se alinean estos con nuestros objetivos iniciales.

El nacimiento de nuestro proyecto comenzó por el creciente interés del deporte tecnológico de la Fórmula 1, una disciplina que combina velocidad, precisión y estrategia. Nuestro objetivo primordial era desarrollar un modelo predictivo avanzado para los resultados de las carreras utilizando técnicas de Machine Learning. Estábamos motivados no solo por el desafío técnico, sino también por el deseo de entender mejor las complejas dinámicas que influyen en este deporte. Aspirábamos a crear un modelo que no solo predijera los resultados con alta precisión, sino que también nos permitiera extraer nuevas perspectivas sobre cómo diversos factores, desde las condiciones climáticas hasta las decisiones estratégicas de los equipos, afectan los resultados de las carreras.

Nuestro primer paso fue sumergirnos en el mundo de la Fórmula 1. Realizamos un análisis exhaustivo de la historia del deporte, las reglas, los cambios técnicos y la evolución de las estrategias de carrera a lo largo de los años. Esta investigación nos permitió comprender no solo los aspectos técnicos y estadísticos del deporte, sino también su contexto histórico y cultural. Identificamos las variables clave que podrían influir en los resultados de las carreras y formulamos hipótesis sobre cómo estas podrían ser modeladas de manera efectiva.

La recolección de datos fue un gran esfuerzo. Nos enfrentamos a la tarea de consolidar un vasto conjunto de datos de múltiples fuentes, incluyendo bases de datos deportivas, registros históricos de carreras y reportes técnicos. La integridad y precisión de los datos eran cruciales, ya que cualquier error o inexactitud podría sesgar nuestro modelo. La limpieza de datos implicó lidiar con incoherencias, datos faltantes y anomalías. Este proceso meticuloso aseguró que la calidad de nuestros datos fuera del más alto nivel, lo que es esencial para el éxito de cualquier proyecto de Machine Learning.

La selección de características fue un proceso detallado y reflexivo. Analizamos una amplia variedad de factores, incluyendo aspectos técnicos como la configuración del coche, factores humanos como la experiencia y habilidad del piloto, y variables externas como la localización y las características de la pista. Utilizamos técnicas avanzadas de análisis de datos para identificar las correlaciones y los patrones subyacentes. Este proceso no solo fue crucial para nuestro modelo, sino que también nos proporcionó una comprensión más profunda de los factores que influyen en el rendimiento en las carreras.

La selección de modelos de Machine Learning fue una tarea estratégica. Comenzamos con modelos más simples como la regresión logística, avanzando gradualmente hacia algoritmos más complejos como Random Forest. Cada modelo se evaluó en función de su capacidad para manejar la complejidad de los datos y su precisión predictiva. A través de un proceso iterativo de prueba y error, refinamos nuestros modelos, ajustamos parámetros y evaluamos el rendimiento utilizando una variedad de métricas. Este enfoque nos permitió no solo mejorar la

precisión de nuestras predicciones, sino también entender mejor cómo diferentes modelos procesan y valoran las variables involucradas.

Los resultados de nuestro modelo fueron reveladores. Logramos no solo una alta precisión en las predicciones, sino que también obtuvimos *insights* profundos sobre la importancia relativa de las variables. Nuestro modelo fue capaz de identificar patrones y tendencias que no eran inmediatamente evidentes. Las métricas de rendimiento, como la precisión, el recall y la sensibilidad, mostraron que nuestro modelo era robusto y confiable en diferentes escenarios de carrera. Estos resultados validaron nuestra hipótesis inicial y demostraron la efectividad de nuestras técnicas de modelado en un entorno deportivo complejo.

Al comparar nuestros resultados con los objetivos iniciales, debemos estar muy satisfechos de lo logrado. El modelo no solo cumplió con su propósito de predecir los resultados con alta precisión, sino que también nos proporcionó una comprensión más profunda de las dinámicas de la Fórmula 1. A pesar de los desafíos enfrentados, como adaptarse a situaciones inesperadas y la integración de datos en tiempo real, estos obstáculos solo sirvieron para profundizar nuestro conocimiento y mejorar nuestro enfoque.

Este proyecto ha sido un hito en nuestro desarrollo profesional y académico. Aprendimos no solo sobre la aplicación práctica de Machine Learning, sino también sobre la importancia de un enfoque meticuloso y detallado en la investigación y análisis de datos. Mirando hacia el futuro, nos surge la curiosidad por las posibilidades de expandir nuestro trabajo. Esto incluye la integración de datos en tiempo real, la exploración de modelos de aprendizaje profundo y la aplicación de redes neuronales. Este proyecto no solo cumplió con nuestros objetivos, sino que también puede aportar interesantes detalles para futuros estudios académicos o incluso profesionales en el campo de la analítica deportiva y el Machine Learning.

A lo largo de este proyecto, aprendimos lecciones valiosas sobre el manejo de grandes conjuntos de datos, la importancia de la limpieza y preparación de datos, y la aplicación efectiva de modelos de Machine Learning. Identificamos áreas de mejora, especialmente en el tratamiento de datos, en la adaptación a situaciones imprevistas como datos irreales, sin sentido o incluso faltantes. Estas lecciones no solo son relevantes para nuestro campo de estudio, sino que también son aplicables a una amplia gama de problemas en la ciencia de datos y la inteligencia artificial.

Nuestro proyecto tiene el potencial de impactar positivamente en la comunidad de la Fórmula 1 y más allá. Nuestro modelo podría ser una herramienta valiosa para equipos y aficionados, proporcionando análisis predictivos y estratégicos para mejorar el rendimiento y la experiencia de carrera. Además, nuestras metodologías y hallazgos pueden ser de gran valor para la comunidad científica, ofreciendo un caso de estudio en la aplicación de Machine Learning en un entorno deportivo dinámico.

Para concluir, este proyecto ha sido un reto muy enriquecedor de aprendizaje y descubrimiento. Logramos nuestros objetivos y, en el proceso, avanzamos en nuestro entendimiento de cómo la tecnología de ML puede aplicarse en el deporte. Este trabajo no solo representa un logro significativo en nuestra formación académico, sino que también es un paso adelante en la integración de la tecnología avanzada en el análisis deportivo.

Las futuras líneas de trabajo deberían considerar la integración de datos en tiempo real para ofrecer análisis más dinámicos y precisos. Además, podrían explorar la ampliación de la base de datos para evaluar si la inclusión de datos más recientes mejora aún más la precisión del modelo en los años siguientes. El empleo de modelos de aprendizaje profundo, como las redes neuronales convolucionales (CNN), podría ser particularmente útil para interpretar datos visuales de las carreras, como el desgaste de los neumáticos o las condiciones de la pista. Además, aplicar algoritmos que consideren secuencias temporales, como las redes neuronales recurrentes (RNN) y las LSTM, podría mejorar la predicción de eventos dinámicos dentro de una carrera. Finalmente, la exploración de la Inteligencia Artificial General (AGI) podría ofrecer estrategias de carrera innovadoras, tomando decisiones en fracciones de segundo similares a las de un piloto o un estratega de equipo.

REFERENCIAS

Amazon Web Services. (n.d.). F1 Insights con tecnología de AWS | Estrategia alternativa. Retrieved from https://aws.amazon.com/es/sports/f1/

Auto Hebdo. (2023). $2.5 billion in revenue, record attendance: F1 figures in 2022. Retrieved from https://www.autohebdof1.com/actualites/f1/25-mds-de-dollars-de-revenus-frequentation-record-les-chiffres-de-la-f1-en-2022.html

AutoBild. (2021). Coches de F1 2022: El nuevo reglamento que lo cambiará todo. Retrieved from https://www.autobild.es/noticias/coches-f1-2022-llega-nuevo-reglamento-899503

Babu, A. (2020, July 3). Towards Data Science. Retrieved from https://towardsdatascience.com/optimization-in-transportation-problem-f8137044b371

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. Applied Computing and Informatics, 15(1), 27-33. https://www.sciencedirect.com/science/article/pii/S2210832717301485

Car and Driver. (n.d.). Los 30 mejores pilotos de Fórmula 1 de la historia. Retrieved from https://www.caranddriver.com/es/formula-1/g39725477/mejores-pilotos-formula-1/

Codling, S. (2017). Speed read F1: The technology, rules, history and concepts key to the sport. Motorbooks. Retrieved from https://books.google.es/books?hl=es&lr=&id=ec85DwAAQBAJ&oi=fnd&pg=PP1&dq=formula+1+racing+evolution+of+technology&ots=xOv0AuwJtI&sig=Q27U7EqcNuRq9mP790VEdcgKGxw

Data Center Dynamics. (n.d.). AWS y la Fórmula 1 renuevan su asociación para impulsar la innovación. Retrieved from https://www.datacenterdynamics.com/es/noticias/aws-y-la-f%C3%B3rmula-1-renuevan-su-asociaci%C3%B3n-para-impulsar-la-innovaci%C3%B3n/

Ergast Developer API. (n.d.). A public open source Formula One Database. Retrieved from https://ergast.com/mrd/open

Formula 1. (n.d.). The Official Home of Formula 1® Racing. Retrieved from https://www.formula1.com/

Foxall, G. R., & Johnston, B. R. (1991). Innovation in Grand Prix motor racing: the evolution of technology, organization and strategy. Technovation, 11(7), 387-402. https://www.sciencedirect.com/science/article/pii/0166497291900205

Friligkos, G., Papaioannou, E., & Kaklamanis, C. (2023). A framework for applying the Logistic Regression model to obtain predictive analytics for tennis matches. Technium Science. Retrieved from https://techniumscience.com/index.php/technium/article/view/9616

FundéuRAE. (2018). «Aprendizaje automático», mejor que «machine learning». Retrieved from https://www.fundeu.es/recomendacion/aprendizaje-automatico-mejor-que-machine-learning

Gao, Z., & Kowalczyk, A. (2021). Random forest model identifies serve strength as a key predictor of tennis match outcome. Journal of Sports Analytics, 7(4), 255-262. Retrieved from https://content.iospress.com/articles/journal-of-sports-analytics/jsa200515

International Automobile Federation (FIA). (n.d.). FIA FORMULA ONE WORLD CHAMPIONSHIP Regulation. Retrieved from https://www.fia.com/regulation/category/110

Jenkins, M., & Floyd, S. (2001). Trajectories in the evolution of technology: A multi-level study of competition in Formula 1 racing. Organization Studies, 22(6), 945-969. Retrieved from https://journals.sagepub.com/doi/abs/10.1177/0170840601226003

Kovalchik, S.A. (2023). Player tracking data in sports. Annual Review of Statistics and Its Application, 10, 677-697. Retrieved from https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-033021-110117

Lalwani, A., Saraiya, A., Singh, A., Jain, A., & Dash, T. (2022). Machine learning in sports: A case study on using explainable models for predicting outcomes of volleyball matches. Retrieved from https://arxiv.org/abs/2206.09258

Löckel, S. A. (2022). Machine learning for modeling and analyzing of race car drivers. Retrieved from https://tuprints.ulb.tu-darmstadt.de/20218/

Marca. (2020). Análisis histórico de pilotos de Fórmula 1. Retrieved from https://www.marca.com/blogs/master-big-data-deportivo/2020/10/26/analisis-historico-de-pilotos-de-formula.html

Marinaro, D. (2021). Lo Stratega: Predicting Formula 1 race strategies with reinforcement learning. Retrieved from https://www.politesi.polimi.it/handle/10589/210056

Marino, A., Aversa, P., Mesquita, L., & Anand, J. (2015). Driving performance via exploration in changing environments: Evidence from Formula One racing. Organization Science, 26(4), 1079-1100. Retrieved from https://pubsonline.informs.org/doi/abs/10.1287/orsc.2015.0984

Patil, A., Jain, N., Agrahari, R., Hossari, M., Orlandi, F., & Dev, S. (2022, December). A data-driven analysis of Formula 1 car races outcome. In Irish Conference on Artificial Intelligence and Cognitive Science (pp. 134-146). Cham: Springer Nature Switzerland. Retrieved from https://link.springer.com/chapter/10.1007/978-3-031-26438-2_11

Piccolomini, E. L., Evangelista, D., & Rondelli, M. (2022). The future of Formula 1 racing: Neural networks to predict tyre strategy. Retrieved from https://amslaurea.unibo.it/27922/1/Tesi_Massimo_Rondelli.pdf

Remonda, A., Krebs, S., Veas, E., Luzhnica, G., & Kern, R. (2021). Formula RL: Deep reinforcement learning for autonomous racing using telemetry data. Retrieved from https://arxiv.org/abs/2104.11106

Scikit-learn developers. (n.d.). About us - scikit-learn 0.24.1 documentation. Retrieved from https://scikit-learn.org/stable/about.html

Shields, B., & Reavis, C. (2020). Formula 1: Unleashing the greatest racing spectacle on the planet. MIT Sloan Working Paper. Retrieved from https://mitsloan.mit.edu/sites/default/files/2021-04/Formula%201.Unleashing%20the%20Greatest%20Spectacle%20on%20the%20Planet.IC_.pdf

Sicoie, H. (2022). Machine learning framework for Formula 1 race winner and championship standings predictor (Doctoral dissertation, Tilburg University). Retrieved from https://arno.uvt.nl/show.cgi?fid=157635

Silverstone Circuit. (2023). The 2023 British Grand Prix in numbers. Retrieved from https://www.silverstone.co.uk/news/2023-british-grand-prix-numbers

SoyMotor. (2022). DAZN estrena el reportaje '2005, el año que cambió nuestra vida', con Lobato y De la Rosa. Retrieved from https://soymotor.com/noticias/dazn-estrena-reportaje-2005-ano-cambio-nuestra-vida-lobato-de-la-rosa-94145

Sports Pro Media. (2021). F1's global TV audience hits 1.55bn in 2021. Retrieved from https://www.sportspromedia.com/news/f1-global-tv-audience-2021/?zephr_sso_ott=9wn7IX

Young, S. (2012). Formula One racing: Driver vs. technology. Intersect: The Stanford Journal of Science, Technology, and Society, 5. Retrieved from https://ojs.stanford.edu/ojs/index.php/intersect/article/view/349

ANEXO. CÓDIGOS DE PYTHON

Código para la recolección de datos

```
import pandas as pd
import numpy as np
import requests

# PRIMERA consulta API en repositorio online ergast F1: contiene información
sobre todos los campeonatos y
# carreras desde 1950 hasta 2022, incluida su ubicación y enlace a la página
de wikipedia

carreras = {'temporada': [],
            'ronda': [],
            'id_circuito': [],
            'lat': [],
            'long': [],
            'pais': [],
            'fecha': [],
            'url': []}

for year in range(1950, 2023): # Se obtiene información sobre las carreras de
Fórmula 1 desde 1950 hasta 2023
    url = f'https://ergast.com/api/f1/{year}.json'
    r = requests.get(url)
    json = r.json()

    for item in json['MRData']['RaceTable']['Races']:
        carreras['temporada'].append(int(item.get('season', None)))
        carreras['ronda'].append(int(item.get('round', None)))
        carreras['id_circuito'].append(item['Circuit'].get('circuitId',
None))

        location = item['Circuit'].get('Location', {})
        carreras['lat'].append(float(location.get('lat', None)))
        carreras['long'].append(float(location.get('long', None)))
        carreras['pais'].append(location.get('country', None))

        carreras['fecha'].append(item.get('date', None))
        carreras['url'].append(item.get('url', None))

carreras = pd.DataFrame(carreras)

# SEGUNDA Consulta API en repositorio online ergast F1: obteniene información
sobre los resultados de todos
# los pilotos. Incluí características como la parrilla y la posición final de
cada piloto, sus equipos y otras
# variables menos relevantes como la fecha de nacimiento, la nacionalidad y
el estado de acabado

# Añade el número de rondas a cada temporada a partir de carreras_df
rondas = []
```

```
for year in np.array(carreras.temporada.unique()):
    rondas.append([year, list(carreras[carreras.temporada ==
year]['ronda'])])

# Query
resultados = {'temporada': [],
              'ronda': [],
              'id_circuito': [],
              'piloto': [],
              'fecha_nacimiento': [],
              'nacionalidad': [],
              'constructor': [],
              'posicion_salida': [],
              'tiempo': [],
              'estado': [],
              'puntos': [],
              'posicion_final': []}

for n in range(len(rondas)):
    for i in rondas[n][1]:

        url = f'http://ergast.com/api/f1/{rondas[n][0]}/{i}/results.json'
        r = requests.get(url)
        json = r.json()
        carrera = json['MRData']['RaceTable']['Races'][0]

        for item in carrera['Results']:
            resultados['temporada'].append(int(carrera.get('season', None)))
            resultados['ronda'].append(int(carrera.get('round', None)))

resultados['id_circuito'].append(carrera['Circuit'].get('circuitId', None))

            driver = item.get('Driver', {})
            resultados['piloto'].append(driver.get('driverId', None))
            resultados['fecha_nacimiento'].append(driver.get('dateOfBirth',
None))
            resultados['nacionalidad'].append(driver.get('nationality',
None))


resultados['constructor'].append(item['Constructor'].get('constructorId',
None))
            resultados['posicion_salida'].append(int(item.get('grid', None)))
            try:
                resultados['tiempo'].append(int(item['Time']['millis']))
            except:
                resultados['tiempo'].append(None)
            # Intenta agregar el tiempo del piloto en milisegundos. Si no
está disponible, agrega None.
            # El tiempo del piloto puede no estar disponible en algunos
casos, como cuando el piloto no
            # completa la carrera o se retira. Por lo tanto, utilizamos un
bloque try-except para manejar
            # estos casos y asegurarnos de que no haya errores al agregar los
tiempos.

            resultados['estado'].append(item.get('status', None))
```

```
            resultados['puntos'].append(int(float(item.get('points', None))))
            # Convierte primero el valor a float y luego a int para evitar
errores con valores como '8.5'
            resultados['posicion_final'].append(int(item.get('position',
None)))

resultados_df = pd.DataFrame(resultados)

import requests
import pandas as pd

# TERCERA Consulta API en repositorio online ergast F1: proporciona el número
de puntos, las victorias y
# la posición de cada piloto a lo largo del Campeonato

clasificacion_pilotos = {'temporada': [],
                         'ronda': [],
                         'piloto': [],
                         'puntos_piloto': [],
                         'victorias_piloto': [],
                         'posicion_clasificacion': [],
                         'constructor': []}

with requests.Session() as session:  # Usar una sesión para hacer las
solicitudes
    for temporada, lista_rondas in rondas:
        for ronda in lista_rondas:
            print(f"Procesando temporada {temporada} - Ronda {ronda}")

            url =
f'https://ergast.com/api/f1/{temporada}/{ronda}/driverStandings.json'
            response = session.get(url)
            json_data = response.json()

            clasificacion =
json_data['MRData']['StandingsTable']['StandingsLists'][0]

            for item in clasificacion['DriverStandings']:
                clasificacion_pilotos['temporada'].append(temporada)
                clasificacion_pilotos['ronda'].append(ronda)

                driver = item['Driver']
                clasificacion_pilotos['piloto'].append(driver['driverId'])

clasificacion_pilotos['puntos_piloto'].append(int(float(item['points'])))

clasificacion_pilotos['victorias_piloto'].append(int(item['wins']))

clasificacion_pilotos['posicion_clasificacion'].append(int(item['position']))

                constructor = item['Constructors'][0]

clasificacion_pilotos['constructor'].append(constructor['constructorId'])

clasificacion_pilotos_df = pd.DataFrame(clasificacion_pilotos)
```

```
# Como los puntos se otorgan después de la carrera, he creado una función de
búsqueda para desplazar los
# puntos de las carreras anteriores dentro del mismo Campeonato

def buscar(df, equipo, puntos):
    # Combinar temporada, equipo y ronda en una sola cadena para crear dos
columnas de búsqueda
    df['busqueda1'] = df.temporada.astype(str) + df[equipo] +
df['ronda'].astype(str)
    df['busqueda2'] = df.temporada.astype(str) + df[equipo] + (df['ronda'] -
1).astype(str)

    # Combinar el DataFrame consigo mismo utilizando las columnas de búsqueda
    nuevo_df = df.merge(df[['busqueda1', puntos]], how='left',
left_on='busqueda2', right_on='busqueda1')

    # Eliminar columnas innecesarias
    nuevo_df.drop(['busqueda1_x', 'busqueda2', 'busqueda1_y'], axis=1,
inplace=True)

    # Renombrar columnas
    nuevo_df.rename(columns={puntos + '_x': puntos + '_despues_carrera',
puntos + '_y': puntos}, inplace=True)

    # Rellenar valores NaN con 0
    nuevo_df[puntos].fillna(0, inplace=True)

    return nuevo_df

# Utilizamos la función buscar() para buscar la información antes de la
carrera
clasificacion_pilotos_df = buscar(clasificacion_pilotos_df, 'piloto',
'puntos_piloto')
clasificacion_pilotos_df = buscar(clasificacion_pilotos_df, 'piloto',
'victorias_piloto')
clasificacion_pilotos_df = buscar(clasificacion_pilotos_df, 'piloto',
'posicion_clasificacion')

# Eliminar columnas innecesarias
clasificacion_pilotos_df.drop(['puntos_piloto_despues_carrera',
                               'victorias_piloto_despues_carrera',
                               'posicion_clasificacion_despues_carrera'],
                              axis=1, inplace=True)



# CUARTA consulta API del repositorio online ergast F1: Query para los
equipos es el mismo que el de
# la clasificación de pilotos, aplicando finalmente la misma función de
búsqueda para obtener los datos
# antes de la carrera

# Comienza desde el año 1958 ya que es desde cuando se premia a los equipos
rondas_constructor = rondas[8:]

clasificacion_constructores = {'temporada': [],
                               'ronda': [],
```

```
                                'constructor': [],
                                'puntos_constructor': [],
                                'victorias_constructor': [],
                                'posicion_clasificacion_constructor': []}

with requests.Session() as session:  # Usar una sesión para hacer las
solicitudes
    for temporada, lista_rondas in rondas_constructor:
        for ronda in lista_rondas:
            print(f"Procesando temporada {temporada} - Ronda {ronda}")

            url =
f'https://ergast.com/api/f1/{temporada}/{ronda}/constructorStandings.json'
            response = session.get(url)
            json_data = response.json()

            clasificaciones =
json_data['MRData']['StandingsTable']['StandingsLists'][0]

            for item in clasificaciones['ConstructorStandings']:
                clasificacion_constructores['temporada'].append(temporada)
                clasificacion_constructores['ronda'].append(ronda)

clasificacion_constructores['constructor'].append(item['Constructor']['constr
uctorId'])

clasificacion_constructores['puntos_constructor'].append(int(float(item['poin
ts'])))

clasificacion_constructores['victorias_constructor'].append(int(item['wins'])
)

clasificacion_constructores['posicion_clasificacion_constructor'].append(int(
item['position']))

clasificacion_constructores_df = pd.DataFrame(clasificacion_constructores)

# Utilizamos la función buscar() para buscar la información antes de la
carrera
clasificacion_constructores_df = buscar(clasificacion_constructores_df,
'constructor', 'puntos_constructor')
clasificacion_constructores_df = buscar(clasificacion_constructores_df,
'constructor', 'victorias_constructor')
clasificacion_constructores_df = buscar(clasificacion_constructores_df,
'constructor', 'posicion_clasificacion_constructor')

# Elimina las columnas innecesarias
columns_to_drop = ['puntos_constructor_despues_carrera',
'victorias_constructor_despues_carrera',
                   'posicion_clasificacion_constructor_despues_carrera']
clasificacion_constructores_df.drop(columns_to_drop, axis=1, inplace=True)


# QUINTA Query en la pagina oficial de F1: extrae la vuelta más rapida
realizada por cada piloto en la sesión
# de clasificacion. Como en la API de Ergast faltaban algunos datos, tuve que
utilizar BeautifulSoup para
```

```
# rastrear el sitio web oficial de la F1 y añadir la tabla que se encuentra
en la página de la parrilla de
# de cada circuito.

from bs4 import BeautifulSoup
import requests
import pandas as pd

resultados_clasificacion_list = []
base_url = 'https://www.formula1.com'

# Utilizar una sesión para las requests
with requests.Session() as session:

    # Los tiempos de clasificación solo están disponibles desde 1983
    for year in range(1983, 2023):
        url = f'https://www.formula1.com/en/results.html/{year}/races.html'
        r = session.get(url)
        soup = BeautifulSoup(r.text, 'html.parser')

        # Encontrar enlaces a todos los circuitos de un año específico
        enlaces_anuales = [page.get('href') for page in soup.find_all('a',
attrs={'class': "resultsarchive-filter-item-link FilterTrigger"}) if
f'/en/results.html/{year}/races/' in page.get('href')]

        # Para cada circuito, cambiar a la página de parrilla de salida y
leer la tabla
        for n, link in enumerate(enlaces_anuales):
            link = link.replace('race-result.html', 'starting-grid.html')
            df = pd.read_html(base_url + link)[0]
            df['temporada'] = year
            df['ronda'] = n + 1
            df = df.loc[:, ~df.columns.str.contains('Unnamed')]
            resultados_clasificacion_list.append(df)

            print(f"Procesado año {year}, ronda {n + 1}")

# Concatenar todas las tablas de todos los años
resultados_clasificacion = pd.concat(resultados_clasificacion_list,
ignore_index=True)

# Renombrar columnas
column_mapping = {
    'Pos': 'posicion_salida',
    'Driver': 'piloto',
    'Car': 'constructor',
    'Time': 'tiempo_clasificacion'
}
resultados_clasificacion.rename(columns=column_mapping, inplace=True)

# Eliminar la columna del número del piloto
resultados_clasificacion.drop('No', axis=1, inplace=True)

# Fusionar los dataframes
import pandas as pd
df_fusionado = pd.merge(resultados_df, carreras, on=['temporada', 'ronda'])
```

```
df_fusionado = pd.merge(df_fusionado, clasificacion_pilotos_df,
on=['temporada', 'ronda', 'piloto'], how='left')

# Si 'constructor_y' existe, eliminarlo y renombrar 'constructor_x' a
'constructor'
if 'constructor_y' in df_fusionado.columns:
    df_fusionado.drop('constructor_y', axis=1, inplace=True)
    df_fusionado.rename(columns={'constructor_x': 'constructor'},
inplace=True)

# Continuar con las fusiones
df_fusionado = pd.merge(df_fusionado, clasificacion_constructores_df,
on=['temporada', 'ronda', 'constructor'], how='left')
df_fusionado = pd.merge(df_fusionado, resultados_clasificacion,
on=['temporada', 'ronda', 'posicion_salida'], how='left')
```

Código para filtrar y escoger datos relevantes

```
#Creamos el dataframe final para escoger las variables clave para predecir
los resultados
df_final = df_fusionado.copy()

import numpy as np

# Convertir la columna 'fecha' a datetime
df_final['fecha'] = pd.to_datetime(df_final['fecha'])

# Extraer características de 'fecha'
df_final['mes_carrera'] = df_final['fecha'].dt.month

# Convertir la columna 'fecha_nacimiento' a datetime
df_final['fecha_nacimiento'] = pd.to_datetime(df_final['fecha_nacimiento'])

# Extraer características de 'fecha_nacimiento'
df_final['ano_nacimiento'] = df_final['fecha_nacimiento'].dt.year

def determine_finish_status(status):
    if status == 'Finished' or status[0] == '+':
        return 'Acabada'
    else:
        return 'No Acabada'

df_final['Carrera'] = df_final['estado'].apply(determine_finish_status)

# Eliminamos las columna 'estado' en df_final
df_final.drop(columns=['estado'], inplace=True)

# Filtrar el DataFrame para mantener solo las filas desde 1983 en adelante
# ya que no existen registros de clasificacion anteriormente
df_final = df_final[df_final['temporada'] >= 1983].reset_index(drop=True)


# Eliminamos las siguientes rondas ya que no tienen tiempo de clasificación
registrado para ningún piloto:
for temporada, ronda in [(2021, 10), (2021, 14), (2021, 19), (2022, 4),
(2022, 11), (2022, 21)]:
```

```
    df_final.drop(df_final[(df_final['temporada'] == temporada) &
(df_final['ronda'] == ronda)].index, inplace=True)

# Calcular el tiempo de clasificacion mínimo en cada ronda
min_qualifying_time = df_final.groupby(['temporada', 'ronda',
'id_circuito'])['tiempo_clasificacion'].transform('min')

# Calcular la diferencia de tiempo solo para las filas donde
'tiempo_clasificacion' no son nulos
df_final['diff_tiempo_clasificacion'] = ((df_final['tiempo_clasificacion'] -
min_qualifying_time) / min_qualifying_time) * 100
```

Código de estratégias de tratamiento de datos

```
# CUARTA ESTRATEGIA
# Eliminar las columnas innecesarias
columnas_eliminar = ['tiempo', 'posicion_final', 'puntos',
'tiempo_clasificacion', "lat", "long", "Carrera", "edad", "pais",
"nacionalidad", 'diff_tiempo_clasificacion']
df_copy = df_copy.drop(columns=columnas_eliminar)



# Función para reemplazar los valores nulos o cero con el máximo valor + 1
para esa ronda y temporada específicas
def replace_zero_or_null_with_max_plus_one(df, column_name):
    for i, row in df.iterrows():
        if pd.isnull(row[column_name]) or row[column_name] == 0:
            # Buscar el máximo valor en la columna específica para la ronda y
temporada
            max_value = df[(df['temporada'] == row['temporada']) &
                           (df['ronda'] == row['ronda'])][column_name].max()
            # Asignar el máximo valor encontrado + 1 a la fila actual
            df.at[i, column_name] = max_value + 1 if pd.notnull(max_value)
else 1

# Rellenar con cero los valores nulos en las columnas de puntos y victorias
columns_to_fill = [
    'puntos_piloto', 'victorias_piloto', 'puntos_constructor',
    'victorias_constructor', 'porcentaje_victorias',
'porcentaje_victorias_constructor'
]
for column in columns_to_fill:
    df_copy[column] = df_copy[column].fillna(0)

# Aplicar la función a las columnas relevantes
replace_zero_or_null_with_max_plus_one(df_copy, 'posicion_salida')
replace_zero_or_null_with_max_plus_one(df_copy, 'posicion_clasificacion')
replace_zero_or_null_with_max_plus_one(df_copy,
'posicion_clasificacion_constructor')

print(df_copy.isnull().sum())
num_filas = df_copy.shape[0]
print("Número de filas en el dataframe:", num_filas)
```

```
#OCTAVA ESTRATEGIA
# Eliminar las columnas innecesarias
columnas_eliminar = ['tiempo', 'posicion_final', 'puntos',
'tiempo_clasificacion', "lat", "long", "Carrera", "edad", "pais",
"nacionalidad"]
df_copy = df_copy.drop(columns=columnas_eliminar)

def replace_zero_or_null_with_previous_or_max_plus_one(df, column_name):
    for i, row in df.iterrows():
        if pd.isnull(row[column_name]) or row[column_name] == 0:
            # Buscar el máximo valor en la columna específica para la ronda y
temporada actuales
            max_value_current = df[(df['temporada'] == row['temporada']) &
                                   (df['ronda'] ==
row['ronda'])][column_name].max()
            if pd.notnull(max_value_current):
                df.at[i, column_name] = max_value_current + 1
            else:
                # Buscar el valor en la temporada anterior para esa misma
ronda
                max_value_previous = df[(df['temporada'] == row['temporada']
- 1) &
                                         (df['ronda'] ==
row['ronda'])][column_name].max()
                if pd.notnull(max_value_previous):
                    df.at[i, column_name] = max_value_previous
                else:
                    # Asignar la última posición posible en esa ronda y
temporada
                    last_position = df[(df['temporada'] == row['temporada'])
&
                                       (df['ronda'] ==
row['ronda'])][column_name].count()
                    df.at[i, column_name] = last_position + 1

# Aplicar la función a las columnas relevantes
replace_zero_or_null_with_previous_or_max_plus_one(df_copy,
'posicion_clasificacion')
replace_zero_or_null_with_previous_or_max_plus_one(df_copy,
'posicion_clasificacion_constructor')

# Función para rellenar los valores nulos en 'diff_tiempo_clasificacion' con
el máximo valor para esa ronda y temporada
def fill_null_with_max_in_round_and_season(df, column_name):
    for i, row in df.iterrows():
        if pd.isnull(row[column_name]):
            # Buscar el máximo valor en la columna específica para la ronda y
temporada
            max_value = df[(df['temporada'] == row['temporada']) &
                           (df['ronda'] == row['ronda'])][column_name].max()
            # Asignar el máximo valor encontrado a la fila actual
            df.at[i, column_name] = max_value if pd.notnull(max_value) else 0

# Rellenar los valores nulos en 'diff_tiempo_clasificacion' con el máximo
valor para esa ronda y temporada
fill_null_with_max_in_round_and_season(df_copy, 'diff_tiempo_clasificacion')

# Rellenar con cero los valores nulos en las columnas de puntos y victorias
```

```
columns_to_fill = [
    'puntos_piloto', 'victorias_piloto', 'puntos_constructor',
    'victorias_constructor', 'porcentaje_victorias',
'porcentaje_victorias_constructor'
]
for column in columns_to_fill:
    df_copy[column] = df_copy[column].fillna(0)

print(df_copy.isnull().sum())
df_copy.to_csv('df_copy.csv', index=False)
```

Código para predecir podio mediante Logistic Regression

```
import pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt
import numpy as np

# Leer el dataframe
df_copy = df_final.copy()

# Crear una columna de podio
df_copy['podio'] = df_copy['posicion_final'].apply(lambda x: 1 if x <= 3 else
0)

# Eliminar las columnas innecesarias
columnas_eliminar = ['tiempo', 'posicion_final', 'puntos',
'tiempo_clasificacion', "lat", "long", "Carrera", "edad", "pais",
"nacionalidad"]
df_copy = df_copy.drop(columns=columnas_eliminar)

# Para la columna posicion_salida:
# Si su valor es 0, lo reemplazaremos con el máximo valor de posicion_salida
para esa ronda específica más uno.
for index, row in df_copy.iterrows():
    if row['posicion_salida'] == 0:
        max_pos = df_copy[(df_copy['temporada'] == row['temporada']) &
                                    (df_copy['ronda'] ==
row['ronda'])]['posicion_salida'].max()
        df_copy.at[index, 'posicion_salida'] = max_pos + 1

#Estrategia para rellenar otros valores nulos de las siguientes variables.
Las rellenamos con 0
columns_to_fill = [
    'puntos_piloto',
    'victorias_piloto',
    'puntos_constructor',
    'victorias_constructor',
    'porcentaje_victorias',
    'porcentaje_victorias_constructor'
]
for column in columns_to_fill:
    df_copy[column] = df_copy[column].fillna(0)
```

Código para predecir podio mediante Random Forest

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier
from sklearn.metrics import confusion_matrix
import matplotlib.pyplot as plt
import seaborn as sns


# Instanciar el normalizador
scaler = StandardScaler()

# DataFrame global para almacenar todos los resultados
global_results = pd.DataFrame()

# Bucle que itera desde 1990 hasta 2022
for year in range(1990, 2023):
    # Dividir el conjunto de datos en entrenamiento y prueba
    df_train = df_copy[df_copy['temporada'] < year]
    df_test = df_copy[df_copy['temporada'] == year]

    X_train = df_train.drop(columns='podio')
    X_test = df_test.drop(columns='podio')

    # Codificar las variables categóricas usando codificación one-hot
    columns_to_encode = ['id_circuito', 'piloto', 'constructor']
    X_train_encoded = pd.get_dummies(X_train, columns=columns_to_encode)
    X_test_encoded = pd.get_dummies(X_test, columns=columns_to_encode)

    # Asegurarse de que ambos conjuntos tengan las mismas columnas
    for col in X_train_encoded.columns:
        if col not in X_test_encoded.columns:
            X_test_encoded[col] = 0
    X_test_encoded = X_test_encoded[X_train_encoded.columns]

    # Normalizar los conjuntos de entrenamiento y prueba
    X_train_encoded = scaler.fit_transform(X_train_encoded)
    X_test_encoded = scaler.transform(X_test_encoded)

    # Entrenar el modelo Random Forest
    # Cambiar aquí para usar uno u otro modelo
    model = RandomForestClassifier()
    model.fit(X_train_encoded, df_train['podio'])

    # Predecir con el modelo
    y_pred_proba = model.predict_proba(X_test_encoded)[:, 1]

    # Crear un DataFrame con los resultados
    df_results = df_test[['temporada', 'ronda', 'piloto', 'id_circuito',
'posicion_salida']].copy()
    df_results['probabilidad_podio'] = y_pred_proba

    # Seleccionar a los tres pilotos con las mayores probabilidades para cada
ronda
```

```
    df_results['prediccion_podio'] = df_results.groupby(['temporada',
'ronda'])['probabilidad_podio'].transform(lambda x: x.nlargest(3).min() <=
x).astype(int)
    df_results['podio_finalmente'] = df_test['podio']

    # Añadir los resultados al DataFrame global
    global_results = pd.concat([global_results, df_results], axis=0)

# Fuera del bucle, calcular la precisión por año y luego la precisión global
accuracies = []
for year in range(1990, 2023):
    df_year = global_results[global_results['temporada'] == year]
    correct_ones = ((df_year['prediccion_podio'] == 1) &
(df_year['podio_finalmente'] == 1)).sum()
    total_prediccion_podio = df_year['prediccion_podio'].sum()
    if total_prediccion_podio > 0:
        accuracy = correct_ones / total_prediccion_podio
        accuracies.append(accuracy)
        print(f"Porcentaje de aciertos en {year}: {accuracy * 100:.2f}%"
```