Título

Ainé Fernández Gregorio

Agosto 2024

Indice

Chapter 1

Introducción

1.1 Objetivo

El objetivo de esta tesis es determinar la estrategia óptima de carrera para los equipos de Fórmula 1. Esto incluye identificar la vuelta ideal para realizar las paradas en pits, el tipo de neumáticos más adecuado y el número de paradas necesarias. La metodología empleada para alcanzar este objetivo consiste en el uso de modelos de regresión múltiple para estimar los tiempos de vuelta de cada coche, así como el tiempo requerido para las paradas en pits.

1.2 Descripción del problema

1.2.1 ¿Qué es la Fórmula 1?

La Fórmula 1 es la categoría reina del automovilismo y la competición de carreras de coches más prestigiosa del mundo. Actualmente, 20 pilotos y 10 equipos conforman la parrilla de la Fórmula 1, con cada equipo compuesto por dos pilotos y sus respectivos monoplazas. Los pilotos compiten por el Campeonato de Pilotos, mientras que como compañeros de equipo suman puntos para su equipo en el Campeonato de Constructores.

La primera temporada de Fórmula 1 se celebró en 1950, y la carrera inaugural tuvo lugar en el circuito de Silverstone, Gran Bretaña, un trazado que sigue siendo parte del calendario actual. Cada temporada se organiza en una serie de Grandes Premios, que se llevan a cabo en distintos circuitos alrededor del mundo.

Cada Gran Premio se desarrolla a lo largo de un fin de semana, de viernes a domingo, y sigue generalmente la siguiente estructura:

- **Prácticas Libres:** Tres sesiones de una hora cada una en las que los equipos tienen la oportunidad de ajustar la configuración de los coches y recopilar datos clave. Durante estas sesiones, los pilotos realizan vueltas rápidas para preparar la clasificación y tandas largas para evaluar el ritmo de carrera. Las Prácticas Libres 1 y 2 se celebran el viernes, mientras que la Práctica Libre 3 tiene lugar el sábado antes de la clasificación.

- **Clasificación:** El sábado, después de la última sesión de prácticas, se lleva a cabo la clasificación para determinar el orden de la parrilla de salida para la carrera del domingo. La clasificación se divide en tres sesiones: Q1, Q2 y Q3. Todos los coches salen a pista para intentar registrar su vuelta más rápida; al final de cada sesión, los cinco coches más lentos son eliminados. El piloto más rápido en Q3 gana la *pole position* y parte desde la primera posición en la carrera del domingo.

- **Carrera:** El evento principal de cada Gran Premio se celebra el domingo. La carrera suele durar entre una hora y media y dos horas, dependiendo de las interrupciones, como la salida del coche de seguridad o las banderas rojas. El número de vueltas varía según el circuito, ya que la distancia total de cada carrera debe exceder los 305 kilómetros. Al finalizar, los pilotos que terminan en las primeras diez posiciones suman puntos para el campeonato.

1.2.2 ¿Por qué nadie quiere parar en pits?

Este proyecto se centrará en la carrera principal de cada Gran Premio. Una vez que se ha completado la clasificación y los coches están en la parrilla de salida, todos los equipos y pilotos tienen un único objetivo: cruzar la línea de meta lo más rápido posible, es decir, completar la distancia de carrera en el menor tiempo posible.

En Fórmula 1, la distancia de carrera es de al menos 305 kilómetros. El número de vueltas se establece de acuerdo con el número mínimo necesario para alcanzar dicha distancia, con la excepción del circuito de Mónaco, cuya carrera cubre poco más de 260 kilómetros. Esto se debe a que Mónaco es un circuito corto y lento, lo que permite que la carrera dure entre 80 y 100 minutos.

El problema radica en que los neumáticos pierden rendimiento tras varias vueltas, lo que provoca una disminución en los tiempos de vuelta. Para contrarrestar esto, los equipos pueden optar por realizar una parada en pits y cambiar a un nuevo juego de neumáticos. Sin embargo, surge una pregunta clave: ¿por qué los equipos no desean realizar paradas en pits? La respuesta es simple: el tiempo perdido en boxes se traduce directamente en tiempo perdido en pista, lo que puede resultar en la pérdida de posiciones.

Por tanto, los equipos se enfrentan a un *trade-off*: perder tiempo realizando una parada en pits y salir con neumáticos nuevos, que en teoría ofrecerán un mejor rendimiento, o continuar con neumáticos desgastados que seguirán perdiendo rendimiento.

El costo de una parada en pits se compone de tres factores principales:

- **Inlap:** Es el tiempo de la vuelta en la que un coche ingresa a los pits. Este tiempo suele ser más lento que una vuelta normal debido a la reducción de velocidad antes de entrar al carril de pits.
- **Tiempo transcurrido en pits:** Este tiempo abarca desde el momento en que un coche entra al carril de pits (*pitlane*) hasta que sale de él. Durante este período, el coche debe desplazarse a través del *pitlane*, donde en la mayoría de los circuitos de

Fórmula 1 hay un límite de velocidad de 80 kilómetros por hora, aunque en algunos circuitos puede ser incluso menor. El tiempo transcurrido en pits incluye el tiempo total desde la entrada al *pitlane*, el tiempo que el coche pasa detenido en el *box* para el cambio de neumáticos, y el tiempo hasta que finalmente sale del *pitlane* de regreso a la pista.

- **Outlap:** Es el tiempo de la primera vuelta que un coche realiza después de salir de los pits. Esta vuelta también suele ser más lenta que una vuelta normal, ya que los neumáticos nuevos necesitan calentarse y el coche debe recuperar el ritmo de carrera.

1.2.3 La estrategia de carerra en la Fórmula 1

Aquí es donde entra la estrategia de carrera, por reglamento de la Fórmula 1, cada piloto debe usar al menos 2 compuestos de neumáticos diferentes durante la carrera, lo que esto significa es que tienen que parar al menos una vez en los pits aunque a muchos no les guste. Ha habido ocasiones donde los pilotos se meten a pits en la penúltima vuelta para evitar una descalificación de la carrera, por ejemplo el caso de Alex Albon en el Gran Premio de Australia de 2022 donde corrió 57 vueltas con el compuesto *Hard* antes de cambiar compuesto en la penúltima vuelta y logró terminar décimo scoring a point for Williams.

La estrategia de carrera se basa en elegir cuantas paradas y que compuesto poner, en la Fórmula 1 actual hay 5 compuestos diferentes, 3 para carreras en seco y 2 para lluvia:

... terminar esta sección

Explicar estatus de la pista

1.3 Trabajos previos

En esta sección se incluirá la revisión de la literatura, con una recopilación de papers y artículos relevantes. Se presentará un resumen de los avances realizados hasta ahora para abordar el problema de encontrar la estrategia óptima de carrera en la Fórmula 1.

Chapter 2

Obtención y procesamiento de datos

2.1 Extracción de datos

La Fórmula 1 es un deporte en el que la tecnología y los datos juegan un rol crucial en el desarrollo de una carrera. Durante un fin de semana de competencia, cada coche puede tener más de 250 sensores que generan información sobre la temperatura, presión, desplazamiento, inercia, etc. Todos estos datos, así como los tiempos de vuelta y resultados, están disponibles a través de la API[1] del Live Timing de la Fórmula 1.

En este trabajo de investigación, se utilizó la librería de Python FastF1 para acceder a la API y descargar los datos. Esta librería proporciona los datos en forma de DataFrames de Pandas[2], lo que permite su manipulación utilizando todas las herramientas que Pandas ofrece.

Usualmente, el flujo de trabajo con FastF1 comienza con la creación de un objeto de Sesión (`session`) utilizando la función `get_session()`. Esto permite acceder a todos los datos dentro de este objeto, que corresponde usualmente a una sesión específica de un Gran Premio. Por ejemplo, para acceder a los datos de la carrera del Gran Premio de

[1] Explicación de API
[2] Explicación de Pandas

México de 2023, se puede crear el objeto de esta sesión de la siguiente manera:

session = fastf1.get_session(2023, 'Mexico', 'Race')

session.load()

A partir de este punto, la sesión ya está creada, y se pueden descargar o analizar todos los datos de la misma. También existen diferentes indicadores para la clasificación y las tres prácticas libres. Los datos disponibles dentro de una sesión incluyen:

- **Información del evento:** Lugar, hora de inicio, fechas, etc.
- **Resultados:** Posiciones, nombres de los equipos y pilotos, estado al finalizar (finished status).
- **Timing Data:** Tiempos de vuelta, tiempos por sectores, información sobre los neumáticos y paradas en pits. Para este trabajo de investigación, estos son los datos de mayor importancia, ya que aportan la información necesaria para analizar las estrategias de carrera.
- **Estatus de la pista:** Usualmente los status posibles son: bandera verde, bandera amarilla, coche de seguridad, coche de seguridad virtual o bandera roja.
- **Telemetría:** Velocidades, revoluciones por minuto, posición en la pista, cambios de marcha, etc.

Los datos de telemetría y timing data están disponibles desde 2018 en adelante, pero es posible acceder a los datos de resultados desde la temporada de 1950 a través de la API Ergast, otra funcionalidad que ofrece la librería FastF1.

2.2 Creación de variables y limpieza de datos

Se descargaron los datos de todas las carreras de las temporadas de 2019 a 2023. Los datos se dividieron en tres partes importantes para una estrategia de carrera exitosa:

- **Tiempos de vuelta:** Incluyen tiempos de vuelta, tipo de neumático, vida del neumático, posición, número de stint, número de vuelta, piloto y equipo.
- **Información del circuito:** Incluye las características del circuito.
- **Costo de la parada en pits:** Incluye el tiempo de parada en pits, el tiempo de vuelta de la inlap y el tiempo de vuelta de la outlap.

2.2.1 Tiempos de vuelta

Vueltas rápidas

Se filtraron las vueltas de todos los pilotos mediante el método:

```
pick_quicklaps()
```

Esto garantiza que todas las vueltas sean más rápidas que el 107% de la vuelta más rápida[3] en la sesión. Filtrar las vueltas de esta manera asegura que todas las vueltas son consideradas rápidas, es decir, que no se hayan realizado bajo un coche de seguridad o una bandera amarilla. El objetivo es analizar los tiempos de vuelta de cada piloto en condiciones normales.

Tiempo de vuelta

Se transformó la variable Tiempo de vuelta (LapTime) de su estado original que era un pandas.Timedelta[4] a segundos para que fuera más fácil su manipulación.

Vueltas en seco

Un componente muy interesante que puede llegar a ser esencial en la estrategia de carrera es el clima, en especial la lluvia, por lo que la Fórmula 1 cuenta con neumáticos para estas condiciones.

[3]Regla del 107% en carrera de la F1: Para que una vuelta rápida durante la carrera sea considerada competitiva, debe estar dentro del 107% del tiempo de la vuelta más rápida registrada durante la carrera. Si una vuelta es más lenta que este umbral, no se considera competitiva para estos fines.

[4]Tipo de objeto en Pandas que representa una diferencia de tiempo

Para este trabajo de investtigación solo se considerará estrategias de carreras para carreras que no estén afectadas por la lluvia por lo que se filtró por tipo de compuesto de neumático para que solo quedaran las vueltas con compuestos de condiciones secas, es decir:

- **HARD:** Compuesto duro.
- **MEDIUM:** Compuesto medio.
- **SOFT:** Compuesto blando.

Equipos

En la Fórmula 1, es común que los equipos sufran *rebrandings* o cambios de nombre a lo largo de los años debido a diferentes patrocinadores, adquisiciones de equipos, entre otros factores. Esto puede convertirse en un problema, ya que los datos de un mismo equipo pueden estar registrados con nombres diferentes simplemente porque el equipo cambió de patrocinador en un año determinado. Para resolver este problema, se utilizaron los nombres de los equipos en 2023 para no perder información valiosa sobre estos equipos.

Nombre antiguo	**Nombre en 2023**
Toro Rosso (2006-2019)	AlphaTauri
Renault (2002-2020)	Alpine
Alfa Romeo Racing (2019-2022)	Alfa Romeo
Racing Point (2018-2020)	Aston Martin

Table 2.1: Cambio de nombres de equipos en la Fórmula 1

CAMBIAR TABLA

Longitud del stint (***StintLength***)

Un stint es el número de vueltas que un piloto realiza con los mismos neumáticos antes de hacer una parada en pits o también entre paradas. Para crear la variable *StintLength*, se agrupó la información por número de stint, piloto, tipo de neumático, Gran Premio (GP) y año.

Tiempo de vuelta por kilómetro (*LapTimePerKM*)

Dado que no todos los circuitos de la Fórmula 1 tienen la misma longitud en kilómetros, se creó la variable *LapTimePerKM*. Esta variable se calculó dividiendo el tiempo de vuelta por el número de kilómetros del circuito.

Porcentaje de carrera (*RacePercentage*)

Conocer cuántas vueltas faltan en una carrera es crucial para decidir la estrategia de carrera, por lo que se creó la variable *RacePercentage*. Esta variable se calculó dividiendo el número de la vuelta actual entre el número total de vueltas de la carrera.

2.2.2 Costo de la parada en pits

El tiempo combinado de estos tres factores es lo que los equipos deben considerar como la pérdida total durante una parada en pits, y es fundamental para planificar la estrategia de carrera.

Inlaps

Al descargar los datos de las vueltas, no existía un indicador específico para identificar las *Inlaps*. No obstante, los datos incluían la variable *PitInTime*, que registra el tiempo exacto en que un coche entra a pits. Cualquier vuelta con un valor de *PitInTime* mayor a cero fue clasificada como una *Inlap*. Posteriormente, se aplicó un filtro adicional para conservar únicamente las *Inlaps* realizadas con neumáticos de compuesto seco. Finalmente, se añadió la variable *LapTimePerKM*.

Tiempo transcurrido en pits

Los datos sobre el tiempo transcurrido en pits se obtuvieron a través de la API de Ergast, accesible también mediante la librería *FastF1*. Un desafío común es la inconsistencia en la forma en que se registran los datos entre las API. En Ergast, los tiempos transcurridos en pits están asociados al nombre del circuito, en lugar de la ciudad o el país donde se

corre la carrera, como ocurre en la otra API. Para evitar problemas en el análisis y la modelización, se reemplazaron los nombres de los circuitos por los nombres de los países o ciudades (en caso de que haya más de un Gran Premio en el mismo país) y se añadió el equipo del piloto.

Outlaps

Similar a las *Inlaps*, no existía un indicador específico para las *Outlaps*, pero se pudo utilizar la variable *PitOutTime*, que registra el tiempo en que un coche sale de pits, para identificarlas. Se filtraron todas las vueltas con un valor de *PitOutTime* mayor a cero. Después, se aplicó un filtro adicional para mantener únicamente las *Outlaps* con neumáticos de compuesto seco, y finalmente se añadió la variable *LapTimePerKM*.

2.2.3 Coches de seguridad (*Safety Cars*)

Una fuente de incertidumbre que puede afectar el diseño de una estrategia de carrera son los Coches de Seguridad (*Safety Cars*).

El *Safety Car* suele salir a la pista después de una bandera amarilla. Las banderas amarillas se despliegan cuando ocurre un incidente en la pista que representa un peligro, alertando a los pilotos para que reduzcan la velocidad. Estos incidentes pueden variar, desde un coche que se sale de la pista, un choque, hasta la presencia de escombros en la pista.

Una bandera amarilla puede escalar a un *Safety Car* cuando el incidente no puede resolverse rápidamente o representa un peligro significativo. En ese caso, el *Safety Car* entra en la pista y lidera el grupo a una velocidad reducida mientras el incidente se resuelve.

Lo interesante de los *Safety Cars* es que obligan a todos los coches a mantener una velocidad controlada, lo que generalmente incrementa los tiempos de vuelta en un 60%. Esta reducción de velocidad abre la posibilidad para que los equipos decidan realizar una parada en pits durante la presencia del *Safety Car*, sacrificando la posición en pista, pero

ganando tiempo en la parada. Si un coche tiene una ventaja considerable respecto al coche que le sigue, puede beneficiarse de una parada en pits "gratis", lo que significa que puede entrar a pits y regresar a la pista sin perder posiciones, ya que el tiempo perdido en pits disminuye debido a la velocidad reducida de los demás coches.

En los datos que se tienen cada registro de vuelta tiene la variable *TrackStatus* que contiene el estado de la pista en esa vuelta, por lo que se crearon una etiquetas de para el estado de la pista:

- ***Green Flag*:** Indicador contiene al 1.
- ***Yellow Flag*:** Indicador contiene al 2.
- ***Safety Car*:** Indicador contiene al 4.
- ***Red Flag*:** Indicador contiene al 5.
- ***Virtual Safety Car*:** Indicador contiene al 6.

Por último se creó una variable indicadora de *Safety Car*, para identificar las vueltas en donde salió el *Safety Car* más fácil.

2.2.4 Lluvia

Se tuvieron que descargar por separado las vueltas afectadas por lluvia, ya que hay carreras en las que no todas las vueltas están afectadas por esta condición. La regla del 107% descartaría todas las vueltas con neumáticos de lluvia, debido a que son más lentas que ese corte. Por lo tanto, no podemos usar el método `pick_quicklaps()`.

TERMINAR ESTA EXPLICACIÓN LLUVIA

Chapter 3

Exploración de datos

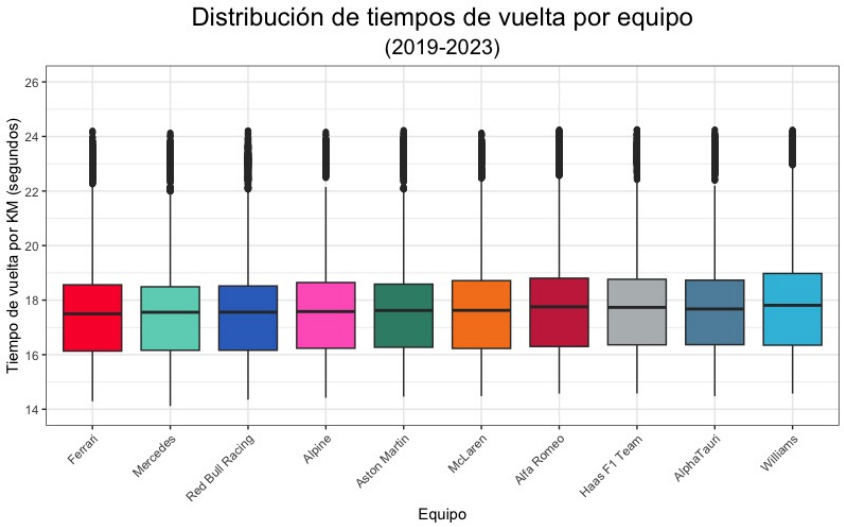
3.1 Vueltas Rápidas en Seco y Neumáticos

Como se mencionó anteriormente, solo se incluyeron en los datos para el modelo las vueltas rápidas realizadas con compuestos de neumáticos para condiciones secas. Esto resulta en un total de 64,516 registros de vueltas.

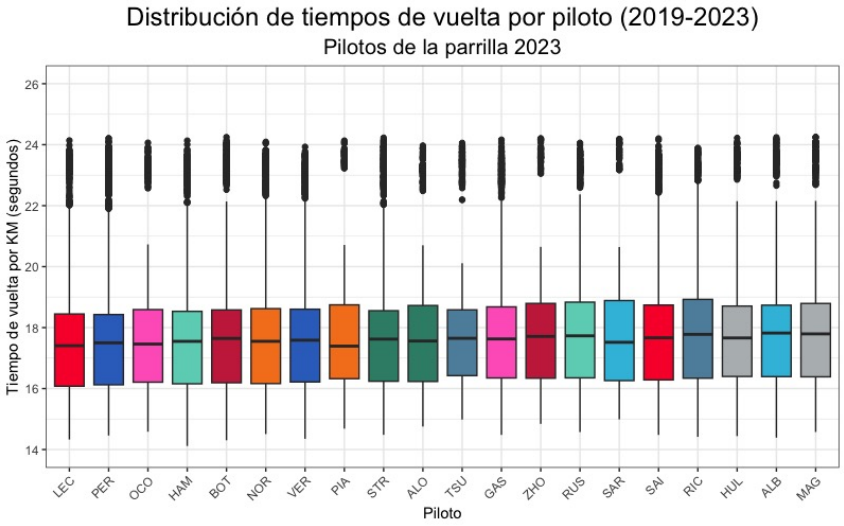
Las variables disponibles en este conjunto de datos son las siguientes:

- **Driver:** Piloto que realizó la vuelta.
- **Team:** Equipo al que pertenece el piloto.
- **LapNumber:** Número de la vuelta en la carrera.
- **LapTime:** Tiempo registrado en la vuelta.
- **Stint:** Número de stint (período de uso continuo de un conjunto de neumáticos).
- **Compound:** Tipo de neumático utilizado.
- **TyreLife:** Edad del neumático, medida en el número de vueltas que ha completado.
- **Position:** Posición del piloto al completar esa vuelta.

3.1.1 Tiempos de vuelta por Equipo y Piloto



(a) Distribución del tiempo de vuelta por KM por equipo



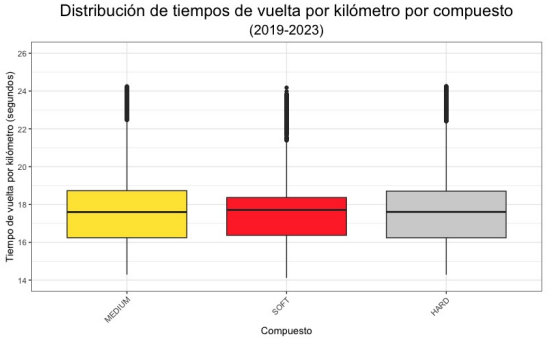
(b) Distribución del tiempo de vuelta por KM por piloto

Figure 3.1: Comparación entre tiempos de vuelta por equipo y por piloto

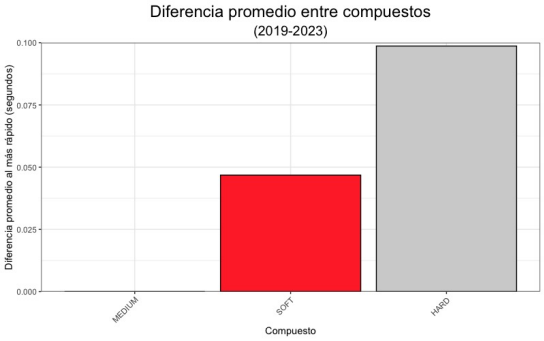
Estos diagramas de caja y brazo están ordenados por el tiempo de vuelta por kilómetro promedio. Se pueden observar diferencias en los tiempos de vuelta por kilómetro tanto entre equipos como entre pilotos. Nico Rosberg, campeón de la Fórmula 1 en 2016, ha mencionado en varias ocasiones su regla del 80-20. Según Rosberg, el 80% del éxito en Fórmula 1 se debe al coche y al equipo, mientras que el 20% se debe a las habilidades del piloto. La validez de estos porcentajes no es el foco de este proyecto de investigación; lo

relevante es que tanto el equipo como el piloto pueden influir de manera significativa en los tiempos de vuelta por kilómetro.

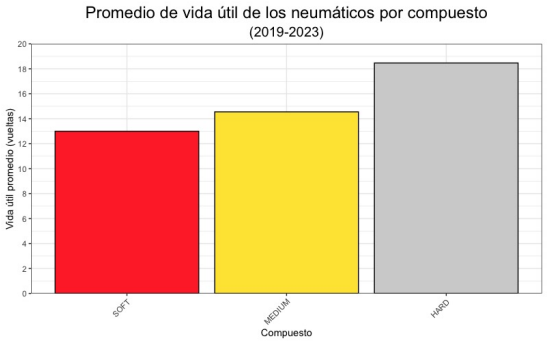
3.1.2 Neumáticos



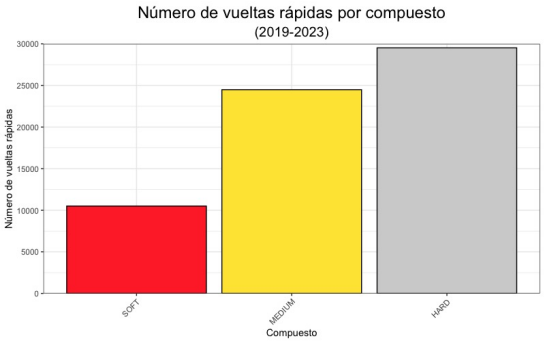
(a) Distribución del tiempo de vuelta por KM por compuesto



(b) Diferencia promedio entre compuestos



(c) Vida útil promedio por compuesto



(d) Número de vueltas por compuesto

Figure 3.2: Visión general de los neumáticos

Se observa en la gráfica (a) una clara diferencia en el tiempo de vuelta por kilómetro entre los distintos compuestos. El compuesto *Medium* es, en promedio, el más rápido durante la carrera, a pesar de que, teóricamente, el compuesto *Soft* debería ser el más veloz a una vuelta. Sin embargo, como se muestra en la gráfica (c), el *Soft* tiene la vida útil promedio más corta debido a su rápida degradación, lo cual explica por qué es el segundo compuesto más rápido en carrera. El compuesto más utilizado es el *Hard*, lo cual es comprensible dado que es el más duradero. Por todo esto, el tipo de llanta podría tener un efecto significativo sobre los tiempos de vuelta por kilómetro.

3.2 Circuitos y pitstops

Pirelli, el proveedor oficial de neumáticos de la Fórmula 1, clasifica todos los circuitos de acuerdo a sus características específicas.

Se dispone de datos sobre las características de los circuitos que pueden influir en la estrategia de carrera, especialmente aquellos aspectos que afectan el desgaste de los neumáticos, un factor crucial para determinar el momento óptimo de realizar paradas en pits:

- **Length:** Longitud del circuito en kilómetros.
- **Abrasiveness:** Nivel de abrasividad del asfalto de la pista.
- **Traction:** Exigencia del circuito en términos de tracción, crucial para salir con velocidad de las curvas lentas.
- **Braking:** Exigencia del circuito sobre los frenos. Las llantas se desgastan también al frenar, ya que colaboran con los frenos para desacelerar el coche.
- **Track Evolution:** Evolución de la pista a medida que los coches depositan caucho sobre el asfalto. En circuitos con alta evolución, esto puede tener un impacto significativo en los tiempos de vuelta.
- **Grip:** Nivel de adherencia del asfalto. Este factor puede depender de la frecuencia de uso del circuito o de características como la presencia de trampas de grava junto a la pista.
- **Lateral Forces:** Exigencia del circuito en términos de fuerzas laterales que actúan sobre el coche al tomar curvas. Estas fuerzas pueden influir en el desgaste de los neumáticos, ya que son los que mantienen al coche en la pista resistiendo las fuerzas centrífugas.
- **Downforce:** Nivel de carga aerodinámica recomendado para el circuito, que empuja al coche hacia el asfalto.

- **Tyre Stress:** Velocidades y cargas a las que están sometidos los neumáticos durante la carrera.

TERMINAR EDA

Chapter 4

Modelos

4.1 Modelos lineales generalizados

Los modelos lineales generalizados, al igual que los modelos lineales clásicos, tienen como objetivo expresar la relación entre una variable dependiente Y y variables explicativas $\mathbf{X} = (X_1, X_2, \ldots, X_n)$.

La principal ventaja de los modelos lineales generalizados sobre los modelos lineales clásicos es que no requieren que la variable dependiente siga una distribución normal. En su lugar, la variable dependiente puede seguir cualquier distribución de la familia exponencial. Además, estos modelos incorporan una función liga, la cual permite modelar relaciones no lineales dentro de un marco lineal, lo que hace que el análisis sea más flexible en comparación con el enfoque clásico.

Los modelos de regresión lineal clásicos están restringidos a datos que cumplen los supuestos de linealidad, normalidad, homocedasticidad e independencia.

- **Linealidad:** En los modelos clásicos, la relación entre Y y las variables explicativas $\mathbf{X}$ es lineal. En los modelos lineales generalizados, esta relación se modela a través de una función de enlace, lo que permite que la relación no sea necesariamente lineal, ya que la función de enlace transforma la media de Y.

- **Normalidad:** En los modelos clásicos, los residuos (la diferencia entre el valor observado y el estimado) deben seguir una distribución normal. En los modelos lineales generalizados, no es necesario que los residuos se distribuyan normalmente, ya que la distribución de Y puede pertenecer a la familia exponencial, sin ser necesariamente normal.

- **Homocedasticidad:** En los modelos clásicos, la varianza de los residuos debe ser constante. En los modelos lineales generalizados, la varianza de los residuos puede no ser constante, y puede depender de la media de la variable dependiente.

- **Independencia:** Los residuos deben ser independientes entre sí. La independencia entre observaciones también es necesaria para los modelos lineales generalizados.

Los modelos lineales generalizados tienen 3 componentes principales:

- **Distribución de los errores (Familia):** Se refiere a la distribución de la variable dependiente Y. La elección de esta distribución está guiada por la naturaleza de la variable dependiente, que puede ser continua, binaria, de conteo, etc. Esta distribución debe pertenecer a la familia exponencial.

- **Predictor lineal:** Es una combinación lineal de las variables independientes que establece la relación entre estas variables y la variable dependiente. Cada una de las variables explicativas está multiplicada por coeficientes que cuantifican esta relación. La expresión del predictor lineal es:

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{4.1}$$

- **Función de enlace:** La función de enlace conecta el predictor lineal con la media de la variable dependiente. Transforma el valor esperado de la variable dependiente a la escala del predictor lineal.

4.1.1 Estimación y ajuste del modelo

A diferencia de los modelos lineales clásicos que utilizan el método de mínimos cuadrados, los modelos lineales generalizados emplean el método de máxima verosimilitud para estimar sus parámetros ($\boldsymbol{\beta}$). Este método busca los valores de los parámetros que maximizan la función de verosimilitud, es decir, la probabilidad de observar los datos bajo el modelo propuesto. La forma de la función de verosimilitud depende de la distribución seleccionada para la variable dependiente. Maximizar la log-verosimilitud, el logaritmo de la función de verosimilitud, en lugar de la verosimilitud es más práctico gracias a la naturaleza aditiva del logaritmo.

Una vez estimados los parámetros, el modelo ajustado se expresa de la siguiente manera:

$$Y = g^{-1}(X\boldsymbol{\beta}) + \varepsilon \tag{4.2}$$

donde:

- Y es la variable dependiente,
- g es la función de enlace,
- X es la matriz de variables independientes,
- $\boldsymbol{\beta}$ es el vector de parámetros,
- ε es el vector de errores.

4.1.2 Evaluación del ajuste y selección de modelos

Determinar si el modelo propuesto representa adecuadamente los datos es crucial para garantizar un buen desempeño predictivo. Asimismo, es fundamental comparar diferentes modelos para seleccionar aquel que mejor se ajuste a los datos y proporcione una mejor capacidad de predicción.

Bondad de Ajuste

Para evaluar la bondad de ajuste en los modelos lineales generalizados se utiliza la **devianza**, que es la diferencia entre la log-verosimilitud del modelo ajustado y la de un modelo saturado con un ajuste perfecto.

Un modelo saturado tiene la misma distribución de la variable dependiente que el modelo propuesto y la misma función de enlace. En el modelo saturado se estima un parámetro para cada observación, por lo que tiene un ajuste perfecto.

La log-verosimilitud del modelo saturado es siempre estrictamente mayor que la del modelo propuesto. Cuanto mejor sea el ajuste del modelo propuesto, más se parecerán entre ellas. Por lo tanto, el estadístico conocido como devianza que está dado por:

$$D = 2\left[l(\boldsymbol{\beta}_{\max}; y) - l(\boldsymbol{\beta}; y)\right] \tag{4.3}$$

Debe ser lo más cercano a cero y, por lo tanto, si el modelo tiene un ajuste adecuado, D se distribuirá como $D \sim \chi^2_{n-p}$, donde n es el número de parámetros en el modelo saturado y p es el número de parámetros en el modelo propuesto. Una devianza baja indica un buen ajuste.

Comparación entre modelos

El **Criterio de Información de Akaike** (AIC) es útil para comparar modelos estadísticos. Este criterio busca balancear la bondad de ajuste con la complejidad del modelo, penalizando aquellos con un alto número de parámetros. El modelo con el menor valor de AIC es generalmente preferido. El AIC se define como:

$$\text{AIC} = -2\ln(L) + 2k \tag{4.4}$$

donde L es la verosimilitud del modelo y k es el número de parámetros en el modelo propuesto.

El **Criterio de Información Bayesiano** (BIC), similar al AIC, también busca balancear la bondad de ajuste y la complejidad del modelo, pero penaliza el número de parámetros de manera más estricta. Al igual que con el AIC, se prefieren los modelos con menores valores de BIC. El BIC se define como:

$$\text{BIC} = -2\ln(L) + k\ln(n) \tag{4.5}$$

donde:

- L es la verosimilitud del modelo propuesto,
- k es el número de parámetros en el modelo propuesto,
- n es el número de observaciones o datos.

También se puede utilizar la **devianza** para comparar modelos, pero los modelos comparados deben ser anidados, es decir, deben tener la misma distribución para los errores y utilizar la misma función de enlace, solo difieren en el número de variables independientes.

4.2 LapTimePerKM

El primer modelo lineal generalizado que ayudará a determinar la estrategia óptima es el de LapTimePerKM. La variable dependiente es el tiempo de vuelta kilómetro, esta variable es útil, ya que como los circuitos de la Fórmula 1 tienen diferentes lengths, si se quiere el tiempo de vuelta dado un circuito determinado solo se necesita multiplicar el número de kilómetros del circuito por esta estimación del tiempo de vuelta por kilómetro.

4.2.1 Selección de variables

Se seleccionaron las variables que podrían tener un impacto en el tiempo de vuelta por kilómetro de acuerdo con el análisis exploratorio de datos (EDA). Posteriormente, se utilizó la función `bestglm`, que compara todas las combinaciones posibles de esas variables para seleccionar el mejor modelo, utiliza el BIC para la selección del modelo óptimo.

A continuación, se realizó un análisis de ANOVA, partiendo del modelo propuesto por `bestglm` hasta llegar a un modelo con solo una variable independiente. El objetivo de este análisis era evaluar si alguna de las variables podía eliminarse para simplificar el modelo. Sin embargo, se comprobó que todas las variables eran significativas y contribuían a reducir la devianza. Por lo tanto, se decidió mantener el modelo completo propuesto por la función `bestglm`.

$$\begin{aligned} LaptimePerKM = {} & \beta_0 + \beta_1 GP + \beta_2 RacePercentage + \beta_3 Driver \\ & + \beta_4 Team + \beta_5 TyreLife + \beta_6 Compound \\ & + \beta_7 Position + \beta_8 Stint + \varepsilon \end{aligned} \tag{4.6}$$

4.2.2 Selección de familia y función de enlace

Una vez seleccionadas las variables, se evaluó el modelo utilizando diferentes familias de distribución para los errores y diversas funciones de enlace:

	Normal	Gamma	Gaussiana Inversa
Función de enlace	Logaritmo	Inversa	Inversa
AIC	84642	79481	77758
BIC	85267.27	80106.1	78383.15

Table 4.1: Comparación de diferentes familias de distribución

Como se puede observar, el modelo con la distribución Gaussiana inversa presenta el menor AIC y BIC, seguido por el modelo con la distribución Gamma.

Dado que el objetivo es modelar los tiempos de vuelta por kilómetro, los cuales son siempre positivos, la elección de una distribución Gamma para los errores es adecuada,

ya que facilita la interpretación del modelo. Por esta razón, se optó por continuar con el modelo basado en la distribución Gamma.

4.2.3 Resultados

	Variable dependiente:
	LapTimePerKM
GPAustin	−0.001*** (0.0000)
GPAustralia	0.01*** (0.0000)
GPAustria	0.01*** (0.0000)
GPAzerbaijan	0.001*** (0.0000)
GPBahrain	0.001*** (0.0000)
GPBelgium	0.01*** (0.0000)
GPBrazil	0.003*** (0.0000)
GPCanada	0.001*** (0.0000)
GPGreat Britain	0.01*** (0.0000)
GPHungary	−0.002*** (0.0000)
GPImola	0.01*** (0.0001)
GPJapan	0.01*** (0.0000)
GPMexico	−0.003*** (0.0000)
GPMiami	0.003*** (0.0000)
GPMonaco	−0.01*** (0.0000)
GPMonza	0.01*** (0.0000)
GPNetherlands	0.001*** (0.0001)
GPQatar	0.01*** (0.0001)
GPSaudi Arabia	0.01*** (0.0000)
GPSingapore	−0.01*** (0.0000)
GPSpain	0.001*** (0.0000)
RacePercentage	0.002*** (0.0000)
DriverALO	0.0001 (0.0001)
DriverBOT	0.0003*** (0.0001)
DriverDEV	0.001*** (0.0001)
DriverFIT	−0.01*** (0.0002)
DriverGAS	0.0002*** (0.0001)
DriverGIO	0.0004*** (0.0001)
DriverGRO	−0.001*** (0.0001)
DriverHAM	0.0002** (0.0001)
DriverHUL	−0.0002** (0.0001)
DriverKUB	−0.001*** (0.0001)
DriverKVY	−0.0000 (0.0001)
DriverLAT	−0.0003*** (0.0001)
DriverLAW	0.001*** (0.0001)
DriverLEC	0.0000 (0.0001)
DriverMAG	−0.001*** (0.0001)
DriverMAZ	−0.001*** (0.0001)
DriverMSC	−0.001*** (0.0001)
DriverNOR	0.0003*** (0.0001)
DriverOCO	0.0000 (0.0001)
DriverPER	0.0002*** (0.0001)
DriverPIA	0.001*** (0.0001)
DriverRAI	0.0004*** (0.0001)
DriverRIC	0.0001 (0.0001)
DriverRUS	−0.0002*** (0.0001)
DriverSAI	0.0001 (0.0001)
DriverSAR	0.0005*** (0.0001)
DriverSTR	−0.0001** (0.0001)
DriverTSU	0.0003*** (0.0001)
DriverVER	0.0002*** (0.0001)
DriverVET	−0.0002** (0.0001)
DriverZHO	0.0005*** (0.0001)
TeamAlphaTauri	0.0002** (0.0001)
TeamAlpine	0.0004*** (0.0001)
TeamAston Martin	0.001*** (0.0001)
TeamFerrari	0.001*** (0.0001)
TeamHaas F1 Team	0.001*** (0.0001)
TeamMcLaren	0.0002** (0.0001)
TeamMercedes	0.001*** (0.0001)
TeamRed Bull Racing	0.001*** (0.0001)
TeamWilliams	0.0004*** (0.0001)
TyreLife	−0.0000*** (0.0000)
CompoundMEDIUM	−0.0003*** (0.0000)
CompoundSOFT	−0.0004*** (0.0000)
Position	−0.0000*** (0.0000)
Stint	−0.0001*** (0.0000)
Constant	0.05*** (0.0001)
Observations	63,694
Akaike Inf. Crit.	79,480.83
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 4.1: Resultados del modelo LapTimePerKM

4.3 Pitstops

Poder estimar el tiempo pérdido en el *pitlane*, es decir, desde que un coche cruza la línea de los pits, entra a su *box* para cambiar llantas y sale del *pitlane*, es crucial para determinar la estrategia óptima, ya que el tiempo pérdido aquí es tiempo que los pilotos pierden en la pista y podría resultar en pérdida de posiciones.

4.3.1 Selección de variables

No había tantas variables en el conjunto de datos de paradas en pits entonces se utilizó `bestglm` con todas las variables y el mejor modelo propuesto fue:

$$PitstopT = \beta_0 + \beta_1 Circuit + \varepsilon \tag{4.7}$$

4.3.2 Selección de familia y función de enlace

Las diferentes familias comparadas para el modelo propuesto son:

	Normal	Gamma	Gaussiana Inversa
Función de enlace	Logaritmo	Inversa	Inversa
AIC	18468	17675	17469
BIC	18663.18	17870.89	17664.68

Table 4.2: Comparación de diferentes familias de distribución

Se decidió continuar con la familia Gamma, aunque tiene el segundo mejor AIC y BIC, para los errores por la misma razón que el otro modelo.

4.3.3 Resultados

Table 4.3: PitstopT

	Variable Dependiente:
	PitstopT
CircuitAustin	-0.01^{***} (0.001)
CircuitAustralia	0.004^{***} (0.001)
CircuitAustria	-0.0001 (0.001)
CircuitAzerbaijan	0.003^{***} (0.001)
CircuitBahrain	-0.01^{***} (0.001)
CircuitBelgium	-0.004^{***} (0.001)
CircuitBrazil	-0.002^{**} (0.001)
CircuitCanada	-0.005^{***} (0.001)
CircuitChina	-0.01^{***} (0.001)
CircuitEifel	-0.002^{*} (0.001)
CircuitFrance	-0.02^{***} (0.001)
CircuitGermany	0.002 (0.001)
CircuitGreat Britain	-0.01^{***} (0.001)
CircuitHungary	-0.004^{***} (0.001)
CircuitImola	-0.02^{***} (0.001)
CircuitJapan	-0.01^{***} (0.001)
CircuitLas Vegas	-0.003^{**} (0.001)
CircuitMexico	-0.003^{***} (0.001)
CircuitMiami	0.001 (0.001)
CircuitMonaco	-0.01^{***} (0.001)
CircuitMonza	-0.01^{***} (0.001)
Circuitmugello	-0.0001 (0.001)
CircuitNetherlands	0.001 (0.001)
CircuitPortugal	-0.01^{***} (0.001)
CircuitQatar	-0.01^{***} (0.001)
CircuitRussia	-0.01^{***} (0.001)
CircuitSaudi Arabia	-0.0003 (0.001)
CircuitSingapore	-0.01^{***} (0.001)
CircuitSpain	-0.002^{**} (0.001)
CircuitTurkey	-0.005^{***} (0.001)
Constant	0.05^{***} (0.001)
Observations	3,330
Akaike Inf. Crit.	17,675.35
Note:	*p<0.1; **p<0.05; ***p<0.01

4.4 Inlaps y Outlaps

TERMINAR ESTA SECCIÓN, SELECCIÓN DE VARIABLES Y FAMILIA

4.4.1 Inlaps

$$\begin{aligned} LaptimePerKM = \beta_0 + \beta_1 GP + \beta_2 Compound + \beta_3 TyreLife \\ + \beta_4 Stint + \varepsilon \end{aligned} \tag{4.8}$$

Resultados

Table 4.4: Inlaps

	Variable dependiente:
	LapTimePerKM
GPAustin	0.002^{***} (0.001)
GPAustralia	-0.004^{***} (0.001)
GPAustria	-0.001 (0.001)
GPAzerbaijan	-0.001 (0.001)
GPBahrain	0.002^{***} (0.001)
GPBelgium	0.01^{***} (0.001)
GPBrazil	-0.01^{***} (0.001)
GPCanada	-0.01^{***} (0.001)
GPChina	0.002^{**} (0.001)
GPGreat Britain	0.01^{***} (0.001)
GPHungary	-0.002^{***} (0.001)
GPImola	-0.003^{***} (0.001)
GPJapan	0.01^{***} (0.001)
GPMexico	-0.01^{***} (0.001)
GPMiami	0.001 (0.001)
GPMonaco	-0.02^{***} (0.001)
GPMonza	0.01^{***} (0.001)
GPNetherlands	-0.005^{***} (0.001)
GPQatar	0.01^{***} (0.001)
GPSaudi Arabia	0.01^{***} (0.001)
GPSingapore	-0.01^{***} (0.001)
GPSpain	-0.0005 (0.001)
CompoundMEDIUM	0.002^{***} (0.0003)
CompoundSOFT	0.002^{***} (0.0004)
TyreLife	0.0003^{***} (0.0000)
Stint	0.0005^{***} (0.0001)
Constant	0.04^{***} (0.001)
Observations	2,729
Akaike Inf. Crit.	12,381.00
Note:	*p<0.1; **p<0.05; ***p<0.01

4.4.2 Outlaps

$$LaptimePerKM = \beta_0 + \beta_1 GP + \beta_2 Compound + \varepsilon \tag{4.9}$$

Resultados

Table 4.5: Outlaps

	Variable Dependiente:
	LapTimePerKM
GPAustin	−0.0002 (0.001)
GPAustralia	0.01*** (0.001)
GPAustria	−0.001 (0.001)
GPAzerbaijan	0.001** (0.001)
GPBahrain	0.0000 (0.001)
GPBelgium	0.01*** (0.001)
GPBrazil	−0.004*** (0.001)
GPCanada	0.002** (0.001)
GPChina	0.01*** (0.001)
GPGreat Britain	0.002*** (0.001)
GPHungary	−0.003*** (0.001)
GPImola	−0.002** (0.001)
GPJapan	0.004*** (0.001)
GPMexico	−0.004*** (0.001)
GPMiami	0.003*** (0.001)
GPMonaco	−0.01*** (0.001)
GPMonza	0.01*** (0.001)
GPNetherlands	−0.002*** (0.001)
GPQatar	0.01*** (0.001)
GPSaudi Arabia	0.003*** (0.001)
GPSingapore	−0.01*** (0.001)
GPSpain	−0.0002 (0.001)
CompoundMEDIUM	−0.001*** (0.0002)
CompoundSOFT	−0.001*** (0.0003)
Constant	0.05*** (0.0005)
Observations	2,570
Akaike Inf. Crit.	11,865.96
Note:	*p<0.1; **p<0.05; ***p<0.01

4.5 SafetyCar

4.6 Lluvia

Chapter 5

Implementación

5.1 Árboles de Decisión

Los datos siempre han jugado un papel crucial en la Fórmula 1, incluso antes de la disponibilidad de herramientas sofisticadas. Unos pocos milisegundos pueden marcar la diferencia entre comenzar desde la pole position o desde la décima posición.

Sin embargo, los datos por sí solos no garantizan decisiones más eficientes. Es necesario contar con herramientas que permitan extraer y utilizar el potencial de estos datos para tomar decisiones más informadas. Los árboles de decisión son una de estas herramientas.

Un árbol de decisión es una representación gráfica de un problema de decisión. Comienza con un nodo raíz, que es el punto de partida del árbol, y representa una decisión inicial. A partir de este nodo, se ramifica en diferentes opciones o alternativas, representadas por ramas.

Cada nodo de decisión en el árbol representa un punto en el que se debe tomar una decisión adicional, con sus propias ramas que representan las opciones disponibles en ese punto. Al final de cada secuencia de decisiones, se llega a un nodo terminal que muestra el resultado específico de esa secuencia.

Para encontrar la mejor decisión, se resuelve el árbol hacia atrás, comenzando por el nodo terminal con el resultado más deseado y rastreando la ruta de decisiones que lleva a ese resultado. De esta manera, el árbol de decisión ayuda a identificar la serie de decisiones más efectiva para alcanzar el objetivo deseado.

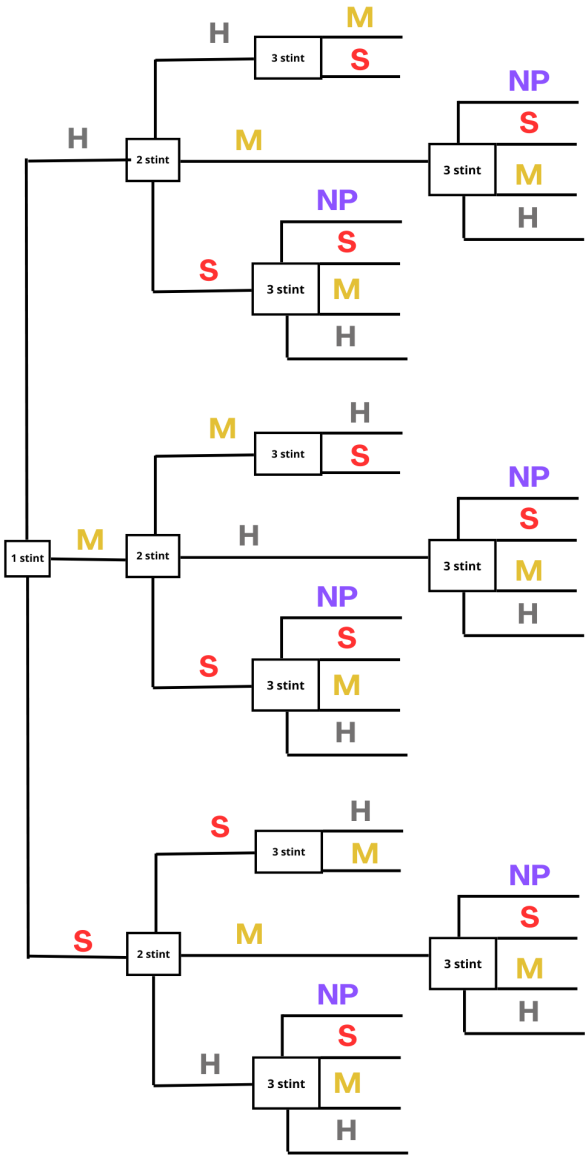


Figure 5.1: Árbol de decisión que siguen los modelos

En el caso de este problema, el nodo raíz se refiere a la decisión sobre qué compuesto de neumático utilizar para el comienzo de la carrera, es decir, para el primer stint. Las opciones disponibles son los tres compuestos de neumático en seco (*Slicks*): *HARD*,

MEDIUM o *SOFT*.

El siguiente nodo de decisión se refiere a la primera parada en pits para cambiar llantas. Nuevamente, se tienen como opciones los tres compuestos de neumático. Esta parada es obligatoria por reglamento, ya que, si la pista se declara seca, todos los pilotos deben usar al menos dos compuestos diferentes durante la carrera.

El último nodo de decisión se refiere al tercer stint. Si se eligió en el segundo stint el mismo compuesto con el que se comenzó la carrera, esta segunda parada en pits es obligatoria. En este caso, las opciones disponibles son los compuestos que no se han usado, para evitar una parada adicional. Si en la primera parada se cambió de compuesto, se tienen cuatro opciones: los tres compuestos restantes o la opción de no parar, es decir, no entrar a pits para cambiar llantas y continuar hasta el final de la carrera.

Al final de cada ruta se encuentran los nodos terminales con el tiempo total estimado de la carrera, que incluye los tiempos de los stints y el tiempo perdido en pits. Con este tiempo estimado, se puede elegir la mejor estrategia basada en el menor tiempo total estimado.

5.2 Funciones

En esta sección se presentan las funciones comunes que serán utilizadas en todos los modelos para realizar cálculos recurrentes de manera eficiente.

5.2.1 vidapromedio

Esta función tiene como objetivo calcular el máximo, promedio y la desviación estándar de la duración de los *stints* por compuesto de neumáticos.

Recibe como argumentos un conjunto de datos con la información de todos los *stints* (2019-2023) y el circuito. La función comienza filtrando los datos de acuerdo con el circuito proporcionado, seleccionando un subconjunto de *stints* correspondientes a dicho circuito.

Posteriormente, calcula el máximo, promedio y desviación estándar de la duración de los *stints* por compuesto de neumático. Al finalizar, devuelve un *dataframe* con las estadísticas de cada tipo de neumático.

5.2.2 tiempoStint

TERMINAR FUNCIONES

5.3 Modelo Determinista

Algunas veces en Fórmula 1 los equipos se preocupan por lo que hacen sus rivales, pero, según Bernie Collins, su enfoque inicial está en sus propios pilotos. "[...] calcularíamos la estrategia más rápida para un coche en particular. En otras palabras, si fueras el único coche en la pista, ¿cuál sería el tiempo más rápido posible para completar la distancia de carrera y cómo lo lograrías?" (Collins, 2024, p. 126).

Esto es lo que se conoce como una *single car race* y constituye el punto de partida para cualquier estrategia de carrera. Según Collins, la estrategia óptima en una carrera con muchas oportunidades de adelantamiento se parecerá mucho a la de una *single car race*, ya que, en teoría, el piloto debería ser capaz de adelantar con llantas más frescas.

El primer modelo se basa precisamente en una single car race con la peculiaridad de que las paradas en pits solo se pueden hacer en la vuelta de la vida esperada del compuesto de neumáticos por eso se le llama determinista a este modelo. Calcula todas las alternativas de estrategia que puede haber y sigue el árbol de decisión descrito anteriormente.

TERMINAR EXPLICACIÓN DEL MODELO

5.4 Modelo Ventana

5.5 Modelo Competidor más cercano

5.6 Safety Cars

5.7 Lluvia

Chapter 6

Resultados

Atinarle a la estrategia

Atinarle al número de vuelta en la que pararon

Si dio diferente estrategia, pensar si es más eficiente la del modelo

Tiempo total de la carrera

Chapter 7

Conclusiones