F1 Q-Learning

Introducción a la inteligencia artificial

Profesor: Arles Ernesto Rodriguez Portela

Tania Julieth Araque Dueñas

Brayan Manuel Rubiano Paramo

Deiver Jair Bernal Garzón

Universidad Nacional de Colombia

Bogotá

Tabla de contenido

1. Descripción del Problema

El problema central es la **toma de decisiones estratégicas** de un piloto (Lando Norris) en un entorno de alta incertidumbre (una carrera de Fórmula 1) en un momento crítico: la **carrera final por el Campeonato Mundial**.

El **agente (Norris)** debe elegir una de **tres estrategias de carrera (Conservadora, Normal o Agresiva)** al inicio de la competición, basándose únicamente en su **posición de salida (estado)**. Esta elección impacta directamente en su resultado probabilístico y, en última instancia, en el resultado del campeonato.

1.1. Descripción General

El proyecto consiste en **simular la carrera final del campeonato** para determinar la **política óptima de carrera** (π^*) que **maximice la probabilidad de que Lando Norris gane el título mundial**.

Utilizamos un entorno de **Aprendizaje por Refuerzo (RL)**, modelado como un proceso de decisión de un solo paso, donde un **agente de Q-Learning** aprende el valor esperado de cada combinación de posición de salida y estrategia elegida.

1.2. Justificación

El uso de RL y Q-Learning se justifica por tres razones principales:

1. **Naturaleza Estocástica**: La F1 es inherentemente probabilística (accidentes, fallas mecánicas, rendimiento variable). El Q-Learning maneja y cuantifica de forma robusta esta incertidumbre.
2. **Optimización de Riesgo/Recompensa**: El algoritmo calcula objetivamente qué estrategia (ej. la agresiva con mayor recompensa pero también mayor riesgo de abandono) es la mejor opción en cada estado inicial.
3. **Cuantificación de la Decisión**: El resultado es una **Q-table** que ofrece una métrica clara (**tasa de victoria esperada**) para cada decisión posible, eliminando la subjetividad.

1.2.1. Modelado de Incertidumbre Estocástica

La incertidumbre se modela mediante **distribuciones de probabilidad** para las posiciones de llegada de los pilotos.

- **Norris:** Su distribución de llegada está condicionada por el **estado (posición de salida) y la acción (estrategia elegida)**. Por ejemplo, una estrategia agresiva aumenta la probabilidad de P1, pero también de un DNF (Posición 20).
- **Rivales (Verstappen/Piastri):** Sus **distribuciones de llegada son fijas e independientes** de la acción de Norris, representando su nivel de rendimiento base en la carrera.

1.2.2. Búsqueda de Política Óptima

La **política óptima** (π^*) se determina después de miles de **episodios de entrenamiento** del agente Q-Learning.

La política se define como la **acción** a^* **que maximiza el valor Q** en cada estado :

$$\pi^*(s) = arg\,max\,Q(s, a)$$

El valor convergente representa la **probabilidad estimada de ganar el Campeonato Mundial** al elegir la acción en el estado inicial.

1.2.3. Dualidad de Objetivos (WDC vs Pista)

El proceso de decisión de Norris implica una dualidad crítica entre dos objetivos que pueden ser mutuamente excluyentes:

1. **Objetivo de Pista (Corto Plazo):** Maximizar la posición de llegada en la carrera individual.
2. **Objetivo de WDC (Largo Plazo):** Garantizar la mayor cantidad de puntos posible para superar a los rivales y ganar el campeonato global.

Enfoque de la Simulación

El agente de Q-Learning será ejecutado bajo dos criterios de recompensa distintos para ilustrar esta dualidad:

Criterio de Recompensa	Definición	Interpretación en la Q-Table
Largo Plazo	$reward = 1$ si el Agente Norris gana el Campeonato	Probabilidad de Ganar el WDC
Corto Plazo	$reward = 1$ si el Agente Norris gana la Carrera Actual	Probabilidad de Superar a los Rivales en la Carrera

1.3. Objetivo Principal del Proyecto

El objetivo principal de este proyecto es **encontrar la política de estrategia óptima** para Lando Norris bajo dos criterios de rendimiento distintos en la carrera final del campeonato.

Específicamente, se busca:

- **Determinar la estrategia óptima** que maximiza la probabilidad de ganar el **Campeonato Mundial de Pilotos (WDC)** (Objetivo de Largo Plazo).
- **Determinar la estrategia óptima** que maximiza la probabilidad de **obtener el mejor resultado en la carrera** (Objetivo de Pista/Corto Plazo).

2. Formulación del MDP

El problema de la estrategia de carrera de Norris se modela como un **Proceso de Decisión de Markov (MDP)**, el marco teórico fundamental para el Aprendizaje por Refuerzo. Un MDP se define por la tupla (S, A, P, R, γ), donde S son los estados, A son las acciones, P función de transición, R la función de recompensa, y γ es el factor de descuento**.**

2.1. Definición de Estados (S)

El conjunto de estados (**S**) representa la información crítica disponible para Lando Norris al tomar su decisión estratégica, que es su **posición de parrilla para la carrera final**.

índice	Símbolo de Estado	Definición
0	s_0	Pole Position $P1$
1	s_1	Primera Fila $P2 - P3$
2	s_2	Segunda Fila $P4 - P6$

Se tiene entonces un Tamaño de Espacio de Estados $N_s = 3$.

2.2. Definición de Acciones (A)

El conjunto de acciones (**A**) representa las **tres estrategias de carrera** mutuamente excluyentes que Norris puede elegir al inicio, las cuales influyen en su distribución de resultados.

índice	Símbolo de Acción	Definición
0	a_0	Conservadora
1	a_1	Normal
2	a_2	Agresiva

Se tiene entonces un Tamaño de Espacio de Acciones $N_a = 3$.

2.3. Definición y Análisis de la Función de Recompensa

La **función de recompensa** es clave, ya que **determina el objetivo del agente**. Se utiliza una **recompensa binaria** que se otorga solo al final del episodio (la carrera), lo que convierte el problema en una tarea de Episodio Único.

$$R(s, a) \epsilon \{0, 1\}$$

Para el entrenamiento, se evaluarán dos funciones de recompensa distintas. Se presenta a continuación el sistema de puntuación utilizado por la **FIA**

Sistema actual de puntuación en Fórmula 1, que reparte los puntos así:

- 1º clasificado (ganador) 25 puntos
- 2º clasificado 18 puntos
- 3º clasificado 15 puntos
- 4º clasificado 12 puntos
- 5º clasificado 10 puntos
- 6º clasificado 8 puntos

2.3.1. Recompensa Global (WDC)

El objetivo es maximizar la probabilidad de ganar el Campeonato Mundial (WDC) dado el puntaje acumulado.

$$R_{wdc} = \{1 \; si \; pts_{total}(Norris) > max(pts_{total}(Verstappen), \; pts_{total}(Piastri))\}$$

Donde $P_{total}(Piloto)$ Incluye los puntos base más los puntos ganados en la carrera. La **Q-Table** resultante (Q_{WDC})estimará la **probabilidad de ganar el campeonato**.

2.3.2. Recompensa Local (Pista)

El objetivo es maximizar el rendimiento en la carrera sin considerar el contexto del Campeonato Mundial.

$$R_{pista} = \{1 \; si \; pts_{carrera}(Norris) > max(pts_{carrera}(Verstappen), \; pts_{carrera}(Piastri))\}$$

Donde $P_{carrera}(Piloto)$ son los puntos ganados en la carrera simulada. La **Q-Table** resultante (Q_{pista})estimará la **probabilidad de superar a los rivales en la carrera individual**.

```
Q-table:
         Conserv.  Normal  Agresiva
Pole:     [0.45]   [0.52]  [0.48]     ← Desde pole, Normal es mejor
P2-P3:    [0.38]   [0.42]  [0.51]     ← Desde P2-3, Agresiva es mejor
P4-P6:    [0.31]   [0.39]  [0.35]     ← Desde P4-6, Normal es mejor
```

3. Modelado de Transiciones y Probabilidades

3.1. Reglas de Transición y Probabilidades de Finalización

Dado que el entorno es un **MDP de un solo paso**, la transición es siempre del estado inicial a un estado terminal virtual, donde se recibe la recompensa.

La función de transición se centra en el resultado probabilístico de la posición final de Norris.

Las probabilidades se basan en la tabla **sample_finish_norris(state, action)** y varían dramáticamente según la estrategia elegida (a) y el punto de partida (s).

Se ilustran algunas:

Estado	Acción	Distribución de Posiciones	Observación
$s_0 - P1$	$a_0 - Conserv.$	$P1\ (40\%)$ $P2(30\%)$ $P3(20\%)$ $P4(10\%)$	Bajo Riesgo. Prioriza consistencia en el podio.
$s_0 - P1$	$a_2 - Agresiva$	$P1(55\%)$ $P2(20\%)$ . . $P20(5\%)$	Mayorp probabilidad de Victoria ($P1$). Se introduce riesgo ($P20 - DNF$)
$s_2 - P4 - P6$	$a_2 - Agresiva$	$P1(25\%)$	Mayor toma de

		$P2(20\%)$ . . $P20(20\%)$	riesgo al partir desde una posición poco favorable buscando el frente. Alto riesgo involucrado ($P20 - DNF\ 20\%$)

Las posiciones de los rivales (Verstappen y Piastri) son **modeladas de manera independiente**, cada uno con una **distribución de probabilidad fija que no depende de la estrategia de Norris**, reflejando su rendimiento promedio esperado.

3.2. Muestreo de Resultados de Carrera (Simulación)

La **incertidumbre de la carrera** se simula mediante la función **random.choices** (implementada en _get_finish_sample). Este proceso es el corazón del entrenamiento Q-Learning:

1. El agente elige la acción a.
2. El entorno utiliza la distribución de probabilidad correspondiente a $(s,\ a)$ para muestrear una posición final para Norris.
3. Simultáneamente, se muestrean las posiciones finales para Verstappen y Piastri de sus distribuciones fijas.
4. La recompensa (R) se calcula con base en estos tres resultados simulados y los puntos de campeonato totales.

La **repetición de este muestreo a lo largo de miles de episodios** permite al agente de Q-Learning **promediar los resultados** y estimar el verdadero valor esperado (probabilidad) de ganar el WDC para cada par**.**

3.3. Determinación del Factor de Descuento

El Factor de Descuento (γ) es un hiper parámetro que determina la importancia de las recompensas futuras en relación con las recompensas inmediatas. Su valor es:

$$\gamma = 1.0 \; o \; 0.95$$

¿Por qué?

Entorno de Un Solo Paso: Dado que el entorno está configurado para que el episodio termine en un solo paso, **no existen recompensas futuras**.

Efecto Práctico: El término de descuento $\gamma \cdot max_{a'}Q(s', a')$ es siempre cero.

Por lo tanto, el valor elegido para γ no tiene ningún impacto en el resultado o la convergencia de la Q-table en esta simulación.

Valor Formal: Se utiliza γ con los valores mostrados por convención, pero la **fórmula de actualización de Q-Learning** se simplifica a la **estimación del valor promedio de la recompensa inmediata** R.

4. Implementación Algoritmo Q-Learning

4.1. Descripción del Algoritmo

La actualización de la **Q-Table** se realiza mediante la **Ecuación de Bellman de Q-Learning**, que ajusta el valor actual $Q(s, a)$ en cada paso con la diferencia entre el valor esperado (Target) y el valor actual (Error de Predicción).

Esta actualización está dada por:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \cdot [R + max_{a'}Q(s', a') - Q(s, a)]$$

En donde:

- $Q(s, a)$: Valor de Q actual
- α : Tasa de Aprendizaje
- R: Recompensa Inmediata Obtenida
- γ: Factor de Descuento
- $max_{a'}Q(s', a')$: Valor Q máximo del Siguiente Estado.

En este caso, Dado que el entorno es de un solo paso, el **término futuro** $max_{a'}Q(s', a')$ es 0. La **ecuación se simplifica a una estimación de la recompensa promedio**:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \cdot [R - Q(s, a)]$$

El valor de $Q(s, a)$ representa la probabilidad estimada de ganar la recompensa (el WDC o la carrera) al tomar la acción a en el estado s.

4.2. Estrategia de Exploración (Greedy) y Decaimiento

Para asegurar que el agente encuentre la verdadera política óptima, debe haber un **equilibrio** entre **Exploración** (probar acciones nuevas y potencialmente mejores) y **Explotación** (usar la mejor acción conocida). Esto se logra mediante la estrategia ϵ-Greedy:

Exploración: Con una probabilidad ϵ (Epsilon), el agente elige una acción aleatoria.

Explotación: Con una probabilidad $1-\epsilon$, el agente elige la acción codiciosa (greedy), que es la acción a con el mayor valor $Q(s, a)$ conocido.

Respecto al decaimiento, la variable ϵ comienza en un valor alto y se **reduce linealmente** a lo largo del entrenamiento:

$$\epsilon_t = max(\epsilon_{final}, \epsilon_{t-1} - TasaDecaimiento)$$

Este **decaimiento lineal** asegura que:

- Al **inicio** ($\epsilon \approx 1.0$), el agente **explora** ampliamente el espacio de estrategias.
- Al **final** ($\epsilon \approx 0.05$), el agente **explota** predominantemente el conocimiento adquirido en la **Q-Table**, optimizando la política.

4.3. Selección y Justificación de Hiper Parámetros

Los siguientes hiper parámetros fueron seleccionados para garantizar una convergencia rápida y precisa en este entorno simplificado:

Hiper parámetro	Valor	Justificación
Tasa de Aprendizaje (α)	0.1	Permite que el **agente incorpore un 10% del nuevo error** de predicción en cada actualización realizada
Factor de Descuento (γ)	0.95	**Recompensas futuras** (si se llegaran a considerar) son de un valor similar a las inmediatas
ϵ_{inicio}	1.0	**Exploración del 100% al inicio** con el objetivo de probar todas las estrategias
e_{final}	0.05	En las etapas finales del entrenamiento el **agente elige la mejor acción un 95% de las veces. Refinamiento.**

5. Resultados

5.1. Tablas Q Finales (WDC vs Pista)

Se presentan las tablas **Q-Tables en escenarios con 5000 Episodios**.

Evidentemente, estas son aquellas con los **resultados más significativos dentro de varias ejecuciones del algoritmo**. No se tienen 2 tablas iguales dada la naturaleza estocástica del proceso

Q-Table WDC (Global)
Tasa de Victorias Promedio cada 500 Episodios

```
Episodio 500/5000 | Win Rate (últimos 500): 87.60% | Epsilon: 0.905
Episodio 1000/5000 | Win Rate (últimos 500): 88.40% | Epsilon: 0.810
Episodio 1500/5000 | Win Rate (últimos 500): 87.40% | Epsilon: 0.715
Episodio 2000/5000 | Win Rate (últimos 500): 91.60% | Epsilon: 0.620
Episodio 2500/5000 | Win Rate (últimos 500): 91.20% | Epsilon: 0.525
Episodio 3000/5000 | Win Rate (últimos 500): 91.00% | Epsilon: 0.430
Episodio 3500/5000 | Win Rate (últimos 500): 86.40% | Epsilon: 0.335
Episodio 4000/5000 | Win Rate (últimos 500): 88.80% | Epsilon: 0.240
Episodio 4500/5000 | Win Rate (últimos 500): 90.40% | Epsilon: 0.145
Episodio 5000/5000 | Win Rate (últimos 500): 89.00% | Epsilon: 0.050
```

Resultado Final

```
Entrenamiento completado!
Victorias totales: 4459/5000 (89.18%)
Q-table final:
[[0.94826247 0.80284852 0.83507402]
 [0.81277798 0.88317827 0.81531134]
 [0.70416145 0.70499374 0.71114468]]
------------------------------------------------------------

🏆 🏎️ ESTRATEGIA ÓPTIMA APRENDIDA:
============================================================
🥇 Desde Pole Position:
   → Estrategia: Conservadora
   → Probabilidad de victoria: 94.83%
   → Q-values: Cons=0.948, Norm=0.803, Agre=0.835

🥈 Desde P2-P3:
   → Estrategia: Normal
   → Probabilidad de victoria: 88.32%
   → Q-values: Cons=0.813, Norm=0.883, Agre=0.815

🥉 Desde P4-P6:
   → Estrategia: Agresiva
   → Probabilidad de victoria: 71.11%
   → Q-values: Cons=0.704, Norm=0.705, Agre=0.711
```

Q-Table Pista (Local)
Tasa de Victorias Promedio cada 500 Episodios

```
Episodio 500/5000 | Win Rate (últimos 500): 18.60% | Epsilon: 0.905
Episodio 1000/5000 | Win Rate (últimos 500): 22.60% | Epsilon: 0.810
Episodio 1500/5000 | Win Rate (últimos 500): 24.00% | Epsilon: 0.715
Episodio 2000/5000 | Win Rate (últimos 500): 21.80% | Epsilon: 0.620
Episodio 2500/5000 | Win Rate (últimos 500): 24.20% | Epsilon: 0.525
Episodio 3000/5000 | Win Rate (últimos 500): 21.00% | Epsilon: 0.430
Episodio 3500/5000 | Win Rate (últimos 500): 18.40% | Epsilon: 0.335
Episodio 4000/5000 | Win Rate (últimos 500): 22.80% | Epsilon: 0.240
Episodio 4500/5000 | Win Rate (últimos 500): 21.40% | Epsilon: 0.145
Episodio 5000/5000 | Win Rate (últimos 500): 25.00% | Epsilon: 0.050
```

Resultado Final

```
Entrenamiento completado!
Victorias totales: 1099/5000 (21.98%)
Q-table final:
[[0.1048515  0.22276978 0.11573784]
 [0.1369903  0.10490233 0.42753115]
 [0.01668335 0.06955975 0.23393384]]
------------------------------------------------------------

🏆 🏎️ ESTRATEGIA ÓPTIMA APRENDIDA:
============================================================
🥇 Desde Pole Position:
   → Estrategia: Normal
   → Probabilidad de victoria: 22.28%
   → Q-values: Cons=0.105, Norm=0.223, Agre=0.116

🥈 Desde P2-P3:
   → Estrategia: Agresiva
   → Probabilidad de victoria: 42.75%
   → Q-values: Cons=0.137, Norm=0.105, Agre=0.428

🥉 Desde P4-P6:
   → Estrategia: Agresiva
   → Probabilidad de victoria: 23.39%
   → Q-values: Cons=0.017, Norm=0.070, Agre=0.234
```

5.2. Análisis de Resultados (Tasa de Victorias)

5.2.1. Convergencia de la Tasa de Victorias Global (WDC)

La tasa de victoria converge a un valor extremadamente alto **(89%)**.

Este alto valor confirma la ventaja estructural inicial del Agente Norris (408 puntos base). La política óptima aprendida por el agente (Conservadora/Normal en S0 y S1 respectivamente) es una política de gestión de riesgo. El **agente maximiza la recompensa**

protegiendo su colchón de puntos, y la alta tasa de convergencia refleja que esta política de bajo riesgo es casi infalible para ganar el WDC en este modelo.

5.2.2. Convergencia de la Tasa de Victorias Local (Pista)

La tasa de victoria converge a un valor significativamente menor que el del WDC (22%).

El valor más bajo demuestra la verdadera **dificultad de superar a ambos rivales** en una carrera individual, sin la ventaja del campeonato, esto dado por el hecho de que tanto Verstappen como Piastri tienen una **alta probabilidad de ganar la carrera** (35% y 30% respectivamente) desde el inicio.

La tasa de convergencia se estabiliza en la probabilidad real de que el Agente Norris consiga más puntos que Verstappen y Piastri en ese día, con su política de maximización de rendimiento (Normal si parte desde $P1$ dado cuenta con ventaja, Agresiva en los demás casos).

5.2.3. Eficacia Computacional y Episodios de Entrenamiento

Tiempo de Entrenamiento: El entrenamiento es computacionalmente eficiente (rápido) debido a la **simplicidad del MDP** (solo 3 estados y 3 acciones) y su naturaleza de un solo paso.

Episodios: Se utilizaron **5000 episodios**. La convergencia de los resultados de victoria confirma que este número de episodios fue suficiente para que los valores de $Q(s,\ a)$ se estabilizaran, evitando el sobreentrenamiento o la falta de exploración.

6. Conclusiones

Los resultados del entrenamiento del agente de Q-Learning confirman que la **estrategia óptima del Agente Norris es altamente sensible a la definición del objetivo** (Recompensa Global del WDC vs. Recompensa Local de la Pista).

6.1. Política Óptima por Estado

Estado	Posición Inicial	Política Óptima Global π_{WDC}^{*}	Política Óptima Local π_{Pista}^{*}	Comportamiento del Agente
0	$P1$	**Conservadora**	**Normal**	Mantener Ventaja del Campeonato sin riesgo vs Competir sin arriesgar en exceso
1	$P2 - P3$	**Normal**	**Agresiva**	Equilibrio entre Riesgo y Posiciones Altas
2	$P4 - P6$	**Agresiva**	**Agresiva**	**Objetivo**: Superar a los rivales tomando el riesgo de DNF

6.2. Impacto de la Función de Recompensa en la Estrategia

6.2.1. Estrategia para Mantener la Ventaja (WDC)

$P1$: **Conservadora**. El valor de $Q(s_0, Conservadora) = Q(0, 0)$ es el más alto, ya que elimina por completo el 5% de riesgo de DNF que introducen las estrategias Normal y Agresiva. El agente no necesita el pequeño aumento en la probabilidad de $P1$. No hay necesidad de arriesgar dada la ventaja que se tiene.

$P2 - P3$: **Normal**. Se necesita un poco más de empuje que la Conservadora, pero se mantiene un enfoque de bajo riesgo para asegurar los puntos.

$P4 - P6$: **Agresiva**. Es la única estrategia donde el riesgo es aceptable. La necesidad de ganar puntos importantes para compensar la mala posición inicial anula el peligro del DNF (5%). **La opción más arriesgada es la única que ofrece una probabilidad viable de obtener el título desde dichas posiciones iniciales.**

6.2.2. Estrategia para Maximizar Riesgo/Recompensa (Pista)

$P1$: **Normal**. El agente se conforma con la estrategia Normal (50% de P1) en lugar de la Agresiva (55% de P1, con 5% de DNF). Esto sugiere que, aunque el objetivo es local, el 5% de riesgo de DNF aún reduce el valor esperado de la recompensa lo suficiente para que la Normal, más estable, sea la elección óptima. Además, aún se considera el riesgo dado por la alta probabilidad de que los rivales terminen por delante, por lo que una estrategia Conservadora no es viable.

$P2 - P3 \;\; y \;\; P4 - P6$**: Agresiva.** Este resultado confirma que la alta probabilidad de que los rivales obtengan $P1$ o $P2$ (Verstappen 35% P1) obliga a Norris a elegir la acción que maximiza su propia probabilidad de $P1$, incluso desde posiciones cercanas a la punta.

6.3. Limitaciones y Mejoras Futuras

Limitaciones Actuales:

- Modelo simplificado de una carrera con estados discretos
- Distribuciones de probabilidad estimadas, no basadas en datos reales
- No considera factores dinámicos (clima, desgaste, Safety Cars)

Mejoras Propuestas:

Corto plazo:

- Incorporar múltiples carreras restantes
- Agregar estados más detallados (gaps de tiempo, neumáticos)

Largo plazo:

- Implementar Deep Q-Learning para estados continuos
- Integrar datos históricos reales de F1
- Modelar múltiples agentes interactuando (Multi-Agent RL)

6.4. Video Funcionamiento

A continuación se encuentra un video del funcionamiento del proyecto:

https://drive.google.com/file/d/1QihqV09MbbzD0GW9i46cjithev8GZNao/view?usp=sharing

6.5. Enlace al Repositorio

A continuación, se proporciona el enlace directo al código fuente, donde se pueden revisar la claridad del código, la estructura e implementación

https://github.com/TaniaAraque/Proyecto-3-IA-F1-ML

7. Referencias

- Curso Introducción a la Inteligencia artificial, Universidad Nacional, 2025-2
- https://es.motorsport.com/f1/news/sistema-puntos-posiciones-reparto-formula1/6505476/