

---

# ORACLE

---

## Oracle

Isabel Piedrahíta Vélez  
Laia Font Solanich

June 2024

# Contents

<b>1</b>	<b>Introducción</b>	<b>1</b>
<b>2</b>	<b>La Opinión de un LLMs</b>	<b>1</b>
2.1	¿Cuáles son las variables que el aficionado percibe que tienen un impacto positivo en el interés de una carrera de Formula 1? . . . . .	1
2.1.1	Primer Experimento Usando LLMs . . . . .	1
2.1.2	Segundo Experimento Usando LLMs . . . . .	3
2.1.3	Tercer Experimento Usando LLMs . . . . .	4
2.1.4	Cuarto Experimento Usando LLMs . . . . .	5
2.2	¿Qué carrera de Formula 1 puede ser la más interesante para el usuario el siguiente año? . . . . .	5
<b>3</b>	<b>Análisis de datos</b>	<b>7</b>
3.1	Análisis de Correlación . . . . .	7
3.2	Análisis Temporal . . . . .	8
3.3	Análisis de Condiciones de Pista . . . . .	9
3.4	Análisis de Cambio de Posiciones . . . . .	10
<b>4</b>	<b>Modelado de Machine Learning</b>	<b>10</b>
4.1	Modelo Grande . . . . .	11
4.2	Modelo Pequeño . . . . .	12
<b>5</b>	<b>Predicción de puntuaciones 2021</b>	<b>14</b>
<b>6</b>	<b>Conclusions</b>	<b>14</b>

## List of Figures

1	Datos de la mejor y peor carrera basado en la puntuación . . . . .	7
2	Matriz Correlación RACES.csv . . . . .	8
3	Diagrama de caja y bigotes mostrando la puntuación por año . . . . .	9
4	Diagrama de caja y bigotes de score por año segregado en seco y mojado . . . . .	9
5	Diagrama comparativo de cambio de posiciones en la mejor y peor carreras basándonos en la puntuación . . . . .	10
6	Diagrama de puntos con la correlación entre adelantamientos y la puntuación . . . . .	11
7	Diagrama comparativo del Ghost Car . . . . .	11
8	Resultado de Experimentos con más Features . . . . .	12
9	Predicciones de puntuaciones 2020 con la distancia entre lo esperado y la predicción .	12
10	Resultado de Experimentos con menos Features . . . . .	13
11	Predicciones de puntuaciones 2020 con menos variables . . . . .	13
12	Predicciones de puntuaciones 2021 con los dos modelos . . . . .	14

## List of Tables

# 1 Introducción

La Fórmula 1 (F1) se ha convertido en uno de los deportes automovilísticos más vistos del mundo, con una media de 1.11 millones de espectadores por Gran Premio [1]. A medida que aumenta su popularidad, es importante analizar qué es lo que hace que crezca en popularidad tan rápidamente, no sólo para entender qué es lo que hace que los espectadores la adoren, sino también para intentar aplicarlo a deportes similares que no son tan populares en todo el mundo. Adicionalmente, la información sobre cómo perciben los espectadores una carrera puntual de la Fórmula 1 puede ser increíblemente valiosa para varios grupos interesados, tales como compañías publicitarias, equipos de corredores e incluso los mismos espectadores pueden beneficiarse de esta información.

En este archivo encontraréis un análisis relacionado con datos de carreras de F1 para entender mejor qué hace que a los espectadores les guste más una carrera o menos, y al final una predicción con las variables relevantes para predecir futuras puntuaciones de carreras.

## 2 La Opinión de un LLMs

Los Large Language Models o LLMs pueden ser una herramienta inestimable para procesar grandes cantidades de datos de forma digerible, por lo cual es interesante experimentar con diferentes formas de usarlos durante la etapa de análisis exploratorio de datos.

### 2.1 ¿Cuáles son las variables que el aficionado percibe que tienen un impacto positivo en el interés de una carrera de Formula 1?

En esta sección proponemos cuatro experimentos.

1. Preguntar a ChatGPT-4 cuáles variables tienen un impacto positivo en el interés de una carrera de Fórmula 1.
2. Repetir la misma pregunta a ChatGPT-4 en un chat nuevo, pero en inglés. Al ser el lenguaje hegemónico en los datos con los que se entrenó el modelo, es interesante ver cómo puede diferir la respuesta dependiendo del idioma de la pregunta.
3. Usar la función para cargar archivos a ChatGPT-4 y ver si esto afecta la respuesta.
4. Hacer un poco de *prompt engineering* para intentar conseguir la respuesta más prometedora posible.

#### 2.1.1 Primer Experimento Usando LLMs

Empezaremos haciendo la pregunta tal cual fue planteada en la definición del reto a ChatGPT-4.

**¿Cuáles son las variables que el aficionado percibe que tienen un impacto positivo en el interés de una carrera de Formula 1?**

R/ "El interés de los aficionados en una carrera de Fórmula 1 puede verse influenciado por varias variables percibidas como positivas. Algunas de las más destacadas son:

1. Competitividad de los equipos y pilotos:
  - La igualdad de rendimiento entre los equipos y pilotos hace que las carreras sean más

impredecibles y emocionantes.

- Las luchas cerradas por las primeras posiciones mantienen a los espectadores al borde de sus asientos.

2. Calidad de los adelantamientos:

- La habilidad y frecuencia con la que los pilotos pueden adelantar a sus competidores en la pista.
- Adelantamientos espectaculares y estrategias de carrera que permiten cambios constantes en las posiciones.

3. Rendimiento y estrategias de los equipos:

- Diferentes estrategias de pit stops y el uso de neumáticos que añaden un elemento de incertidumbre y emoción.
- Decisiones estratégicas inteligentes que pueden cambiar el curso de la carrera.

4. Circuitos emocionantes:

- Circuitos con diseños desafiantes, que incluyan curvas difíciles y rectas largas para alta velocidad.
- Pistas históricas o icónicas que tienen un prestigio especial y generan gran expectativa.

5. Clima y condiciones de la pista:

- Carreras con condiciones meteorológicas variables, como lluvia, que pueden alterar el rendimiento de los coches y las estrategias.
- Cambios inesperados en el clima que añaden un elemento de imprevisibilidad.

6. Desempeño de pilotos favoritos o carismáticos:

- Actuaciones destacadas de pilotos populares o con gran carisma.
- Historias de superación personal o rivalidades entre pilotos que añaden un componente narrativo atractivo.

7. Innovaciones tecnológicas y técnicas:

- Introducción de nuevas tecnologías y su impacto en el rendimiento de los coches.
- Evolución de las reglas y regulaciones que fomenten la innovación sin comprometer la competitividad.

8. Cobertura mediática y comentarios:

- Calidad de la transmisión, incluyendo ángulos de cámara emocionantes y comentarios informativos y apasionantes.
- Acceso a contenido detrás de escenas y análisis detallados que enriquecen la experiencia del espectador.

9. Eventos y actividades paralelas:

- Actividades adicionales durante el fin de semana de la carrera, como sesiones de autógrafos, eventos promocionales y conciertos.
- Experiencias interactivas para los fans en el circuito o a través de plataformas digitales.

#### 10. Seguridad y deportividad:

- Altos estándares de seguridad para pilotos y espectadores, asegurando que la emoción no comprometa la integridad física.
- Un espíritu deportivo que promueva el respeto y la competencia justa entre los pilotos y equipos.

Cada uno de estos factores puede contribuir significativamente a la percepción positiva del interés en una carrera de Fórmula 1 por parte de los aficionados.”

Notamos que la respuesta cuenta con una cantidad importante de factores que como espectadoras de la Formula 1 sabemos que son relevantes, pero que como desarrolladoras sabemos que son difíciles de cuantificar y explorar por medio de datos, especialmente a la luz de los datasets que tenemos a nuestra disposición.

Factores como las condiciones de la pista, el clima, el interés de la pista, paradas en boxes, y adelantamientos son datos que podemos explorar. Otros factores como las actividades paralelas a la carrera, la narrativa alrededor de la carrera o la cobertura mediática quedan por fuera de lo que podemos hacer con los datos disponibles.

La respuesta dada por el modelo no es mala, sin embargo, carece de contexto, algo que intentaremos mejorar durante los experimentos 3 y 4, al proveer al modelo con más información sobre lo que estamos intentando lograr mediante la provisión de los datasets o más información en el prompt.

### 2.1.2 Segundo Experimento Usando LLMs

En este experimento intentamos traducir la pregunta de la forma más leal al original posible. Así obtuvimos el siguiente prompt para ChatGPT-4:

**Which of the variables a formula 1 fan can percieve have a positive impact on the interest of a Formula 1 race?** R/ "The interest of a Formula 1 race is influenced by several factors that fans can perceive. Here are some key variables that positively impact their interest:

1. Competitive Racing: Close and competitive racing, where multiple drivers and teams have a chance to win, significantly enhances interest. Battles for positions, frequent overtakes, and strategic play among drivers and teams make the race exciting.
2. Driver Personalities and Rivalries: Engaging and charismatic drivers with strong personalities, as well as rivalries between top drivers, draw fans' attention. Stories and tensions off the track add to the drama and intrigue.
3. Team Strategies: The strategic elements of the race, including pit stop strategies, tire choices, and fuel management, add depth and suspense. Fans enjoy speculating and analyzing these decisions.
4. Track Layout and Conditions: Certain circuits are known for producing more thrilling races due to their layout. Tracks with a mix of high-speed sections, tight corners, and overtaking

opportunities are particularly popular. Weather conditions, such as rain, can also make races unpredictable and exciting.

5. Technological Innovations: The constant evolution of car technology and how teams adapt to new regulations can be fascinating for fans. Innovations in aerodynamics, power units, and other technical aspects keep fans engaged.
6. Historical Significance: Races held at historically significant tracks or those that commemorate important milestones in F1 history tend to attract more interest.
7. Broadcast Quality: High-quality broadcast coverage, including expert commentary, on-board camera footage, detailed analytics, and immersive graphics, enhances the viewing experience.
8. Fan Engagement: Interactive elements such as fan voting for driver of the day, social media interactions, and other forms of engagement help fans feel more connected to the race and its participants.
9. Unexpected Events: Unpredictable elements such as crashes, safety cars, and unexpected technical failures add an element of surprise and keep fans on the edge of their seats.
10. Storylines and Narratives: Ongoing storylines throughout the season, such as a driver's quest for a championship, a rookie's rise, or a veteran's comeback, keep fans invested in each race.

Each of these factors contributes to the overall excitement and appeal of a Formula 1 race, making it more engaging for fans.”

Notamos que la respuesta es prácticamente igual a la obtenida en el experimento descrito en la Sección 2.1.1. Básicamente consiste en una traducción casi que textual de la respuesta previamente obtenida. Esto nos deja saber que realmente no hay mayor diferencia entre usar ChatGPT-4 en inglés o español en este caso. Sin embargo, esto no es suficiente información para garantizar que este siempre será el caso, por lo que seguiremos usando el modelo en inglés para las siguientes secciones traduciendo al español para facilidad de lectura.

### 2.1.3 Tercer Experimento Usando LLMs

En esta ocasión repetimos la misma pregunta planteada en la Sección 2.1.2, sin embargo también cargamos RACES.csv y RESULTS.csv usando la funcionalidad para analizar archivos de ChatGPT-4. Desafortunadamente, esta funcionalidad está mucho más enfocada a analizar y resumir archivos tipo PDF, por lo que la respuesta fue un poco decepcionante.

En este caso la respuesta del LLM es un breve análisis de las columnas de los datasets, definiendo que representa cada una de ellas. Posteriormente recomienda hacer un merge de ambas fuentes de datos y conducir un análisis de correlación, complementa esta sugerencia con un borrador de cómo el código para realizar esta operación se vería en Python.

Sería interesante llevar este concepto más allá e intentar introducir información de estos datasets mediante finetuning a un modelo como Cohere Light usando OCI Generative AI Playground. Sin embargo, los datasets que se usan para hacer finetuning de LLMs deben estar en formato prompt y respuesta, por lo que al no tener nosotros conocimiento suficiente sobre el dominio para escribir estos prompts y respuestas por nuestra cuenta no será posible realizar finetuning de este tipo.



#### 2.1.4 Cuarto Experimento Usando LLMs

Finalmente, intentaremos darle un poco más de contexto al modelo modificando el prompt para reflejar los datos que tenemos disponibles. Así obtenemos el siguiente prompt.

**Which of the following variables are most likely to predict spectator score of a formula 1 race:**

**RACES.csv:**

**RACEID, YEAR, ROUND, NAME, F1DATE, TIME, URL, SCORE, DNF\_COUNT, DNF\_DUE\_TO\_ACCIDENT\_COUNT, WEATHER, WEATHER\_WET, CIRCUITREF, YEAR\_C, RACE\_COUNT, NAME\_YEAR, OVERTAKEN\_POSITIONS\_TOTAL**

**RESULTS.csv:**

**RESULTID, RACEID, DRIVERID, CONSTRUCTORID, F1NUM, GRID, POSITION, POSITIONTEXT, POSITIONORDER, POINTS, LAPS, TIME, MILLISECONDS, FASTESTLAP, RANK, FASTESTLAPTIME, FASTESTLAPSPEED, STATUSID, YEAR, ROUND, NAME, POINTS\_CHAR**

**LAPTIMES.csv:**

**RACEID, DRIVERID, LAP, POSITION, MILLISECONDS, DRIVERREF, RACE\_NAME, PIT\_STOP, MILLISECONDS\_CUMULATIVE, MEDIAN\_MILLISECONDS\_CUMULATIVE**

R/ "Las variables que capturan la dinámica de la carrera, el rendimiento del conductor y las condiciones ambientales son las más propensas a predecir la puntuación del espectador. Específicamente, se deben considerar para modelar la puntuación del espectador de una carrera de Fórmula 1 variables relacionadas con el número de abandonos (DNF), las condiciones meteorológicas, las posiciones adelantadas, la vuelta de la carrera, la posición en la parrilla, los cambios de posición, las vueltas más rápidas y las paradas en boxes."

Dado este prompt ChatGPT-4 provee una respuesta relativamente útil. Además de seleccionar las variables que más relevantes considera de cada uno de los datasets, da una breve explicación sobre por qué fueron seleccionados.

## 2.2 ¿Qué carrera de Formula 1 puede ser la más interesante para el usuario el siguiente año?

En esta sección no conduciremos tantos experimentos debido a que ya la mayoría del contexto ha sido dado por las preguntas anteriores. Lo único diferente que haremos será empezar desde el mismo chat usado para la Sección 2.1.4.

De esta forma obtenemos el siguiente intercambio con ChatGPT-4.

**¿Qué carrera de Formula 1 puede ser la más interesante para el usuario el siguiente año?**

R/ "Determinar cuál será la carrera de Fórmula 1 más interesante el próximo año puede depender de varios factores, incluyendo las expectativas basadas en eventos pasados, características del

circuito, y otros elementos que suelen atraer a los aficionados. Aquí hay algunas carreras que habitualmente son consideradas emocionantes y que podrían ser particularmente interesantes el próximo año:

#### 1. Gran Premio de Mónaco

- **Características:** Circuito urbano con curvas cerradas y muy poco espacio para adelantar.
- **Razón de interés:** Prestigio histórico, glamour y el desafío técnico extremo que representa para los pilotos.

#### 2. Gran Premio de Bélgica (Spa-Francorchamps)

- **Características:** Uno de los circuitos más largos, conocido por sus cambios de elevación y el famoso sector de Eau Rouge y Raidillon.
- **Razón de interés:** Frecuentemente ofrece carreras emocionantes debido a su trazado rápido y técnico, además de ser impredecible en términos de clima.

#### 3. Gran Premio de Italia (Monza)

- **Características:** Circuito de alta velocidad conocido como el "Templo de la Velocidad".
- **Razón de interés:** Historia rica, velocidad extrema y la pasión de los tifosi (fans italianos).

#### 4. Gran Premio de Gran Bretaña (Silverstone)

- **Características:** Circuito con una mezcla de curvas rápidas y lentas.
- **Razón de interés:** Considerado uno de los circuitos más emocionantes del calendario, con un público apasionado y carreras frecuentemente impredecibles.

#### 5. Gran Premio de Brasil (Interlagos)

- **Características:** Circuito con varias oportunidades de adelantamiento y cambios de elevación.
- **Razón de interés:** Históricamente ha sido escenario de carreras dramáticas y finales de temporada emocionantes.

#### 6. Gran Premio de Singapur

- **Características:** Carrera nocturna en un circuito urbano.
- **Razón de interés:** La espectacularidad de la carrera nocturna y la complejidad del circuito urbano que presenta un desafío único.

#### 7. Gran Premio de Japón (Suzuka)

- **Características:** Circuito con diseño de figura en ocho, conocido por sus curvas desafiantes.
- **Razón de interés:** Uno de los circuitos favoritos de los pilotos, frecuentemente ofrece carreras emocionantes y cambios climáticos.

Además de estos clásicos, también hay nuevas incorporaciones al calendario que podrían ser muy interesantes, como el Gran Premio de Las Vegas, que debutará con una carrera nocturna en el famoso Strip de Las Vegas. Esta carrera promete ser un espectáculo impresionante y podría ser una de las más comentadas del año.

Elegir la "más interesante" también puede depender de cómo se desarrollen las dinámicas de la temporada, las rivalidades emergentes, y cualquier cambio reglamentario o técnico que pueda influir en la competitividad de los equipos. "

En esta respuesta nuevamente notamos que, a pesar de ser útil desde la perspectiva de un humano, no puede guiar mucho nuestro análisis. Lo que llevaremos de esta respuesta a nuestro análisis es un interés por la pista en la que se conduce la carrera, ya que claramente tendrá un alto impacto en el interés de los espectadores.

### 3 Análisis de datos

Se nos proporciona un conjunto de datos que incluye datos de muchas carreras diferentes de distintas épocas del automovilismo. Las tablas disponibles son las clasificaciones de pilotos, tiempos por vuelta, paradas en boxes, carreras, resultados y coche de seguridad.

Basándonos en los datos y en el reto propuesto, comenzamos analizando las puntuaciones más alta y más baja obtenidas en una carrera, que son el GP de Brasil de 2012 con un 9.45 sobre 10 y el GP de Abu Dhabi de 2020 con un 3.20 sobre 10. Las principales diferencias (ver Figura 1) entre estas dos carreras son el clima, ya que en Brasil llovió y en Abu Dhabi no, la cantidad de coches retirados tanto por causa externa como por accidente, y si el accidente causó que el *Safety Car* estuviera presente. Finalmente, hubo una cantidad mayor de adelantamientos en Brasil que en Abu Dhabi, ¡en total casi 5 veces más!

FDATE	RACED	YEAR	NAME	TIME	URL	WEATHER	WEATHER_VET	CIRCUITREF	YEAR_C	NAME_YEAR	ROUND	SCORE	DNF_COUNT	DNF_DUE_TO_ACCIDENT_COUNT	RACE_COUNT	OVERTAKEN_POSITIONS_TOTAL	LAPS_WITH_CHANGE_POS_1	LAPS_WITH_CHANGE_POS_1_5	RANK_VERSUS_POSITION	SAFETY_CAR
20/12/2012 12:00:00.000 a.m.	876	2012	Brazilian Grand Prix	16:00:00	<a href="http://en.wikipedia.org/wiki/2012_Brazilian_Grand_Prix">http://en.wikipedia.org/wiki/2012_Brazilian_Grand_Prix</a>	Intermittent drizzle at start, light rain intermittent drizzle at the end [2] Air Temp 20 °C Sea Temp 23 °C Track Temp 23 °C (T0-F)(2)	Y	Interlagos	2012	2012 Brazilian Grand Prix	20	9.45	6	6	1	367	6	25	17	1
13/12/2020 12:00:00.000 a.m.	1047	2020	Abu Dhabi Grand Prix	15:30:00	<a href="https://en.wikipedia.org/wiki/2020_Abu_Dhabi_Grand_Prix">https://en.wikipedia.org/wiki/2020_Abu_Dhabi_Grand_Prix</a>	Clear	N	yas_vicenna	2020	2020 Abu Dhabi Grand Prix	17	3.20	1	0	1	79	0	8	12	0

Figure 1: Datos de la mejor y peor carrera basado en la puntuación

#### 3.1 Análisis de Correlación

Un primer paso lógico sería crear diferentes matrices de correlación para verificar si alguna de las variables en nuestros conjuntos de datos está particularmente correlacionada con el *Score*. Esto nos da una idea general de en qué dirección mover nuestro análisis.

Primero analizaremos el conjunto de datos, creando la matriz de correlación para todas las variables numéricas, como se puede ver en la figura 2. Haciendo este análisis notamos que no hay ninguna variable muy fuertemente correlacionada con *Score*. La correlación positiva más grande es de 0.52 con *MAX\_COMEBACKSCORE\_POSITION* y *LAPS\_WITH\_CHANGE\_POS\_1\_5*. Mientras que los coeficientes de correlación más bajos son -0.06 con *ROUND* y 0 con *RACE\_COUNT*. Estos últimos dos, al ser tan cercanos a cero, nos hacen sospechar que estas dos variable realmente no tendrán un impacto positivo sobre nuestra predicción (como mucho, crearán ruido y confusión) y podrán fácilmente ser removidas del dataset de entrenamiento.

También podemos notar que en el cuartil de abajo a la derecha hay varias correlaciones bastante fuertes entre múltiples variables. Cuando sea hora de entrenar el modelo, puede que sea buena idea remover algunas de estas variables, ya que tener múltiples variables altamente correlacionadas puede degradar el rendimiento del modelo.

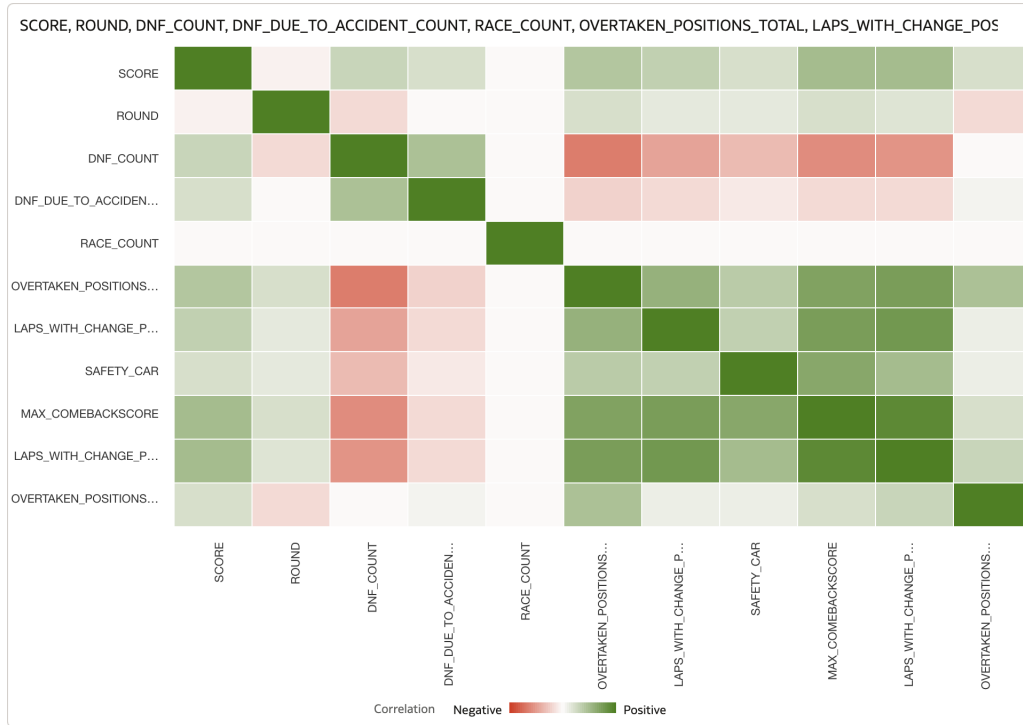


Figure 2: Matriz Correlación RACES.csv

### 3.2 Análisis Temporal

Un factor importante a considerar en lo que respecta a las carreras de Fórmula 1 es el paso del tiempo. Sabemos que muchos factores pueden impactar el puntaje que los espectadores dan a una carrera particular. Entre estos factores se encuentran variables como los corredores, el circuito o la tecnología que se esté usando. Todas estas variables están correlacionadas con el paso del tiempo, ya que los corredores tienen períodos de actividad limitados a cierta cantidad de años, y las nuevas tecnologías se introducen cada cierto número de años.

Para ver cómo se comporta la puntuación a lo largo del tiempo, podemos hacer un diagrama de cajas (ver Figura 3). Este nos permite ver un resumen bastante completo de la información, ya que contiene varias métricas importantes como máximos, mínimos, la mediana, y el segundo y tercer cuartil.

Al mirar el diagrama (ver Figura 3) es claro que el rating de las carreras fluctúa a lo largo de los años, en un patrón aparentemente ondulatorio; sin embargo, no hay suficiente información como para definir si este patrón es consistente a lo largo del tiempo. Igualmente, notamos que la fluctuación en la percepción de las carreras puede estar marcada por muchos factores externos, como los pilotos y la campaña alrededor de dicha carrera, que pueden variar de año en año a medida que pilotos conocidos y amados salen, y nuevos pilotos entran y se hacen conocer y querer por el público.

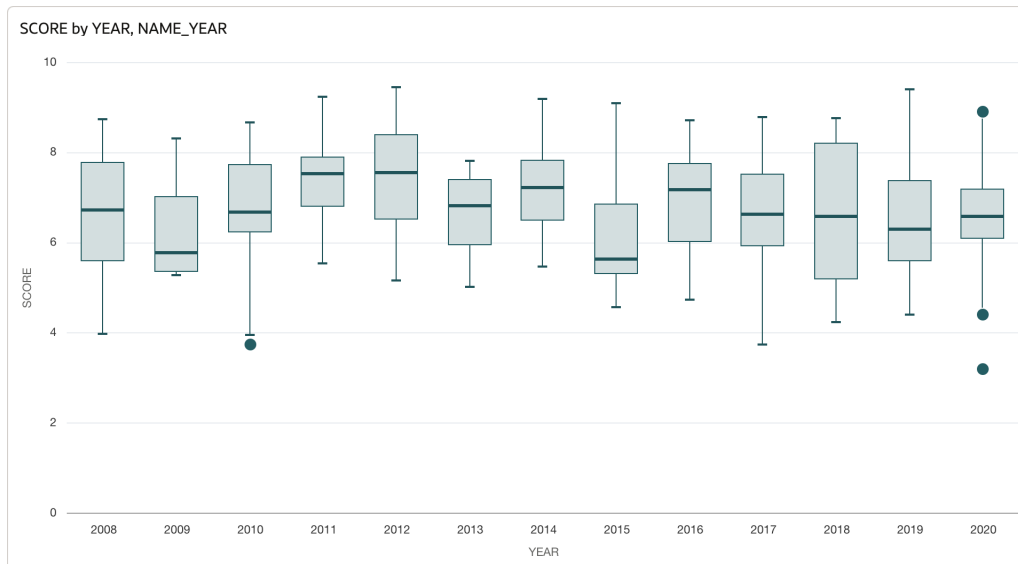


Figure 3: Diagrama de caja y bigotes mostrando la puntuación por año

### 3.3 Análisis de Condiciones de Pista

Teniendo en cuenta que las carreras más relevantes son de distintos años, con coches con propiedades distintas, hemos decidido analizar por año las distintas puntuaciones de las carreras y segregarlas entre carreras en seco y mojado, ya que en el breve análisis inicial hemos visto que una carrera en mojado mostraba una mayor puntuación. La Figura 4 muestra el diagrama en el cual se puede ver como las carreras en mojado suelen obtener una puntuación más alta que las carreras en seco, confirmando nuestra hipótesis. Esto nos indica que las carreras en mojado son más interesantes para el espectador y gustan más que la mayoría en seco. Esto se puede relacionar a que es más usual que una carrera sea en seco que no en mojado, transmitiendo emoción de que haya una situación especial al espectador.

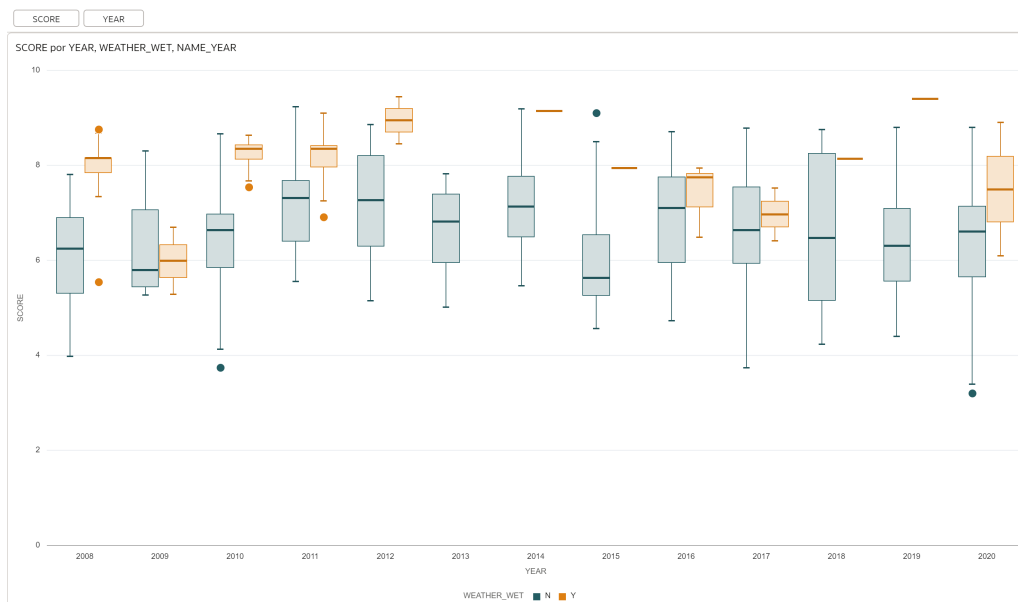


Figure 4: Diagrama de caja y bigotes de score por año segregado en seco y mojado

### 3.4 Análisis de Cambio de Posiciones

Basándonos en el primer breve análisis comparando los datos globales de las carreras, la Figura 5 muestra el cambio de posiciones que hay en total en la pista, los cuales se relacionan con adelantamientos tanto dentro de la pista como en el *pit lane*. En el diagrama se puede ver claramente como en la carrera de Brasil hubo una cantidad mayor de cambio de posiciones que en la carrera de Abu Dhabi. Por ejemplo, en la vuelta 54 de Brasil vemos como Hamilton desaparece del diagrama, indicando que retiraron el coche. Lo mismo se puede apreciar en la vuelta 8 de Abu Dhabi como Pérez desaparece.

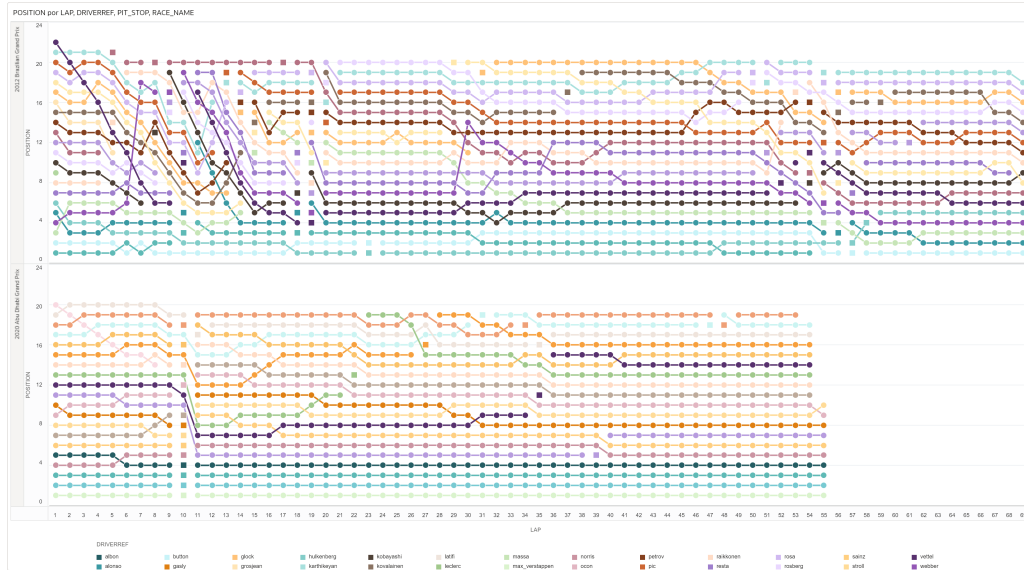


Figure 5: Diagrama comparativo de cambio de posiciones en la mejor y peor carreras basándonos en la puntuación

Para validar aún más nuestra hipótesis, hacemos un diagrama de puntos con la línea de tendencia de los adelantamientos totales de la carrera. En la Figura 6 podemos ver como esperado, como más adelantamientos (o cambio de posición) la puntuación aumenta.

Aparte de comprobar los cambios de posición, queremos ver como el Safety Car puede influenciar en el movimiento de los coches, para eso añadimos el *Ghost Car*, el cual marca el ritmo medio de los coches y obtenemos restando de la media acumulativa en milisegundos los milisegundos acumulativos. La Figura 7 muestra el diagrama final, en el cual se puede ver claramente como en el caso de Brasil en la vuelta 29 todos los coches se vuelven a juntar, indicado por como las velocidades vuelven a ser iguales ya que el Safety Car lo que hace es agrupar a todos los coches juntos. En el caso de Abu Dhabi, podemos ver como en la vuelta 13 también se juntan los coches pero sabemos por los datos que no hubo Safety Car, esto indica que se ralentizaron por accidente en pista pero el Safety Car no llegó a salir, indicando bandera amarilla i/o Virtual Safety Car.

## 4 Modelado de Machine Learning

Para desarrollar esta sección usamos AutoML Experiments. Nuestra intención es crear un modelo base siguiendo las instrucciones propuestas en la guía y un modelo más pequeño orientado a priorizar las variables que detectamos como relevantes durante la Sección 3.

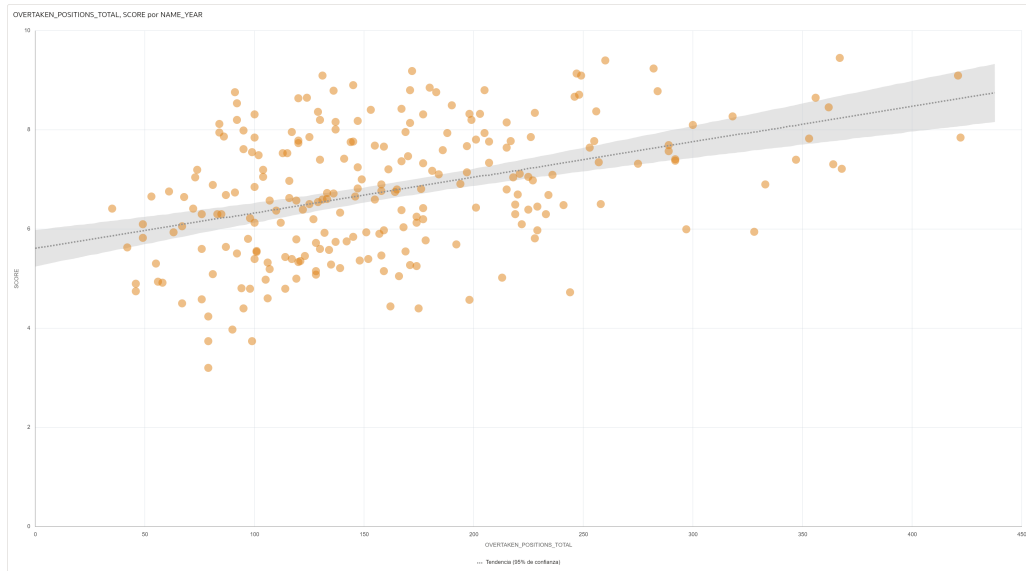


Figure 6: Diagrama de puntos con la correlación entre adelantamientos y la puntuación

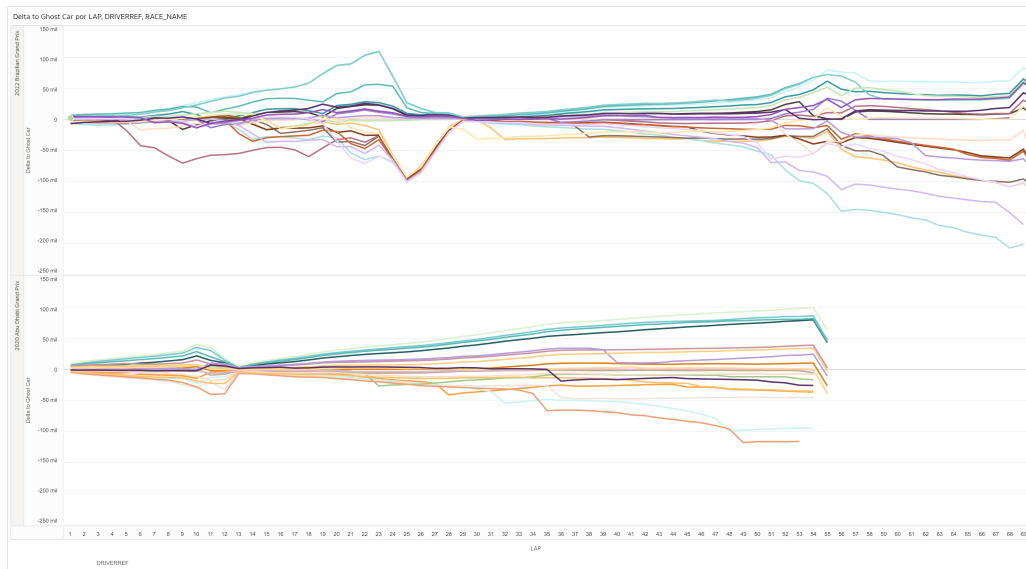


Figure 7: Diagrama comparativo del Ghost Car

## 4.1 Modelo Grande

Para el primer modelo se llevaron a cabo los experimentos usando las siguientes feautres:

- CIRCUITREF
- DNF\_COUNT
- DNF\_DUE\_TO\_ACCIDENT\_COUNT
- LAPS\_WITH\_CHANGE\_POS\_1
- LAPS\_WITH\_CHANGE\_POS\_1\_5
- MAX\_COMEBACKSCORE
- OVERTAKEN\_POSITION\_DUE\_TO\_PITSTOP

- OVERTAKEN\_POSITION\_REAL\_TOTAL
- RANK\_VERSUS\_POSITION
- SAFETY\_CAR
- WEATHER\_WET

Al realizar los experimentos notamos que el modelo que obtuvo el mejor resultado es el SVML (ver Figura 8), motivo por el cual lo usamos para llevar a cabo las predicciones de puntuación para el 2021 descritas en la Sección 5.

Máquina de vector de soporte (lineal)	SVML_CE1D4D48C	0.7891
Modelo lineal generalizado (regresión Ridge)	GLMR_5135D59D5A	0.7546
Modelo Lineal Generalizado	GLM_DDAFDA7163	0.7252
Red neuronal	NN_CAA49892CD	0.6943
Máquina de vector de soporte (Gauss)	SVMG_73B1856D62	0.6466

Figure 8: Resultado de Experimentos con más Features

Los experimentos fueron hechos usando MSE como métrica y se obtuvieron los resultados descritos en la Figura 9 para el año 2020. Notamos que las predicciones (Linea Azul) son bastante acertadas con respecto al *Score* (Linea Naranja), estando a menos de una unidad de diferencia en para cada carrera.

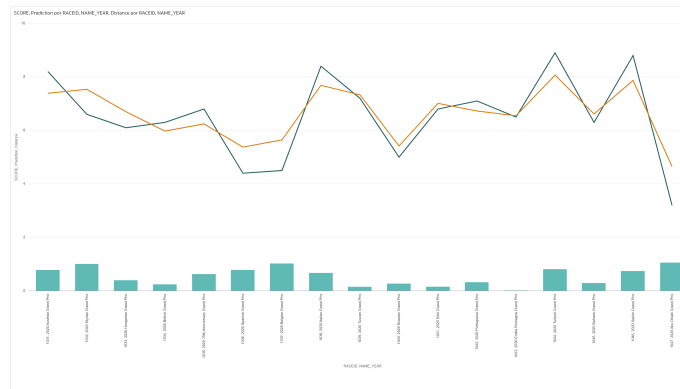


Figure 9: Predicciones de puntuaciones 2020 con la distancia entre lo esperado y la predicción

## 4.2 Modelo Pequeño

Para el construir el segundo modelo se llevaron a cabo los experimentos usando las siguientes features:

- DNF\_COUNT
- OVERTAKEN\_TOTAL
- LAPS\_WITH\_CHANGE\_POS\_1\_5
- LAPS\_WITH\_CHANGE\_POS\_1
- SAFETY\_CAR
- WEATHER\_WET
- OVERTAKEN\_POSITION\_DUE\_TO\_PITSTOP\_TOTAL



En este caso nuevamente SVML resulta ser el modelo con menor error (ver Figura 10), por lo que lo usaremos para hacer las predicciones del 2020 (ver Figura 11) y del 2021 (ver Sección 5).

Al hacer las predicciones para el 2020 notamos que hay más error al compararse con los resultados obtenidos con el modelo más grande en la Sección 4.1. Esto claramente se debe al trade-off entre la cantidad de datos usada contra la precisión.

En caso de que haya limitaciones de computo, recomendaríamos usar el modelo pequeño, ya que es más rápido de entrenar y que a pesar del error introducido por usar menos datos sigue las mismas tendencias que el modelo grande. Es decir, aún si no proporciona un número tan preciso como el modelo anterior, conserva las relaciones de preferencia. Esto lo notamos ya que el orden de las carreras por *Score* es casi el mismo, lo cual es evidente al ver la forma de la curva de predicciones (Línea Naranja).

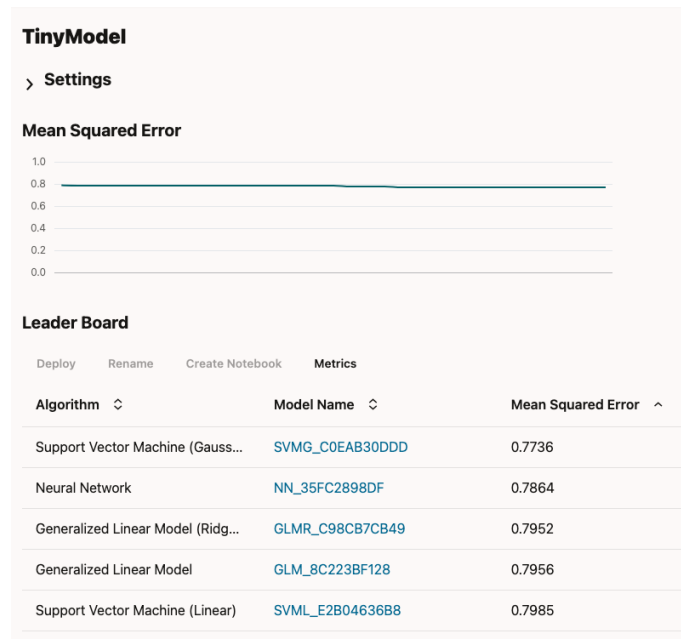


Figure 10: Resultado de Experimentos con menos Features

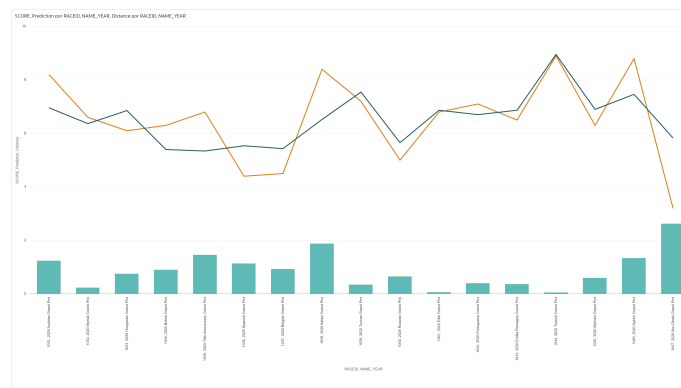


Figure 11: Predicciones de puntuaciones 2020 con menos variables

## 5 Predicción de puntuaciones 2021

Finalmente, usamos nuestros modelos para realizar las predicciones para el año 2021 usando el modelo grande (Linea Azul) y pequeño (Linea Naranja) descritos anteriormente. En la Figura 12 se pueden ver las predicciones que hemos obtenido. De los resultados podemos extraer que la mejor carrera debería ser la de Hungría con un aproximadamente un 8.4 sobre 10, mientras que la peor la de Qatar con aproximadamente un 4 sobre 10.

Notamos que ambos modelos siguen las mismas tendencias en lo que respecta al puntaje de las carreras, lo cual es una buena señal sobre la selección de variables del modelo más pequeño, ya que no aparenta haber perdido mucha robustez a comparación del más grande. Sin embargo, es importante notar que la diferencia absoluta entre ambos aún es grande para algunas carreras, como es el caso de las de Qatar, Belgia y Rusia.

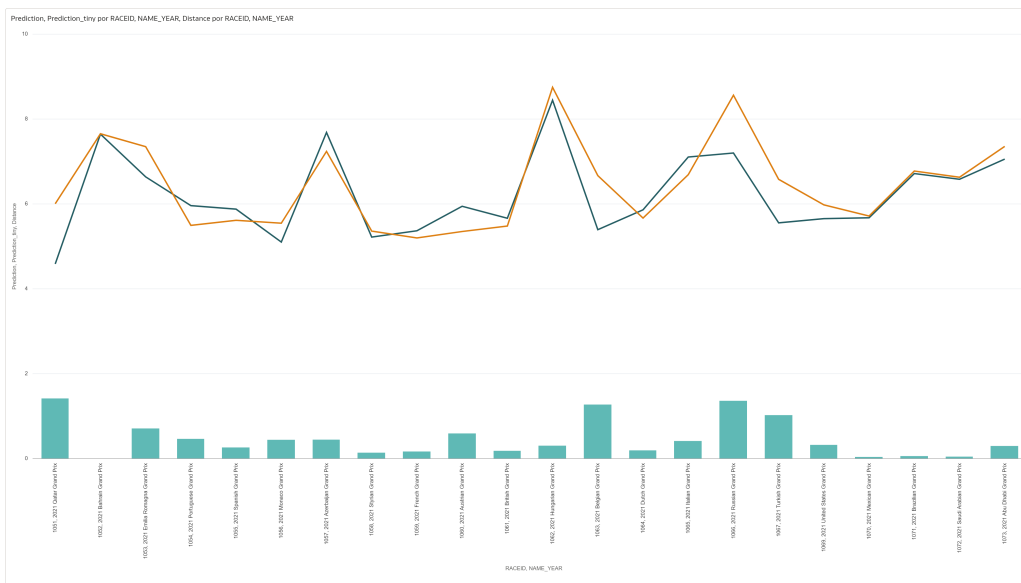


Figure 12: Predicciones de puntuaciones 2021 con los dos modelos

## 6 Conclusions

Durante el proceso de desarrollo pudimos aprender varias lecciones valiosas. Para empezar, tener una noción sobre el dominio al cual pertenecen los datos facilita el proceso de análisis de datos, al permitirte como desarrollador hacer preguntas más atinadas e interesantes sobre los datos que tienes a tu disposición.

El proceso de análisis de datos es una base fundamental del proceso de machine learning, ya que si no se conduce cuidadosamente pueden malgastarse recursos a la hora de entrenar, por hacer uso de datos innecesarios, y estos datos innecesarios pueden incluso llegar a disminuir el rendimiento del modelo.

Finalmente, también pudimos notar cómo el uso de un stack de desarrollo agiliza bastante el proceso de desarrollo y despliegue de soluciones de inteligencia artificial. Nuestra experiencia anterior como equipo ha sido usando herramientas de pequeña escala, entrenando modelos en nuestras máquinas personales y haciendo nuestros gráficos en Jupyter Labs. Estas herramientas son buenas, pero cuando

los modelos empiezan a crecer en tamaño deja de ser viable desarrollar usando solo estas. Disfrutamos mucho de poder experimentar con Oracle.

## References

- [1] Lydia Mee. 2023 formula 1 season smashes viewership records: ESPN's second-highest watched ever. <https://www.si.com/fannation/racing/f1briefings/news/2023-formula-1-season-smashes-viewership-records-espns-second-highest-watched-ever-lm22>, November 2023. Accessed: 2024-6-2.