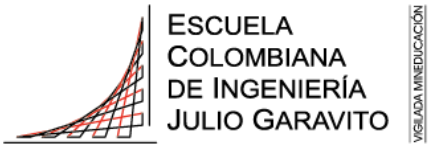
MODELO DE TOMA DE DECISIONES UTILIZANDO APRENDIZAJE POR REFUERZO CUÁNTICO

SANTIAGO SASTOQUE GRANADOS

Trabajo Dirigido

Tutor
Juan Manuel López López PhD

Contutora
Laura Andrea León Anhuaman PhD



UNIVERSIDAD DEL ROSARIO
ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO
PROGRAMA DE INGENIERÍA BIOMÉDICA
BOGOTÁ D.C
2021

AGRADECIMIENTOS

Agradezco al equipo que tuve en el desarrollo de esta investigación, mis tutores, Juan Manuel López y Laura Andrea León. Fueron bastantes consejos, ideas y tiempo dedicado, su paciencia y experiencia fueron piedras angulares, gracias.

Juan Manuel, han sido varias las experiencias y aprendizajes en esta etapa, el inicio de mi vida académica.

"La naturaleza solo nos muestra la cola del león, pero no tengo duda de que el león pertenece a ella incluso aunque no pueda mostrarse de una vez debido a su enorme tamaño"

Albert Einstien.

Índice General

Índice de figuras

Índice de tablas

Capítulo 1

INTRODUCCIÓN

En este capítulo se presenta la motivación para realizar el proyecto, se exponen los objetivos y las contribuciones de la investigación. Adicionalmente, se introducen brevemente las ideas y conceptos que se desarrollan a lo largo del documento, finalmente, se aclaran algunos aspectos sobre la interpretación de los resultados de esta investigación.

1.1 Motivación

El aprendizaje por refuerzo (Reinforcement Learning, RL por sus siglas en inglés) es un tipo de algoritmo de aprendizaje automático que busca entrenar un modelo computacional con la interacción entre el agente (puede ser más de uno) y un entorno, del cual no es necesario tener información previa y por ende no se requiere tener una base de datos anotada previamente. El RL utiliza un escalar, llamado recompensa, para evaluar el desempeño de cada acción en el entorno y por medio de la optimización de una función, se logra hallar una política (secuencia de acciones) dentro del entorno con las que se consigue una recompensa máxima. Es decir, el agente aprende y explora el entorno en un proceso de prueba y error, de manera similar a la que los humanos aprenden de un entorno desconocido. Este tipo de algoritmos han sido responsables de exitosos avances recientes en el área del aprendizaje automático, por ejemplo, el computador que venció al campeón mundial en el juego Go[1] [1].

El uso de algoritmos de RL se ha incrementado en diferentes aplicaciones, principalmente en robótica [2] [3], debido a la buena adaptación con los sistemas en tiempo real y a su capacidad de aprendizaje de sistemas complejos no lineales [4]. Aunque, el RL también es utilizado en campos como la neurociencia [5], donde se modelan matemáticamente algunos procesos neuronales de interés. Esto con el fin de entrenar modelos y predecir acciones humanas, como también para entender la base conceptual de estos procesos. El RL y la toma de decisiones humanas comparten su influencia en la recompensa, penalidad y el recuerdo de situaciones similares en el pasado, además del hecho de que parte del entendimiento del pensamiento humano está basado en modelos probabilísticos [6], es por esta razón que este proyecto se centra en el estudio del proceso de toma de decisiones y en su modelado a partir de algoritmos de RL.

Los modelos que utilizan probabilidad clásica no explican algunas paradojas del comportamiento, por ejemplo, la falacia de la conjunción y el efecto del orden (para más información, ver sección 2.3), y recientemente se ha encontrado que la teoría de probabilidad cuántica [44] puede dar explicación a estas inconsistencias [6] [7]. Además, existen aspectos de la teoría cuántica que se pueden relacionar estratégicamente con la toma de decisiones para buscar una explicación desde otro punto de vista a estos fenómenos. La idea general que se encuentra a lo largo del documento es que la teoría cuántica aporta más información y entendimiento de la cognición y toma de decisiones humana, que los modelos probabilísticos tradicionales [6], esta idea se desarrollará en base a la evidencia de modelos cuánticos para el comportamiento humano, una prueba

[1] Este evento ocurrió en el año 2016

psicológica diseñada para analizar la toma de decisiones e impulsividad y un algoritmo que simula los principios de la computación cuántica. Combinando el RL y las ventajas de la computación cuántica se propuso el Aprendizaje por refuerzo cuántico (QRL, por sus siglas en inglés) [4], y debido a las similitudes entre la toma de decisiones con la teoría cuántica y los estudios anteriores que relacionan el RL con procesos cognitivos y de toma de decisiones se han propuesto algoritmos de Aprendizaje por refuerzo cuántico aplicados a la toma de decisiones [8]. En esta investigación se implementa este tipo de algoritmos para evaluar si este reciente enfoque puede ser útil para generar nuevos puntos de vista en el modelado de la toma de decisiones humanas, para esto se utiliza un protocolo experimental de toma de decisiones.

En el protocolo experimental diseñado para analizar la toma de decisiones se tuvieron en cuenta los datos de participantes de dos grupos de rangos de edades diferentes, esto con el fin de analizar cómo cambian los resultados entre estos, ya que existe evidencia del aumento de toma de decisiones de alto riesgo en edades jóvenes y la disminución de estas después de los 25 años [9][10]. Las simulaciones de los algoritmos se realizaron en computadores clásicos, inicialmente un algoritmo de QRL en un entorno tipo laberinto, donde el objetivo es que el agente logre llegar a la meta de un laberinto en la menor cantidad de pasos posibles, después de esto se modificó el algoritmo para utilizar los resultados obtenidos en la tarea psicológica y cambiar el entorno para poder crear un modelo por cada participante que pueda predecir las decisiones tomadas, este algoritmo se comparó con cuatro modelos de toma de decisiones diferentes usando RL para evaluar su rendimiento con respecto al estado del arte del RL. Finalmente, como las simulaciones se están realizando en un computador clásico, esto conlleva a preguntarse ¿Si el algoritmo de QRL resulta tener mejores resultados que el de RL, a qué se debe esto? Y ¿Qué aspectos del QRL ayudan a mejorar los resultados?

1.2 Objetivos del proyecto

1.2.1 Objetivo general

Plantear un modelo matemático de toma de decisiones en adultos sanos, utilizando procesamiento digital de señales, y aprendizaje por refuerzo (RL).

1.2.2 Objetivos específicos

1. Plantear un protocolo experimental que permita analizar el proceso de toma de decisiones basadas en el valor a través de un modelo de RL.

2. Diseñar un algoritmo con los resultados del protocolo experimental basado en el aprendizaje por refuerzo.

3. Analizar los resultados obtenidos del algoritmo de aprendizaje por refuerzo en relación con el protocolo experimental planteado.

1.3 Contribuciones

En primera instancia, la implementación del algoritmo de QRL con la prueba Iowa Gambling Task (IGT) (para más información, ver sección 3.1.3) para el análisis de la toma de decisiones en grupos de distintas edades, es la principal contribución de esta investigación y ayuda a ver la relación entre la toma de decisiones e impulsividad con la edad, por otro lado, la implementación de la prueba IGT con el QRL contribuye con la verificación experimental del desempeño de estos algoritmos, pues, se necesitan más estudios que corroboren los resultados debido a lo reciente de la investigación, también es importante relacionar las predicciones teóricas de este marco de referencia con pruebas experimentales, y así encontrar con resultados de más investigaciones nuevos aspectos de la naturaleza de las decisiones humanas.

Esta implementación requiere nuevas ideas con respecto al RL, principalmente porque en este caso el entorno es un proceso cognitivo (toma de decisiones) y es necesario ajustar los procesos de aprendizaje en estos nuevos entornos. Debido a que existen diferentes modelos de toma de decisiones, se prueban los cuatro que reportan mejores resultados para RL [8] y la implementación de QRL, con el fin de obtener los mejores resultados posibles y al mismo tiempo, poder definir cuál de los modelos es más efectivo para ser utilizado en futuras investigaciones que involucren igualmente la prueba IGT.

Para poder comparar los dos algoritmos basados en RL y QRL, se utilizó un algoritmo base de RL en el entorno tipo laberinto, lo que permite comparar ambos algoritmos en igualdad de condiciones del entorno y analizar las mismas variables de resultados. Finalmente, este estudio muestra la utilidad y ventajas de simular circuitos cuánticos incluso en computadores clásicos (donde no se aprovechan las ventajas en los tiempos de procesamiento y otras características de las implementaciones físicas) y abrir las posibilidades a próximas investigaciones utilizando estos conceptos [8].

1.4 Interpretación de los resultados

Los modelos cuánticos de cognición y toma de decisiones son un campo de investigación activo [6], y es importante aclarar que el objetivo de este enfoque para el estudio de estos fenómenos es utilizar un marco teórico innovador que permite explicar inconsistencias en los modelos clásicos de toma de decisiones. También se puede utilizar como una nueva fuente de herramientas, pero no se intenta modelar el cerebro usando mecánica cuántica, ni comparar el cerebro con un computador cuántico. Este tipo de idealizaciones son usadas todo el tiempo, por ejemplo, en modelos de dinámica neuronal del cerebro, donde ecuaciones diferenciales son usadas para modelar el crecimiento de la activación neuronal, aunque solo existen un número finito de neuronas [6]. Es decir, los modelos cuánticos de toma de decisiones aportan nuevas perspectivas y abstracciones a problemas existentes sin asumir que estos procesos físicos están ocurriendo en el cerebro, por otro lado, cómo estos mecanismos cuánticos pueden ser factibles en el cerebro es aún una pregunta abierta [6] [8].

Capítulo 2

MARCO TEÓRICO

2.1 Aprendizaje por Refuerzo Clásico

Los métodos de aprendizaje automático están usualmente clasificados en aprendizaje supervisado, no supervisado y por refuerzo. Los dos primeros tienen que pasar por un proceso de entrenamiento en el que utilizan la información de una base de datos que contiene muestras relevantes para resolver el problema que se desee (ya sea de clasificación o de regresión). Por otro lado, el aprendizaje por refuerzo busca aprender de la interacción con el entorno, este método está más relacionado con la naturaleza del aprendizaje, ya que, esta interacción provee información de causa y efecto necesaria para considerar tomar de nuevo una misma acción, si lo que se desea es lograr un objetivo, adicionalmente, recordar esta información es útil para formar un conjunto de acciones que facilite alcanzar el objetivo.

El hecho de que las respuestas del entorno a las decisiones tomadas modifiquen el comportamiento humano crea una conexión entre las teorías de aprendizaje y el RL (el RL aplicado a la toma de decisiones se expone en la sección 2.2). El objetivo del aprendizaje por refuerzo es aprender a tomar un conjunto de acciones (toma de decisiones), que maximicen una recompensa, estas acciones se deben descubrir por medio de prueba y error en el entorno; esto da paso a uno de los desafíos más importantes en el RL: El equilibrio entre la exploración y la explotación. Para obtener una buena recompensa el agente debe preferir acciones que ha tomado anteriormente y han mostrado ser efectivas en producir recompensa, pero para encontrar esas acciones se deben probar acciones que no se han tomado anteriormente. El agente debe *explotar* la experiencia que ha obtenido de previas acciones-recompensas, pero igualmente debe *explorar* para encontrar mejores acciones en el futuro, el dilema es encontrar el equilibrio entre ambos conceptos. Este dilema es de interés en el estudio de procesos estocásticos y es aún un problema abierto de las matemáticas [11]. Otro aspecto diferenciador del RL es que considera explícitamente el problema que se desea resolver, es decir, el agente está orientado al objetivo, a diferencia de otros enfoques de aprendizaje automático donde se crean subproblemas o capas que están orientadas a objetivos diferentes al objetivo general del algoritmo y muchas veces se desconoce cómo estos subprocesos pueden ayudar al propósito general. Otro desafío del RL surge cuando la complejidad del entorno aumenta y las posibles acciones con él, la cantidad de parámetros a modificar en el modelo aumenta exponencialmente con la dimensionalidad de los datos, esto se conoce en aprendizaje automático como ‘La maldición de la dimensionalidad’ [4].

El RL ha interactuado estrechamente con la psicología y la neurociencia, ya que, es el paradigma de aprendizaje automático más cercano al tipo de aprendizaje de los humanos y otros animales [11]. Esta relación ha traído beneficios en ambos sentidos. Varios de los algoritmos centrales de RL fueron inspirados en sistemas de aprendizaje biológicos e igualmente el RL ha modelado procesos de aprendizaje que han brindado información del funcionamiento de estos procesos [11]. Ya establecida la importancia de la toma de decisiones en el RL, se procede a describir con más formalidad la teoría y técnicas de RL

utilizadas para lograr el objetivo de la investigación: modelar la toma de decisiones con los principios del RL.

La formulación general de los problemas que se desean resolver con RL se puede describir mediante los Procesos de Decisión de Markov (MDP por sus siglas en inglés). Este marco teórico representa matemáticamente la interacción entre le agente y el entorno.

2.1.1 Procesos Finitos de Decisión de Markov

Cuando la cantidad de estados posibles en el entorno es finita y las acciones en el entorno no solo afectan la recompensa inmediata sino también las acciones siguientes, la formalización de los problemas que involucran toma de decisiones secuenciales se pueden representar con procesos finitos de decisión de Markov. Los elementos que se pueden identificar en un MDP son:

i. Agente: hace referencia al aprendiz; el encargado de tomar las decisiones y de recibir señales externas para modificar su comportamiento.

ii. Entorno: Es con lo que interactúa el agente; define las acciones que puede tomar y genera las señales importantes en el proceso de aprendizaje, además, en algunos casos el entorno puede reaccionar a las acciones que toma el agente.

iii. Política: Corresponde a un conjunto de acciones que el agente toma en un tiempo dado, es decir, la política define el comportamiento del agente. Corresponde a reglas de respuestas ante determinadas situaciones y deben ser estocásticas; especificando probabilidades para tomar cada acción [11].

iv. Recompensa: Es una señal que surge de la interacción entre le agente y el entorno; por cada paso del agente en el entorno, este envía al agente un escalar llamado recompensa y da información del desempeño del agente para resolver el problema. La recompensa define la necesidad de actualizar las políticas

v. Función de valor: Esta función evaluada en un estado dentro del entorno entrega la cantidad total de recompensa que el agente puede esperar acumular en el futuro, si se inicia en ese estado. Esta función representa el desempeño a largo plazo que el agente puede tener siguiendo una política dada. Un valor alto en un estado representa mayor ganancia de recompensas a largo plazo, es decir, si se maximiza esta función, se están maximizando las ganancias [11].

En la Figura 2.1 se muestra cómo es el flujo de información y las interacciones entre los elementos de un MDP. Por cada paso t del agente en el entorno, tiene que realizar una *acción*, $A_t \in \mathcal{A}(s)$, a lo cual el entorno responde devolviendo la recompensa de la acción, $R_{t+1} \in \mathcal{R} \subset \mathbb{R}$[2], y el estado en el que quedó el entorno luego de la acción, $S_{t+1} \in \mathcal{S}$. Con esta información se actualiza la función de valor y de ser necesario también las políticas, y finalmente se escoge la siguiente acción A_{t+1} en base a las políticas π (experiencia obtenida) [11]. Para simplicidad se asume que el agente interactúa con el entorno en pasos

[2] Se utiliza R_{t+1} para denotar la recompensa obtenida debido a A_t

de tiempo discretos, $t = 0,1,2,3, \ldots$,[3]. Los conjuntos de estados y acciones posibles del entorno tienen elementos finitos, por esto se consideran los procesos finitos de decisión de Markov. Donde $\mathcal{S}$ es el espacio de estados; $\mathcal{A}(s)$ es el espacio de acciones para el estado s y se define una política como una secuencia: $\pi = (\pi_0, \pi_1, \ldots)$, que representa probabilidades de seleccionar una acción especifica dado un estado [4].

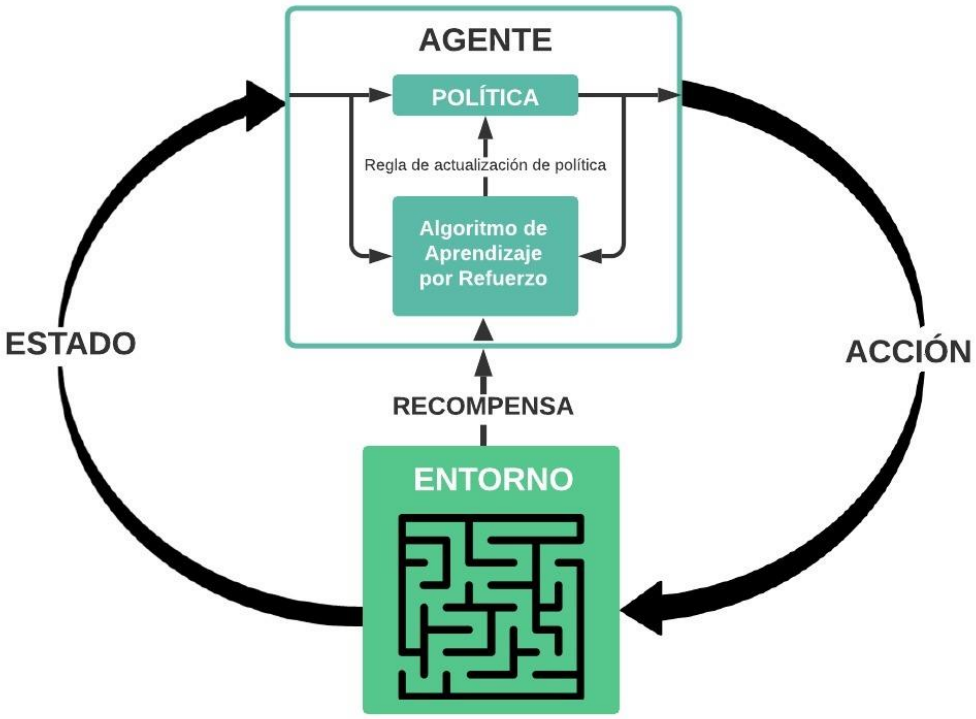


Figura 2.1: Interacción agente-entorno en un proceso de decisión de Markov

El proceso comienza a dar una secuencia que empieza de esta manera:

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \ldots \tag{1}$$

Como los procesos son estocásticos, se debe definir la probabilidad de obtener un par de valores, s' y r en un tiempo t, dados los valores del estado y acción anteriores:

$$p(s', r \mid s, a) \equiv Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\}, \tag{2}$$

$\forall\, s', s \in \mathcal{S},\ r \in \mathcal{R}$ y $a \in \mathcal{A}(s)$, esta distribución condicional p define la dinámica del MDP [11], y es una distribución de probabilidad para cada elección de a y s, es decir:

$$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r \mid s, a) = 1, \quad \forall\, s' \in \mathcal{S}, r \in \mathcal{R} \tag{3}$$

[3] Las técnicas de RL utilizadas en esta investigación se pueden implementar en tiempo continuo [12]

Esta representación de MDP se puede utilizar para describir una gran variedad de problemas, de especial interés para esta investigación se utiliza el caso en el que el entorno es un conjunto de decisiones tomadas por un participante[4].

El RL busca maximizar la obtención de las recompensas acumuladas en un cierto número de pasos en el entorno, específicamente la secuencia de recompensas obtenidas hasta que el agente llegue a un estado terminal del entorno, estos se denominan *episodios*, por ejemplo, un episodio en un entorno tipo laberinto es un viaje desde el estado inicial hasta el estado final en el laberinto. La suma de recompensas en el episodio se puede representar así:

$$G \equiv R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^k R_{t+k+1} \tag{4}$$

Donde γ es un *factor de descuento*, $0 \geq \gamma \leq 1$ y determina el valor actual de futuras recompensas, es decir, se les da menos valor a las recompensas mientras más alejadas estén del estado actual [11]. Para tareas que se necesiten de muchos pasos para lograr el objetivo, se utiliza un valor γ cercano a uno para tener en cuenta las recompensas futuras con más valor.

La *función de valor* de un estado s bajo una política π, se denota $v_\pi(s)$ y es la recompensa esperada a largo plazo empezando desde s y siguiendo π. Esta función se define así:

$$\begin{aligned} v_\pi(s) &\equiv \mathbb{E}_\pi[G_t \mid S_t = s\,] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s\,] \\ &= \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma \mathbb{E}_\pi[G_{t+1} \mid S_{t+1} = s']\right] \\ &= \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_\pi(s')] \end{aligned} \tag{5}$$

Para encontrar las políticas optimas (π_*), se debe optimizar la función de valor:

$$v_*(s) = \max_\pi v_\pi(s) \tag{6}$$

$$v_*(s) = \max_a \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_*(s')] \tag{7}$$

La ecuación (7) es la llamada *ecuación óptima de Bellman* para v_* [4].

[4] Las decisiones y otros datos son obtenidos de la prueba IGT

2.1.2 Aprendizaje por *Diferencia – Temporal* (TD)

El Aprendizaje por Diferencia – Temporal (TD por sus siglas en inglés), es una técnica para solucionar procesos finitos de decisión de Markov, es considerado uno de los enfoques más innovadores y de importancia en la historia del RL, su nombre se debe a que el concepto fundamental para su funcionamiento es la diferencia entre estimaciones temporalmente sucesivas de la misma cantidad. Los orígenes del aprendizaje por TD son en parte la psicología del aprendizaje animal aplicados en la inteligencia artificial, por esta razón también se ha utilizado el aprendizaje por TD en modelos psicológicos de condicionamiento clásico y se ha demostrado que este tipo de aprendizaje es apropiado para solucionar problemas de reconocimiento de patrones donde la información está organizada temporalmente [13].

Sutton [13] sugirió similitudes entre el aprendizaje por TD y el proceso de aprendizaje en los animales, lo cual hace que esta sea la técnica apropiada para solucionar el problema de RL de esta investigación, ya que la prueba IGT es realizada por humanos y sus datos son secuenciales.

El objetivo de este método es predecir valores futuros de v_π basados en la experiencia aprendida y los datos actuales, una de las ventajas más grandes del aprendizaje por TD es que puede hacer predicciones por cada paso en el entorno y, por ende, actualizar las políticas más seguido que otros métodos, además, no se necesita de un modelo del entorno para hacer predicciones. La regla de actualización de TD para $v(S_t)$ es:

$$v(S_t) \leftarrow v(S_t) + \alpha[r + \gamma v(S_{t+1}) - v(S_t)] \tag{8}$$

Donde $\alpha \in (0,1)$ es el factor de aprendizaje. Este tipo de algoritmo de aprendizaje por TD, se denomina $TD(0)$ porque únicamente tiene en cuenta las diferencias entre el estado actual S_t y los posibles estados siguientes S_{t+1}.

Como este método proporciona una estimación, es importante aclarar que para cualquier política π, se ha demostrado que $TD(0)$ converge a v_π con el factor de aprendizaje apropiado [11].

Debido a las relaciones entre RL y el aprendizaje en humanos, este tipo de algoritmos constituyen un enfoque importante para la inteligencia artificial. Existen otros algoritmos efectivos de RL como Q-learning, $TD(\lambda)$, Sarsa, entre otros [4].[5]

2.2 Toma de decisiones

El estudio de la toma de decisiones abarca varios campos como la neurociencia, psicología, economía, estadística, ciencias computacionales, entre otros [14]. A pesar de todas las aplicaciones que tiene, los componentes de las decisiones son los mismos y se encuentran en diferentes situaciones en la naturaleza, por esta razón el entendimiento conceptual de los procesos de toma de decisiones ayuda a modelar procesos biológicos. La inteligencia

[5] Para más información, revisar [11]

artificial se ha beneficiado de los avances en el entendimiento del aprendizaje humano y a su vez también ha ayudado a entender las variables involucradas en la toma de decisiones. Para esta investigación se presentan tres puntos de vista de la toma de decisiones, el primero hace referencia a la base neuronal de la formación y los componentes de un tipo especial de decisiones; la toma de decisiones basadas en el valor, las cuales son las que un agente en un algoritmo de RL tiene que tomar. El segundo enfoque gira en torno a la aplicación de pruebas psicológicas para comprobar hipótesis relacionadas con la toma de decisiones, ya que, existen muchos factores que influyen en la toma de decisiones y son de interés, incluso para ver el efecto de algunas enfermedades como la depresión y la adicción [15][16], además, se puede comparar la toma de decisiones en diferentes poblaciones. Finalmente, en el tercero se presentan los modelos utilizados para explicar y simular la toma de decisiones basadas en el valor y como se relacionan con el RL.[6]

2.2.1 Base neuronal de la toma de decisiones

La toma de decisiones en animales es el resultado de un proceso cognitivo que involucra escoger una acción, este proceso depende de la memoria (experiencia con situaciones similares en el pasado), percepción (la interpretación de estímulos sensoriales), entre otros [14].

2.2.1.1 Toma de decisiones basadas en el valor (*Value-Based Decision Making*)

Este tipo de decisiones está basado principalmente en el valor subjetivo que se asocia con cada una de las posibles alternativas, con estas decisiones se busca estudiar cómo el cerebro asigna, almacena y usa valores para tomar decisiones [17]. Las correlaciones neurobiológicas del valor de las decisiones se han descrito en la corteza orbitofrontal, la corteza cingulada y los ganglios basales, áreas del cerebro tradicionalmente relacionadas con comportamientos de búsqueda de recompensas y toma de decisiones [18]. Por ejemplo, se ha propuesto que la corteza orbitofrontal se encuentra implicada en la formación de expectativas, un proceso importante para actualizar los valores de las posibles alternativas, también se encuentra relacionada con la regulación de la planificación conductual asociada a la sensibilidad, a la recompensa y a las penalidades[7] [19] [20]. Por otro lado se ha propuesto que los ganglios basales interactúan en el proceso de toma decisiones como un mecanismo de selección basado en los valores de las posibles acciones [21].

Debido al impacto de la toma de decisiones en el comportamiento humano, se considera un sello distintivo de la cognición de alto nivel y el foco central de la Neuroeconomía; un campo emergente interdisciplinario que busca estudiar los cálculos que realiza el cerebro para tomar una decisión basada en el valor [24].

Los procesos requeridos para la toma de decisiones basadas en valor se pueden dividir en cinco principales, estos se muestran en la Figura 2.2 y están basados en los modelos teóricos de toma de decisiones [22]. En estos procesos se puede apreciar la relación de la toma de decisiones con el RL y también cómo la toma de decisiones es un proceso

[6] En este estudio también se relaciona a la toma de decisiones basadas en el valor con QRL
[7] La sensibilidad a la recompensa y a las penalidades son parámetros en los modelos de toma de decisiones utilizados en esta investigación

adaptativo, la pregunta que intentan responder estos modelos es: ¿Cómo los humanos y animales adquieren preferencias por diferentes acciones y recompensas?, y aunque es necesario realizar más estudios al respecto de cómo estos mecanismos se implementan en el cerebro se tiene claro que parte de la respuesta es el aprendizaje, en situaciones normales las actualizaciones que se dan durante el aprendizaje mejoran la calidad de futuras decisiones [23].

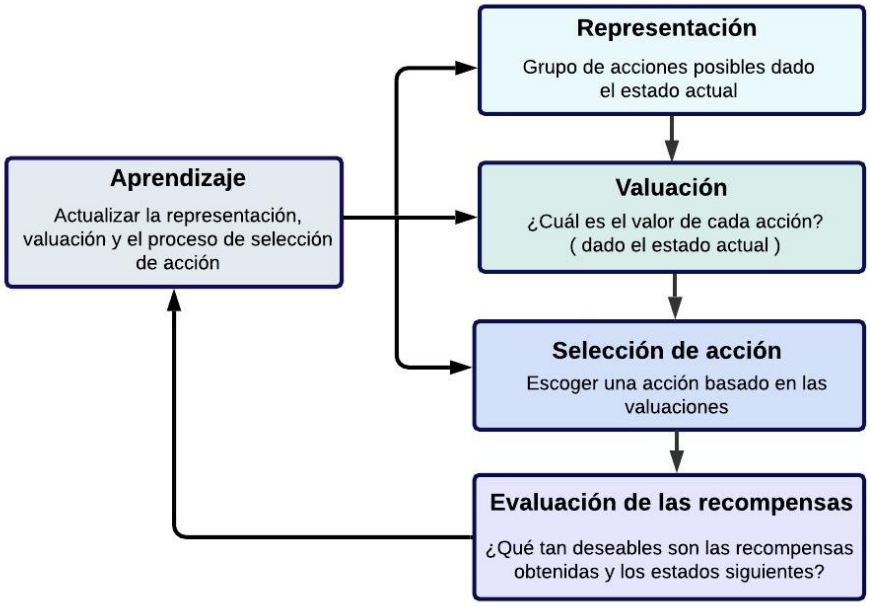


Figura 2.2: Procesos principales de la toma de decisiones basadas en valor

2.2.2 Pruebas psicológicas para el estudio de toma de decisiones

Las pruebas psicológicas como el IGT o el Soochow Gambling Task (SGT) han sido de gran importancia para el progreso en el estudio del comportamiento humano durante la toma de decisiones [25]. Estas pruebas son simulaciones de tareas de azar y el desempeño de los participantes en las pruebas revela información de procesos psicológicos presentes durante la realización de la prueba, como aprendizaje por experiencia y procesos motivacionales como sensibilidad a la recompensa y a la penalidad.

Inicialmente el IGT fue desarrollada para evaluar el grado de déficit en la toma de decisiones en pacientes con daños en la corteza prefrontal ventromedial. Esta área del cerebro está involucrada en el procesamiento de riegos, miedo, emoción y toma de decisiones [20]. Esta prueba posibilitó por primera vez detectar pacientes con déficits en la toma de decisiones en la vida real [20].

La prueba comúnmente se ha utilizado para comparar participantes sanos con participantes adictos a las drogas o con daños cerebrales, resultando en diferencias significativas entre los dos grupos, confirmando el déficit en la toma de decisiones, también, se encontró que estos participantes son inconscientes respecto a las consecuencias futuras de sus actos y

parece que se dejan llevar únicamente por los impulsos inmediatos [16] [26]. Incluso se ha comprobado que diferentes tipos de drogas afectan de distintas formas la toma de decisiones [16].

Existen otros patrones que pueden influenciar en la toma de decisiones que no tienen que ver con enfermedades. Se mostró que durante la adolescencia (12 a 18 años) y un periodo llamado adolescencia tardía (hasta los 25 años), la toma de decisiones riesgosas o de alto riesgo aumenta en comparación a otras edades, lo cual está relacionado con la exploración de decisiones novedosas o sin experimentar. Se ha demostrado que este comportamiento es transitorio para la mayoría de los individuos, es decir, algunos adultos siguen experimentando estos comportamientos de alto riesgo. También se reporta que esta toma de decisiones de alto riesgo puede ser beneficiosa y necesaria en términos de la exploración y experiencia para tomar futuras decisiones [9] [10]. Finalmente, se reportan pocos estudios en la literatura del análisis de la toma de decisiones en grupos sanos de diferentes edades con el IGT.

Se ha demostrado que se obtienen diferentes desempeños según el tipo de prueba psicológica, estas se pueden catalogar en: *independientes de la elección* (pruebas en las que las recompensas no son afectas por elecciones anteriores) y *dependientes de la selección* (pruebas en las que las recompensas dependen de la cantidad de veces que una opción ha sido escogida anteriormente), donde, cada población de diferente rango de edad es más afín a obtener mejor desempeño en una de las categorías de pruebas psicológicas de toma de decisiones [27][8].

2.2.3 Modelos cognitivos de toma de decisiones

Desde el éxito del IGT para evaluar experimentalmente la toma de decisiones, se han creado modelos matemáticos, que buscan explicar los diferentes procesos que ocurren durante la toma de decisiones (Figura 2.2). Modelar la toma de decisiones permite dividir el proceso en subcomponentes, y como resultado, quedan parámetros libres que se tienen que ajustar al modelo y pueden ayudar a entender la fuente de los déficits de toma de decisiones, además, tienen capacidad predictiva de las decisiones, lo que permite comparar cuantitativamente el desempeño de diferentes modelos; lo cual es de gran importancia para el progreso científico en la cognición [28][29].

Desde el punto de vista de esta investigación, el IGT es un problema de RL, donde un participante (el agente) toma decisiones en un entorno, basado en su experiencia y la recompensa de sus acciones, es decir, el entorno de este problema es el IGT. El primer modelo creado específicamente para modelar la toma de decisiones en el IGT es el *Aprendizaje por valencia de la expectativa,* más conocido como EVL (*Expectancy-Valence Learning*) [30], más adelante para mejorar el desempeño del modelo EVL se diseñó el *Aprendizaje por valencia prospectiva*, más conocido como PVL (*Prospect-Valence Learning*) [29][32].

[8] Las diferencias de toma de decisiones según la edad se discutirán en detalle en la sección 5

Todas las variaciones del modelo EVL se basan en supuestos principales:

1. Los participantes utilizan una *función de utilidad* para evaluar lo positivo o negativo de cada elección que hacen en la tarea psicológica.

2. Los participantes actualizan la expectativa de cada opción posible basado en la función de utilidad usando una *regla de aprendizaje por refuerzo*

3. Los participantes utilizan una *función probabilística de elección* para escoger que acción tomar en cada intento.

Existen diferentes funciones de utilidad, reglas de aprendizaje y funciones de elección, las combinaciones de estas funciones dan lugar a distintos modelos de toma de decisiones, las que se usaron en esta investigación se enuncian en la Metodología (sección 3).

2.2.3.1 Funciones de utilidad

1. Modelo EVL

Este modelo asume que la función de utilidad $u(t)$, es la diferencia ponderada de la ganancia y pérdida actuales en el intento t :

$$u(t) = (1 - W) \times ganancia(t) - W \times |perdida(t)| \quad (9)$$

Donde, $W \in (0,1)$ es un parámetro que representa la atención a las pérdidas, denota el peso o importancia que los participantes asignan a las pérdidas. Un valor bajo de W indica búsqueda de recompensa; incluso aunque esto signifique algunas pérdidas, por otro lado, un valor alto de W indica aversión a la pérdida [29]. El uso de esta función asume implícitamente que los participantes procesan las pérdidas y las ganancias por separado, en base a una función de utilidad definida a trozos y lineal [29].

Esta función de utilidad es llamada *función de utilidad de expectativa*, debido a que es la función de utilidad utilizada en el modelo de EVL.

2. Modelo PVL

La función de utilidad de expectativa asume que la utilidad subjetiva es linealmente proporcional a la ganancia o pérdida neta de cada intento, sin embargo, esta aproximación no toma en cuenta el efecto de la frecuencia de ganancia-pérdida. Por ejemplo, obtener una penitencia pequeña repetidas ocasiones pude ser peor que obtener una penitencia mayor únicamente una vez, incluso cuando la suma de las penitencias es equivalente entre ambas situaciones [31], es decir, que la función de utilidad de expectativa predice que ambas situaciones tienen la misma utilidad general. Por esta razón se propuso la *función de utilidad prospectiva*, que utiliza una función de utilidad no lineal:

$$u(t) = \begin{cases} x(t)^{\alpha} & si\ x(t) \geq 0 \\ -\lambda|x(t)|^{\alpha} & si\ x(t) < 0 \end{cases} \quad (10)$$

Donde, $x(t) = ganancia(t) - |perdida(t)|$ es el resultado neto en el intento t, $\alpha \in (0,2)$ y $\lambda \in (0,1)$ son parámetros libres, a diferencia de la función de utilidad de expectativa, esta función asume que los participantes procesan únicamente el resultado neto. El parámetro α gobierna la forma de la función de utilidad (Por ejemplo, la curva de la función es cóncava para resultados positivos y convexa para negativos), y el parámetro λ representa la sensibilidad del participante a las pérdidas [34].

2.2.3.2 Reglas de aprendizaje

1. Regla de aprendizaje delta

También conocida como la regla de Rescorla-Wagner [35], es ampliamente utilizada en aplicaciones de RL [11] y también en la prueba IGT [30] [36].

$$E_j(t) = E_j(t-1) + A\delta_j(t)\left[u(t) - E_j(t-1)\right] \quad (11)$$

Donde, $E_j(t)$ es la expectativa de la opción j en la ronda t. El parámetro $A \in (0,1)$, es el parámetro de actualización e indica cuánto de la expectativa de la opción seleccionada j, en la ronda t es modificada por la predicción del error, $\left[u(t) - E_j(t-1)\right]$. Se puede ver cómo las expectativas de las opciones no seleccionadas en la ronda t, permanecen iguales que en la ronda anterior, finalmente, $\delta_j(t)$ es una función indicadora que vale 1 si la opción j es escogida en la ronda t y 0 de lo contrario.

La expectativa hace referencia a la predicción del valor que tendrán las diferentes decisiones, es utilizada para aprender a tomar decisiones con expectativas más altas. Es interesante ver las semejanzas de la ecuación (11) con la ecuación (8), es decir, el aprendizaje por diferencia temporal y la regla de aprendizaje delta.

2. Regla de refuerzo por decaimiento

En el caso del refuerzo por decaimiento [37], la expectativa de la opción j en la ronda t es:

$$E_j(t) = kE_j(t-1) + \delta_j(t)\,u(t) \quad (12)$$

Esta regla asume que la expectativa pasada siempre se descuenta y la opción escogida en la ronda actual es actualizada por $u(t)$, k $(0 < k < 1)$ es un parámetro de decaimiento y describe la cantidad de descuento de la expectativa [8]. Esta regla se caracteriza por tener más flexibilidad que la regla delta, ya que, permite el cambio de las expectativas de todas las opciones en cada ronda. Por otro lado, se ha reportado que la flexibilidad del modelo puede resultar en un sobreajuste de los datos, lo que afectaría la capacidad de generalización del modelo [29].

2.2.3.3 Funciones de elección

En primera instancia se podría pensar que se debería escoger la opción que tenga mayor valor de expectativa, pero este tipo de decisiones no permitirían la exploración de las otras opciones, por esto, en las funciones de elección se hace importante utilizar una estrategia que tome en cuenta el equilibrio entre exploración y explotación. Una regla utilizada para la prueba IGT se denomina la función exponencial normalizada (más conocida como función *softmax)* o función de exploración de Boltzmann [38]. En esta regla se asume que la probabilidad de seleccionar una opción es proporcional a la expectativa de la alternativa [29].

Si $D(t+1)$ es la opción escogida en la siguiente ronda, entonces la probabilidad de que la opción j sea escogida en la siguiente ronda ($Pr[D(t+1)=j]$) está determinada por:

$$Pr[D(t+1)=j] = \frac{e^{\theta(t)E_j(t)}}{\sum_{k=1}^{4} e^{\theta(t)E_k(t)}} \tag{13}$$

El parámetro $\theta(t)$ (también llamado función de temperatura), determina la sensibilidad a las expectativas que tiene la función probabilidad de elección (por ejemplo, cuando $\theta(t)$ se aproxima a 0, las elecciones se vuelven completamente aleatorias; ayudando a la exploración). El parámetro de sensibilidad puede ser determinado de dos formas:

1. Independiente de la ronda

$$\theta(t) = 3^c - 1 \tag{14}$$

Donde, $c \in (0,3)$ es un parámetro de consistencia. Un valor grande de c indica una elección determinística y un valor pequeño sugiere que la elección es aleatoria [8][29].

2. Dependiente de la ronda

$$\theta(t) = \left(\frac{t}{10}\right)^c \tag{15}$$

La diferencia con esta regla es que depende de la ronda t y en este caso $c \in (0,3)$. Nótese que antes de la ronda 10 el parámetro de sensibilidad favorece la exploración, pero después de la ronda 10, la explotación predomina.

Es importante notar que los parámetros en los modelos cognitivos de toma de decisiones representan características psicológicas de los participantes, por esta razón, se tiene que ajustar un modelo a cada participante.

2.3 Enfoque cuántico a la toma de decisiones

En esta sección se busca responder la siguiente pregunta:

¿Por qué utilizar la teoría cuántica para estudiar la cognición y toma de decisiones?

Los argumentos que se darán en esta sección van a complementar la discusión del desempeño del algoritmo de QRL que se implementó para predecir y modelar la toma de decisiones.

Lo que es importante destacar de la teoría cuántica para esta sección principalmente, es que tiene un conjunto de axiomas que permite asignar probabilidades a eventos, las consecuencias de estos axiomas resultan en dos aspectos principales que permiten estudiar la toma de decisiones desde nuevas perspectivas:

1. Un primer juicio o decisión interfiere en los juicios o decisiones subsecuentes, produciendo un efecto de orden, es decir, las decisiones son *no conmutativas*. Se obtiene un resultado diferente si se cambia el orden de los juicios o de las decisiones, en contraste con la teoría de toma de decisiones clásica [6].

2. El entrelazamiento cuántico permite que una observación realizada en una parte de un sistema afecte instantaneamente el estado en otra parte del sistema, incluso si los sistemas están separados por distancias espaciales [6]. Esta propiedad permite modelar los fenómenos de toma de decisiones en modos no-reduccionistas, pues una señal producida en un área del cerebro puede afectar otras áreas del cerebro que están espacialmente distanciadas (los sistemas entrelazados no pueden ser descompuestos en subsistemas) [6].

Una de las ideas principales es que la toma de decisiones es un sistema no descomponible, esto desencadena que estos modelos se puedan describir con otras propiedades cuánticas como el *paralelismo cuántico*; la cual permite evolucionar varias trayectorias en paralelo [6]. Antes de enunciar los argumentos para el uso de la teoría cuántica, es importante recordar que la teoría de probabilidad clásica puede emerger como un caso restrictivo de la teoría de probabilidad cuántica [6].

2.3.1 Argumentos para un enfoque cuántico a la toma de decisiones

2.3.1.1 Los juicios están basados en estados indefinidos

En los modelos teóricos y computacionales de toma de decisiones, el sistema cognitivo cambia de momento a momento, pero en cada momento específico se encuentra en un estado definido con respecto a un juicio o decisión tomada [6].

Por ejemplo, un jurado que se encuentra ponderando la evidencia en un juicio, tiene que determinar un veredicto de culpabilidad o no. Si se supone que la probabilidad de culpa es $p \in [0,1]$, entonces los modelos clásicos de cognición y toma de decisiones asumen que en cada momento el jurado se encuentra en un estado definido con respecto a la culpabilidad, es decir, $p > 0.5$ o $p \leq 0.5$ [6].

Cuando se realiza una simulación por computador, las diferentes trayectorias que puede tomar un proceso son por el cambio de las condiciones iniciales, pero realmente se tiene una trayectoria definida en un espacio de estados [6]. En contraste a esto, la teoría cuántica permite estar en un estado indefinido (*estado de superposición*) en cada momento antes de

una decisión, es decir, en el juicio todas las opciones posibles (culpable o inocente) tienen *potencial* de ser expresadas en cada momento (esto **no** significa que se encuentre en dos estados definidos al mismo tiempo). Este nuevo enfoque permite explicar la confusión, ambigüedad, o incertidumbre al tomar una decisión [6].

Finalmente, la teoría cuántica permite modelar la toma de decisiones como si fuera una "onda" evolucionando en el tiempo moviéndose por un espacio de estados definidos. La incertidumbre o confusión se resuelve cuando se toma la decisión (colapso de la función de onda) [6].

Se puede pensar la dualidad onda-partícula en este contexto como si el comportamiento ondulatorio representara el conflicto o ambigüedad al tomar una decisión y el comportamiento corpuscular representara la resolución y certeza de una decisión tomada [6].

2.3.1.2 Las mediciones influyen en la toma de decisiones

La función de onda colapsa cuando es medida, es decir, la acción de medir u observar el sistema influye directamente en el resultado. Los modelos clásicos de toma de decisiones y cognición asumen que preguntarle a una persona sobre un sentimiento o un estado actual no cambia el estado de la persona. En los modelos cuánticos, antes de la pregunta la persona se encuentra en un estado indefinido, en el que todas las opciones de respuesta tienen probabilidad de ser expresadas [6].

En otras palabras, la respuesta o decisión es construida por la interacción entre los estados indefinidos y la pregunta (medición). Esto es de hecho la base de teorías psicológicas modernas de la emoción [39] [6]. También, se ha argumentado que las creencias y preferencias son construidas en el camino, más que únicamente ser extraídas de la memoria [40] [6].

2.3.1.3 Los juicios se perturban entre ellos

El colapso de un estado debido a una medición sugiere que los sistemas responden diferente ante mediciones posteriores, en el caso de la toma de decisiones significa que el orden de las preguntas se vuelve importante. La primera pregunta introduce incertidumbre en la segunda pregunta, lo que puede cambiar el resultado final [6]. Para los modelos clásicos de toma de decisiones, el orden de las mediciones no altera el resultado (conmutan).

Los estados definidos se llaman formalmente *eigenvectores* y una superposición de los eigenvectores se denomina *eigenestado*. El conjunto de eigenvectores forman una base en un espacio vectorial y cualquier eigenestado es un punto en este espacio vectorial [41].

El campo de los modelos cuánticos de cognición y decisión es amplio y actualmente hay estudios de modelos cuánticos de procesamiento de información en el cerebro [42], modelos del funcionamiento del cerebro desde la decoherencia; la cual es una propiedad cuántica que explica como un sistema entrelazado puede dar lugar a un sistema clásico (no

entrelazado), en este estudio la decoherencia es inducida por la interacción entre la memoria y el entorno mental externo [43].

2.4 Aprendizaje por Refuerzo Cuántico

2.4.1 Introducción a la computación cuántica

En computación cuántica la unidad de información se denomina *qubit* (Quantum Bit), es el análogo al bit clásico y se representa como un estado cuántico arbitrario en estado de superposición. Los eigenvectores son los estados clásicos (1 y 0). Los qubits se representan utilizando la notación Bra-Ket de Dirac [45]:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \tag{16}$$

Donde α, β son coeficientes complejos. Si alguno de estos coeficientes tiene un valor de cero, entonces el qubit deja de estar en estado de superposición, por otro lado, si se realiza una medición, el estado colapsa en alguno de los eigenvectores [50]. La probabilidad de ocurrencia de cada eigenvector está dada por: $|\alpha|^2$ y $|\beta|^2$. La principal diferencia con un computador clásico es que los qubits pueden simultáneamente almacenar información de los eigenvectores [4] [50]. Debido a que los estados clásicos forman una base ortonormal, cualquier qubit se puede representar como una combinación lineal de los estados clásicos.

Una operación esencial en la computación cuántica es la *transformación unitaria* U. Si se aplica una transformación a un estado en superposición, entonces, todos los eigenvectores se verán afectados y el resultado sería un nuevo estado cuántico en superposición. La propiedad que tiene la computación cuántica de evaluar diferentes valores de una operación simultáneamente, se le llama *paralelismo cuántico* [4]. Sin embargo, no se pueden obtener los resultados de la transformación simultáneamente, debido a que el colapso del estado cuántico genera una única salida [50].

2.4.2 Representación de variables en QRL

Para representar las variables de un MDP en computación cuántica se utiliza un estado cuántico, el cual es una superposición de los eigenvectores. La cantidad de eigenvectores depende de la cantidad de acciones que se pueden tomar en el entorno, en el caso de esta investigación se pueden tomar 4 acciones diferentes [4].

Un estado arbitrario $|S\rangle$ (o acción $|A\rangle$), en un algoritmo de QRL se puede expandir en un conjunto de eigenestados ortogonales $|s_n\rangle$ (o eigenacciones $|a_n\rangle$), de la siguiente manera:

$$|S\rangle = \sum_n \alpha_n |s_n\rangle \tag{17}$$

$$|A\rangle = \sum_n \beta_n |a_n\rangle \tag{18}$$

En los algoritmos de QRL, se colapsa el estado cuántico para determinar cuál de las acciones posibles (eigenacciones) se va a tomar, por esta razón la actualización de las probabilidades de ocurrencia de cada eigenacción es muy importante en el proceso de aprendizaje de este algoritmo. La actualización de estas variables está basada en el algoritmo de Grover [51] (el algoritmo de Grover se expone en la sección 3.3).

Las propiedades de la computación cuántica permiten dar otro enfoque a los retos principales del RL, en primer lugar, la computación cuántica posibilita el aumento de la velocidad de aprendizaje, lo cual abre la posibilidad de explorar entornos más grandes y complejos, en segundo lugar, el mecanismo de actualización y colapso del estado cuántico para escoger una acción en el entorno ha demostrado aportar al equilibrio entre exploración y explotación. En esta investigación se simula un algoritmo de QRL en un computador clásico, con un entorno generado a partir de los datos obtenidos de la prueba IGT.

Capítulo 3

METODOLOGÍA

En este capítulo se enuncia el detalle de los métodos y materiales utilizados para cumplir los objetivos de la investigación. Está organizado en el orden que fueron realizadas las tareas y finaliza con la implementación de los algoritmos de aprendizaje por refuerzo clásico (CRL, por sus siglas en inglés) y el de QRL para un entorno generado por la base de datos recolectada de participantes realizando la prueba psicológica Iowa Gambling Task (IGT).

Toda la metodología está basada en tres tareas principales: Protocolo de adquisición de datos, Implementación de modelos de CRL y QRL en un entorno tipo laberinto y finalmente la implementación de cuatro modelos de CRL y uno de QRL para modelar la toma de decisiones de cada participante. El esquema de los procesos y orden de la metodología se muestra en la Figura 3.1.

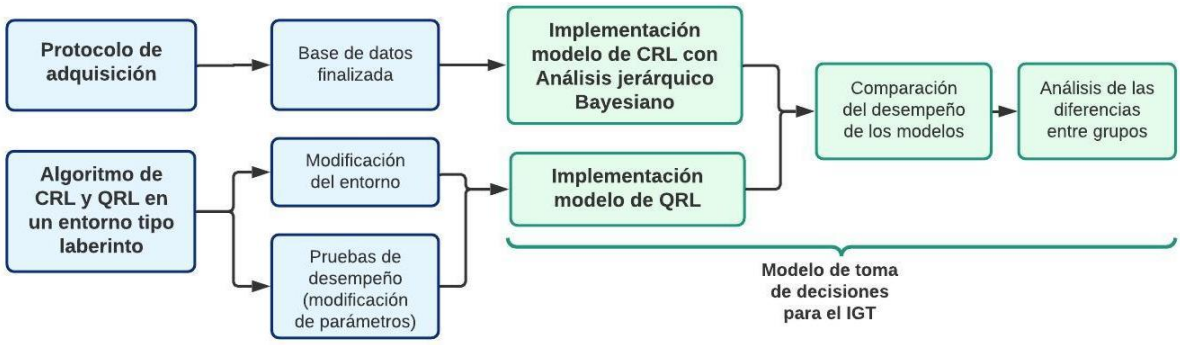


Figura 3.1: Esquema general de la metodología del proyecto

Adicionalmente, la distribución de las tareas en el tiempo de realización del proyecto (Diagrama de Gantt) se encuentra en la sección de anexos.

3.1 Protocolo de adquisición

3.1.1 Participantes

La base de datos recolectada para la realización de esta investigación cuenta con datos de 33 participantes, divididos por su edad en dos grupos: Adolescencia tardía (18 a 25 años) y Adultos jóvenes (26 a 40 años). La distribución de los participantes en los dos grupos se muestra en la siguiente tabla:

	Número de participantes	Promedio de edades (años)	Desviación estándar	Sexo	
				F	M
Adolescencia tardía	20	21.15	1.061	13	7
Adultos jóvenes	13	29.53	3.733	4	9

Tabla 3.1: Información de la población

Debido a que el rango de edades del grupo de adultos jóvenes es mayor, la desviación estándar también es mayor.

3.1.2 Reclutamiento de participantes

Los participantes inicialmente fueron invitados a participar en el proyecto con una pieza divulgativa (sección de anexos), los siguientes pasos se enumeran a continuación:

1. Encuesta en página del semillero PROMISE
2. Firma del consentimiento informado
3. Envío del enlace para realizar la actividad (IGT)

La primera encuesta se realiza con el fin de contactar a los participantes interesados en participar, que pertenezcan a alguno de los dos grupos de edades.

3.1.3 Implementación de la prueba Iowa Gambling Task (IGT)

Para realizar la prueba de forma remota, se utilizó una plataforma llamada PsyToolkit [47] [48], esta permite realizar experimentos psicológicos y cognitivos, como también encuestas.

La prueba IGT consiste en 100 rondas, en las que el participante deberá tomar una de las opciones disponibles (A, B, C o D), el experimento se diseñó inicialmente con una pantalla de bienvenida, seguido por una encuesta que permite obtener la información relevante de cada participante.

La pantalla de bienvenida se muestra en la Figura 3.3

Figura 3.2: Pantalla de bienvenida del experimento

Con el fin de no predisponer a los participantes al indiciar que es una prueba que analiza la toma de decisiones, durante todo el protocolo de adquisición de datos se utilizó como título del proyecto: Aprendizaje por refuerzo cuántico.

La secuencia de preguntas de la encuesta es la siguiente:

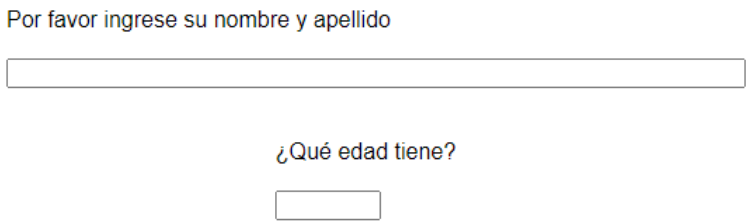


Figura 3.3: Secuencia de preguntas de la encuesta

Posteriormente la página se pone automáticamente en modo de pantalla completa, con el fin de evadir distracciones y empieza el experimento. Toda la información requerida para realizar la prueba se encuentra dentro del experimento, adicionalmente, se le indicó a los participantes realizar la prueba concentrados, sin pausas y no comentar a otros participantes como funciona la prueba.

Las diferentes pantallas disponibles en la prueba se muestran a continuación:

Figura 3.4: Secuencia de pantallas de la prueba IGT

La prueba consta de cuatro opciones (A, B, C y D), los participantes tienen que hacer un total de 100 elecciones de estas opciones. Cada ronda se tiene que escoger una opción y cuando lo hacen se les muestra la realimentación sobre si ganaron o perdieron dinero. Todas las cartas tienen un 50% de probabilidad de tener que pagar una penalidad, para las opciones A y B, la penalidad es de $250, mientras que para las opciones C y D la penalidad es de $50, además, la ganancia en las opciones A y B es de $100 y la ganancia en las opciones C y D es de $50. Por estas razones las opciones A y B se consideran desventajosas y de alto riesgo, por otro lado, las opciones C y D son consideradas ventajosas [20]. Los participantes no tienen conocimiento previo de las reglas del juego.

Para cada participante se obtuvieron los siguientes datos:

	Datos obtenidos
1	Nombre del participante.
2	Edad.
3	Sexo.
4	Tiempo total de ejecución de la prueba.
5	Opción escogida en cada ronda.
6	Ganancia por ronda.
7	Pérdida por ronda.
8	Monto total de dinero por ronda.
9	Tiempo de reacción (click del mouse).

Tabla 3.2: Datos obtenidos en el IGT

Todos los datos son analizados con Python, se cargan los archivos de las encuestas y de los datos de la prueba y se dividen en dos grupos, dependiendo la edad del participante, esto con el fin de analizar los resultados en ambos grupos.

3.2 CRL y QRL en un entorno tipo laberinto

La implementación de los algoritmos de CRL y QRL se realizó en Python, más específicamente, para implementar el algoritmo de CRL se utilizó la librería GYM y para implementar el algoritmo de QRL se utilizó la librería Qiskit.

El entorno tipo laberinto se utilizó inicialmente para probar el algoritmo de aprendizaje por refuerzo cuántico. Debido a que el proceso de entrenamiento en este entorno es más rápido que en el entorno de toma de decisiones, es ideal para hacer pruebas de las constantes del algoritmo, por ejemplo, el factor de aprendizaje y el factor de descuento (ecuación 8). Además, la cantidad de acciones que se pueden tomar en el entorno tipo laberinto es igual a la del entorno de toma de decisiones (cuatro), por lo que en ambos entornos se utilizan 2 qubits, esto permite una primera comparación del desempeño de ambos algoritmos.

Por las similitudes mencionadas resulta conveniente primero probar el algoritmo en este entorno y posteriormente realizar modificaciones pequeñas para implementar el QRL con los resultados del IGT.

El entorno tipo laberinto que se utilizó es una malla de 10x10, que tiene obstáculos y un estado inicial y uno final, la tarea del agente es llegar al estado final obteniendo la máxima recompensa. El entorno es el siguiente:

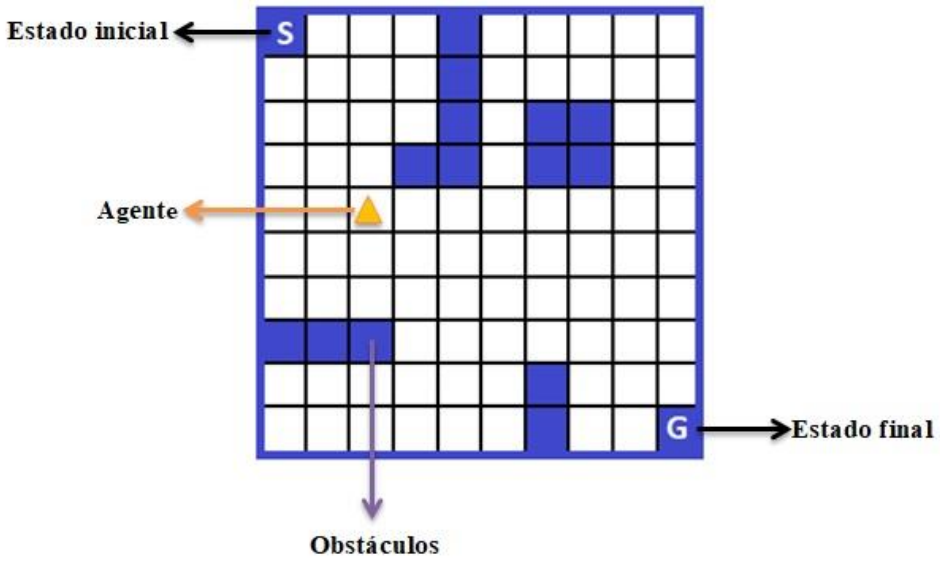


Figura 3.5: Entorno tipo laberinto

El agente puede tomar 4 acciones diferentes (arriba, abajo, izquierda y derecha), si la acción conlleva a la posición de un obstáculo, entonces el agente no realiza la acción. Cada vez que se colapsa el estado cuántico, resulta una de estas 4 acciones. Las recompensas se representan con una matriz:

$$r = \begin{bmatrix} -1 & -1 & -1 \\ -1 & -1 & -1 \\ -1 & -1 & 10 \end{bmatrix} \tag{19}$$

Las dimensiones de esta matriz son las mismas que las del laberinto, cada posición del laberinto que no sea el estado final se penaliza con una recompensa de -1, esto contribuye a buscar acciones que generen más recompensa, el estado final tiene una recompensa de 10. La función de valor $v(s)$ se inicializa en 0 y los estados cuánticos se inicializan de tal forma que todas las acciones tienen igual probabilidad de ocurrencia.

El algoritmo QRL se muestra en la Figura 3.7, nótese que se utiliza el algoritmo de grover para actualizar los estados y que colapse en la acción que genera más recompensa.

Procedimiento QRL:

Inicializar $|s^m\rangle = \sum_{s=00\ldots0}^{11\ldots1} C_s|s\rangle$, $f(s) = |a_s^{(n)}\rangle = \sum_{a=00\ldots0}^{11\ldots1} C_a|a\rangle$ y V(s) arbitrariamente.

Repetir (Por cada episodio)

Para todos los estados $|s\rangle$ en $|s^m\rangle = \sum_{s=00\ldots0}^{11\ldots1} C_s|s\rangle$:

1. Observar $f(s) = |a_s^{(n)}\rangle$ y obtener $|a\rangle$;
2. Tomar acción $|a\rangle$, observar el siguiente estado $|s'\rangle$. Recompensar r, entonces:
   a) Actualice el valor del estado: $V(s) \leftarrow V(s) + \alpha(r + \gamma V(s') - V(s))$
   b) Actualice las amplitudes de probabilidad:
   Repita U_{GROV} para L veces.

$$U_{GROV}|a_s^{(n)}\rangle = U_{a_0^{(n)}} U_a = |a_s^{(n)}\rangle$$

Hasta que para todos los estados $|\Delta V(s)| \leq \varepsilon$.

Figura 3.6: Algoritmo QRL

3.3 CRL y QRL en entorno de toma de decisiones

3.3.1 Modelos de CRL

Se utilizaron cuatro modelos de toma de decisiones y se estimaron los parámetros libres de cada modelo utilizando Análisis Bayesiano Jerárquico (HBA, por sus siglas en inglés), este método ofrece beneficios sobre los métodos convencionales [34]. El HBA utiliza una clase de aprendizaje por refuerzo llamado Monte Carlo Hamiltoniano (HMC, por sus siglas en inglés), el cual es una variante del método de cadenas de Markov Monte Carlo.

Los modelos de toma de decisiones utilizados, con el número de parámetros que contienen se muestran en la tabla 3.3.

Modelo	Número de parámetros
Aprendizaje por representación de resultado (ORL)	5
PVL Decay-RI	4
PVL Delta	4
Value-Plus-Perseverance (VPP)	7

Tabla 3.3: Modelos de toma de decisiones

Todos los modelos se entrenaron para los dos grupos de diferentes edades, es decir, se obtuvieron 8 modelos entrenados.

3.3.2 Modelo de QRL

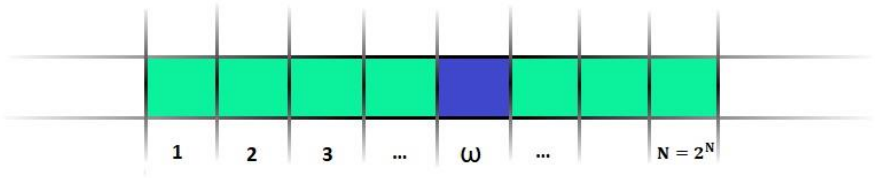


Figura 3.7: Algoritmo de Grover

Si se tiene una lista de elementos como en la Figura 3.8, y se desea encontrar el elemento ω, en computación clásica se tendría que revisar en promedio $N/2$ de los ítems, sin embargo, el algoritmo de Grover, utilizado en mecánica cuántica, permite aumentar la probabilidad de ocurrencia del coeficiente del eigenvector ω. De esta manera se aumenta la velocidad con la que se encuentra el elemento en la lista, en el caso de los algoritmos de QRL, el algoritmo de Grover es el mecanismo de actualización de los estados cuánticos [4] [8].

El algoritmo de QRL en el entorno de toma de decisiones se diseñó en base al previamente explicado algoritmo de QRL para el entorno tipo laberinto. El diagrama de flujo mostrado en la Figura 3.9 permite comparar los elementos de ambos algoritmos.

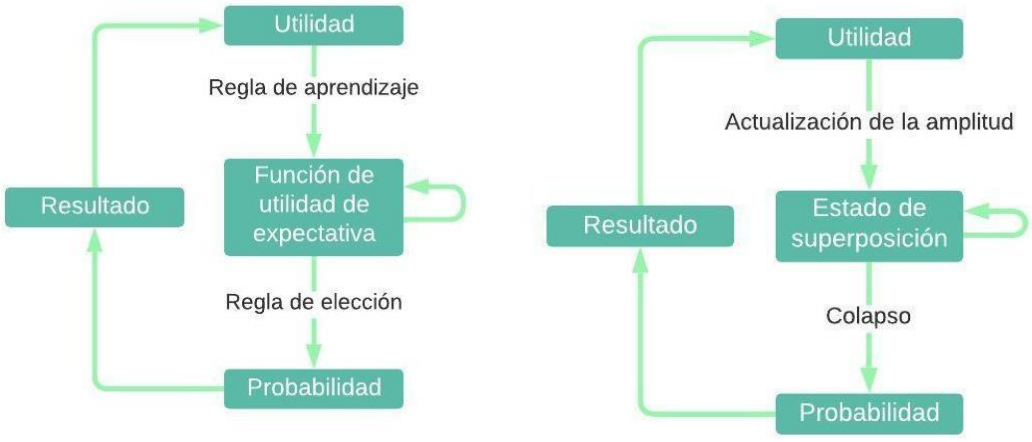


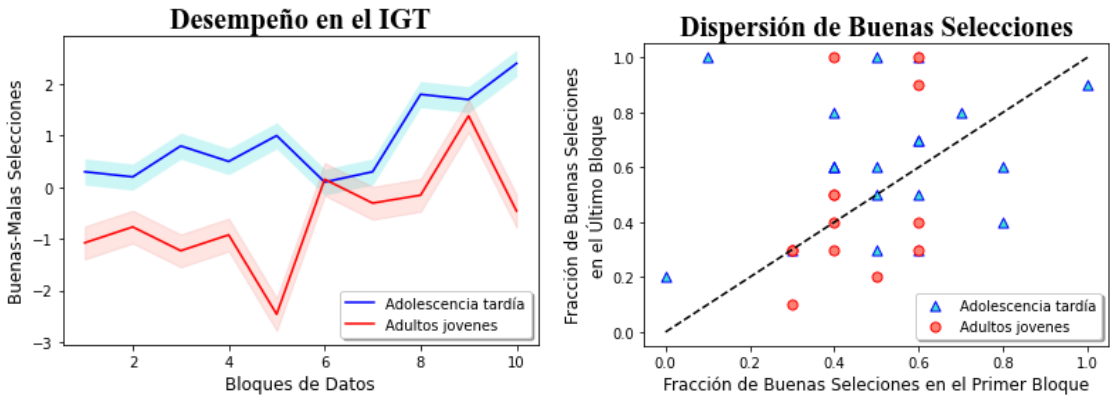
Figura 3.8: Comparación CRL (izquierda) y QRL (derecha)

Capítulo 4

RESULTADOS

4.1 Resultados del IGT

Para evaluar el desempeño de ambos grupos en el IGT primero se comparó la proporción de Buenas-Malas decisiones en bloques de 10 rondas, de esta manera se puede evidenciar el aprendizaje de los participantes mientras avanzaba la prueba. En la Figura 4.1 A, la zona sombreada de las gráficas es el error estándar de la media y la Figura 4.1 B, es un gráfico de dispersión en el que cada ítem es un participante.



A) B)

Figura 4.1: Desempeño en el IGT

De la Figura 4.1 A, se observa que en la mayoria de los bloques el grupo de adolescencia tardía obtuvo mayor número de buenas selecciones que el grupo de adultos jovenes, también es importante resaltar que en el último bloque el grupo de adolecencia tardía obtuvo la mayor cantidad de buenas selecciones durante toda la prueba, lo que demuestra el aprendizaje durante toda la prueba.

También como se observa en la Figura 4.1 B se comparó la cantidad de decisiones ventajosas en el primer bloque con la catidad de decisones ventajosas en el ultimo bloque, en donde principalmente se observa que los participantes de ambos grupos inician sin conocimiento previo de las reglas de la prueba, ya que, la fracción de buenas selecciones está en general en el medio de la gráfica.

Asimismo, se evaluaron las fracciones de decisiones tomadas de cada opción, es la manera de seguir los cambios en las elecciones de cada una de las opciones y así ver tendencias

de decisiones en cada grupo. Las opciones ventajosas (C y D) están con un grosor de línea mayor para hacer énfasis en ellas.

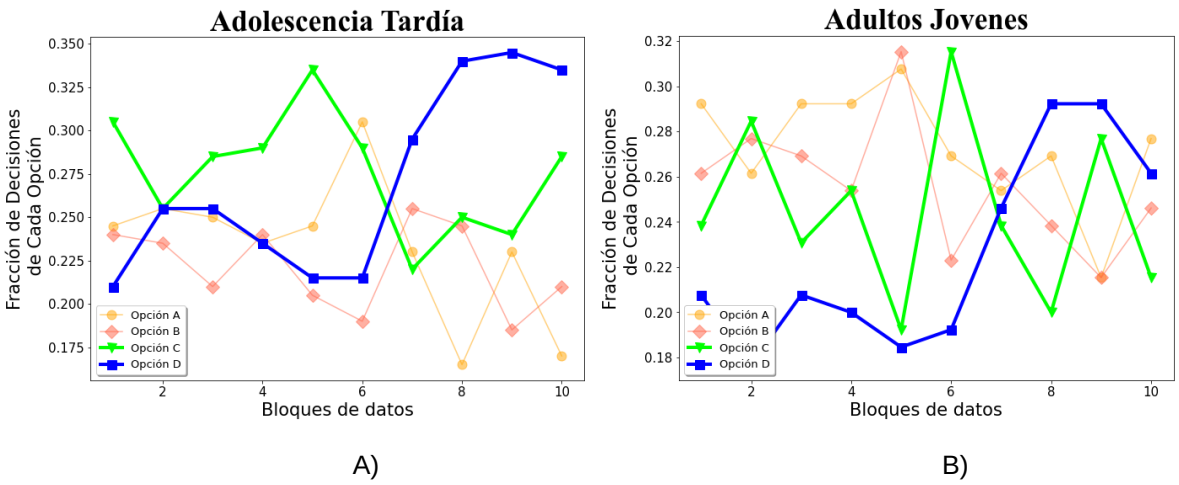


Figura 4.2: Fracción de decisiones ventajosas para ambos grupos

Finalmente en la Figura 4.3 se muestra el promedio sobre todas las rondas de cada opción. Para verificar si existe diferencia significativa de las decisiones tomadas entre ambos grupos se realizó una prueba T de muestras independientes $t = 1.528, p > 0.1$, demostrando que no existe una diferencia significativa en las decisiones de ambos grupos.

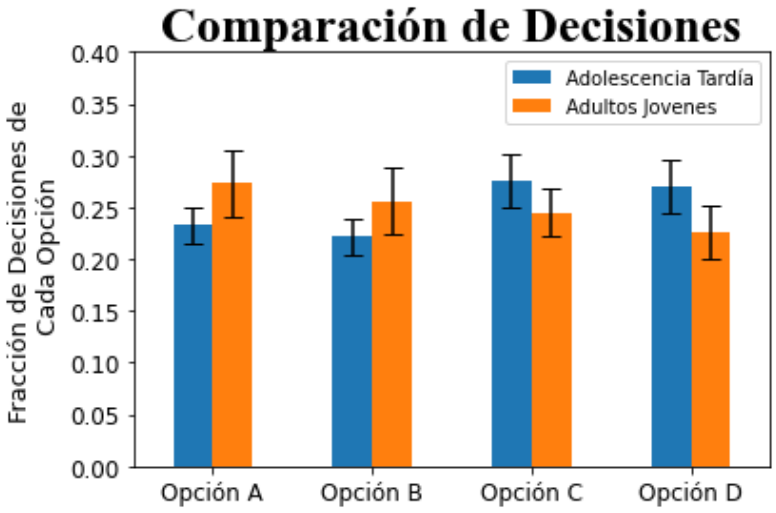


Figura 4.3: Diagrama de barras que representa el promedio de las elecciones en el IGT

4.2 Resultados en el entorno tipo laberinto

La manera usual de presentar resultados de RL es graficar la cantidad de iteraciones que se realizaron en cada episodio, debido a que, si se busca una política óptima, se debería disminuir la cantidad de iteraciones y actualizaciones de la función de valor con los episodios. Esta gráfica se puede ver a continuación:

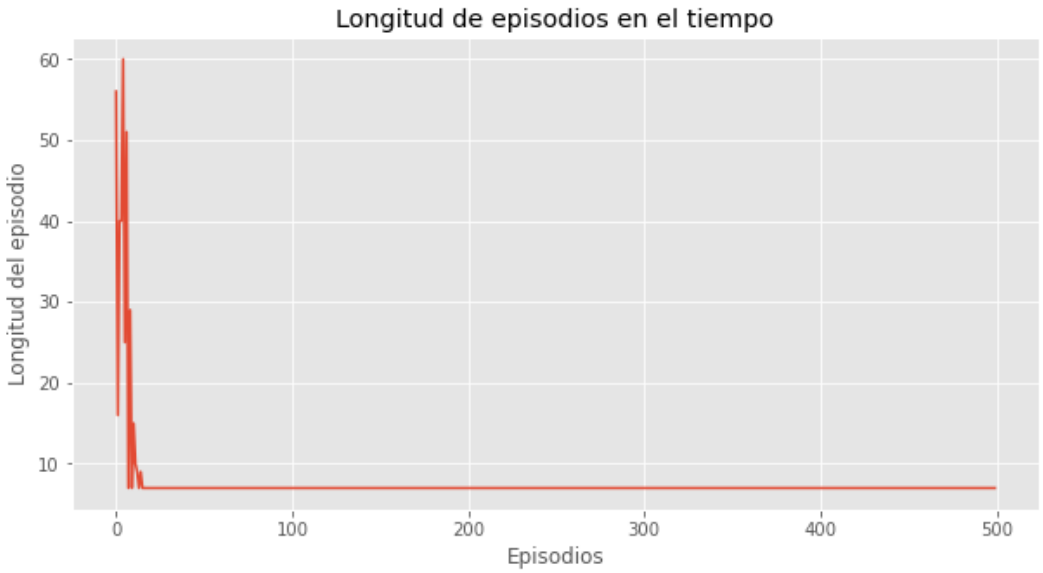


Figura 4.4: Desempeño algoritmo QRL(Laberinto)

También es de interés ver los cambios en las recompensas obtenidas en cada episodio. Si el algoritmo está aprendiendo a resolver el problema se debería ver un aumento de las recompensas, en el caso del algoritmo de QRL se observa que el agente obtiene la recompensa máxima en la misma cantidad de episodios en los que la longitud de los episodios alcanza un valor constante.

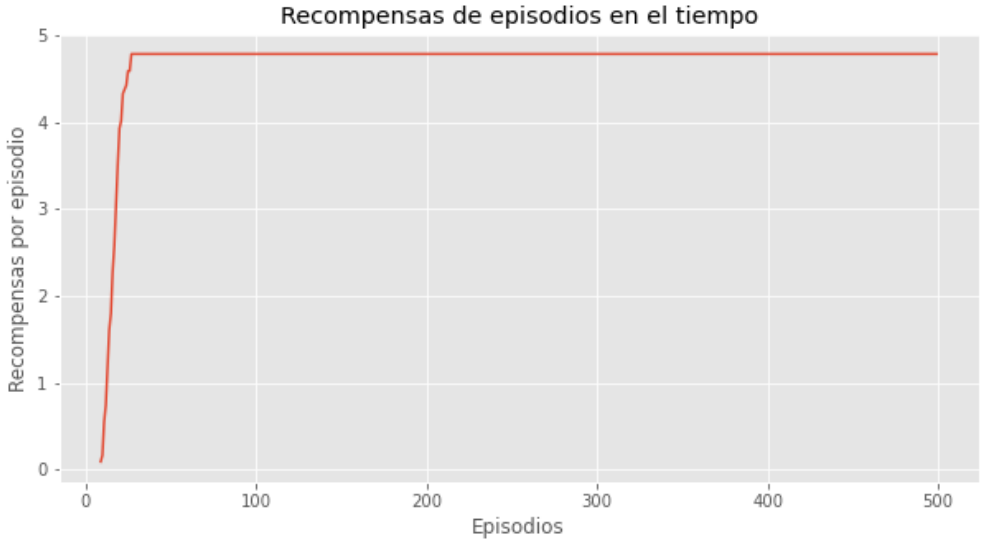


Figura 4.5: Recompensas por episodios QRL(Laberinto)

Finalmente, cuando el algoritmo obtiene una política óptima, se puede trazar un recorrido desde el inicio hasta el final del laberinto en el que obtiene recompensa máxima, y en este entorno, significa que se realiza en la menor cantidad de movimientos posibles. Después de encontrar la política óptima, las variables como la función de valor y las políticas se mantienen constantes en los episodios. En la Figura 4.6 se muestra una política óptima del algoritmo de QRL

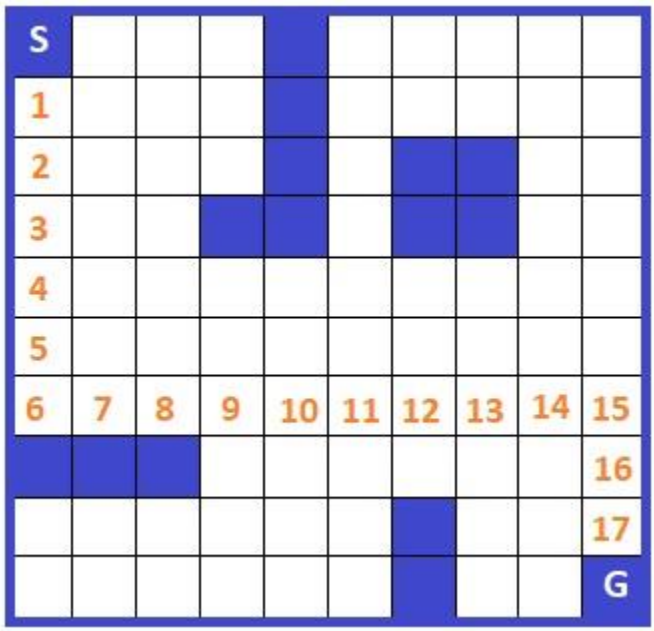


Figura 4.6: Política óptima QRL

La implementación en el entorno tipo laberinto fue realizada inicialmente para comprobar el funcionamiento del algoritmo QRL y el entendimiento de todo el proceso, además, permitió formar la base para la posterior implementación del algoritmo de QRL con la prueba IGT.

4.3 Resultados en la implementación con el IGT

Todos los modelos de CRL y el de QRL fueron entrenados con el 80% de los datos de la base de datos del IGT [6], y se comprobó su desempeño con los datos restantes. El ajuste de los parámetros de los modelos de CRL se realizó con la librería de hBayesDM, de Python [49]. Se utilizaron dos índices de desempeño, LOOIC y WAIC, los cuales son comúnmente usados en modelos de toma de decisiones [49]. Un menor valor en estos índices indica un mejor desempeño.

Debido a que los modelos se entrenaron para cada grupo de edades, se obtuvieron los índices de desempeño para los dos grupos.

Modelo	LOOIC	WAIC
ORL	-2375.39	-2361.13
VPP	-2395.03	-2379.62
PVL_decay	-2548.31	-2538.9
PVL_delta	-2575.82	-2565.7
QRL	-2359.43	-2352.34

Tabla 3.4: Desempeño de modelos en grupo de adolescencia tardía

Modelo	LOOIC	WAIC
ORL	-1608.19	-1593.64
VPP	-1620.42	-1620
PVL_decay	-1723.62	-1718.34
PVL_delta	-1734.86	-1727.1
QRL	-1598.58	-1581.93

Tabla 3.5: Desempeño de modelos en grupo de adultos jóvenes

Las tablas 3.4 y 3.5 muestran que el algoritmo de QRL obtuvo el mejor resultado para los dos índices de desempeño en los dos grupos. Esto significa que la capacidad de predicción del algoritmo de QRL es mejor que la de los otros modelos. En la Figura 4.7 se compara la predicción ronda a ronda del mejor modelo de CRL con las decisiones tomadas por un participante.

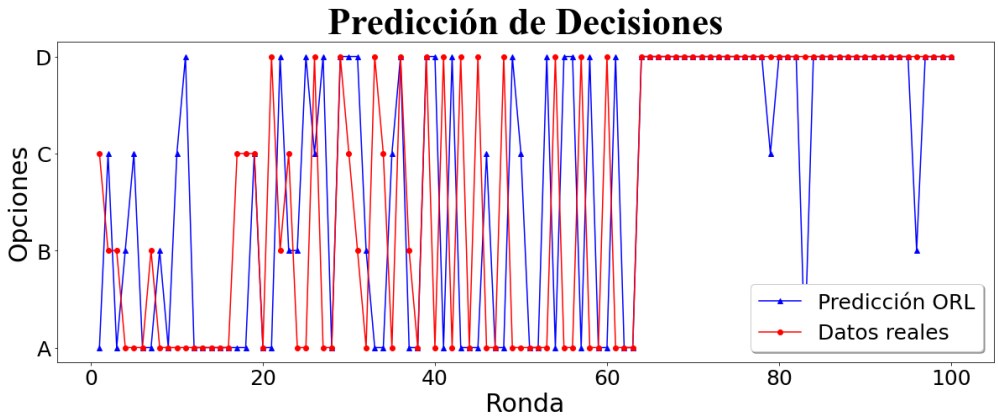


Figura 4.7: Predicción de decisiones del modelo ORL

En la Figura 4.8 se puede ver que las predicciones del modelo de QRL se ajustan mejor a las decisiones tomadas por el participante, como se esperaba por los índices de desempeño.

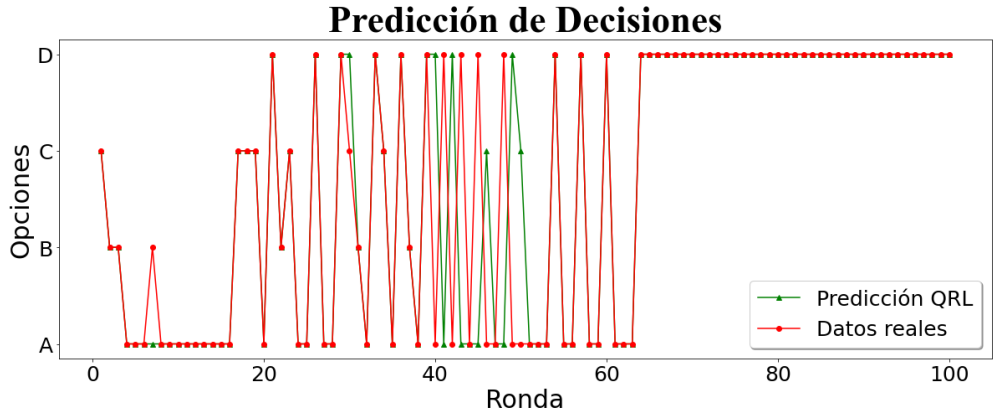


Figura 4.8: Predicción de decisiones del modelo QRL

La distribución de los parámetros ajustados para los dos grupos, con el mejor modelo de CRL se muestran en las Figuras 4.9 y 4.10.

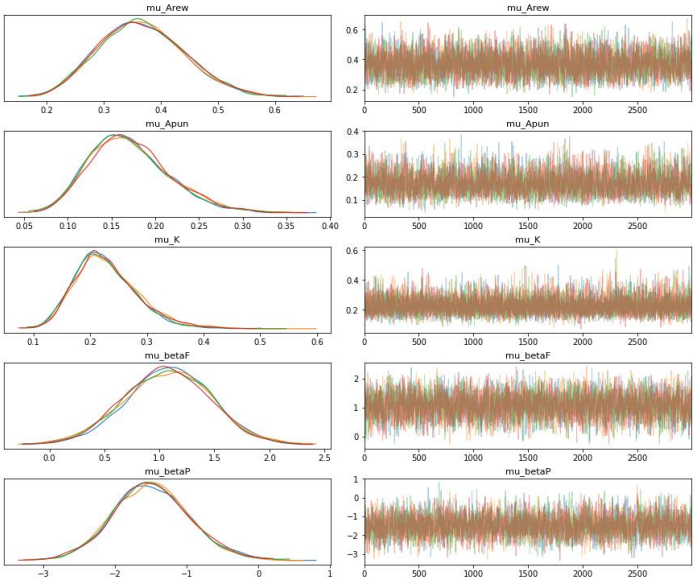


Figura 4.9: Distribución de parámetros de los adolescentes tardíos

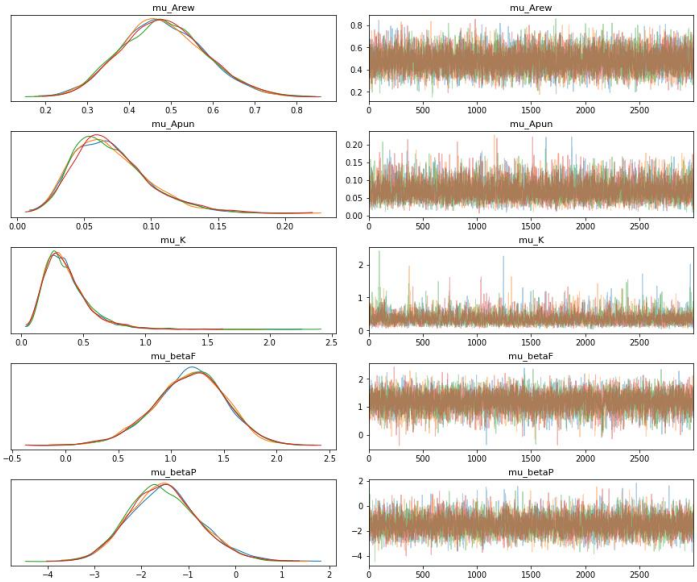


Figura 4.10: Distribución de parámetros de los adultos jóvenes

Las distribuciones de los parámetros también ayudan a verificar la convergencia del algoritmo, las curvas de diferentes colores representan las diferentes cadenas de Markov. Por otro lado, el promedio de los parámetros para los dos grupos se muestra en las Figuras 4.11 y 4.12.

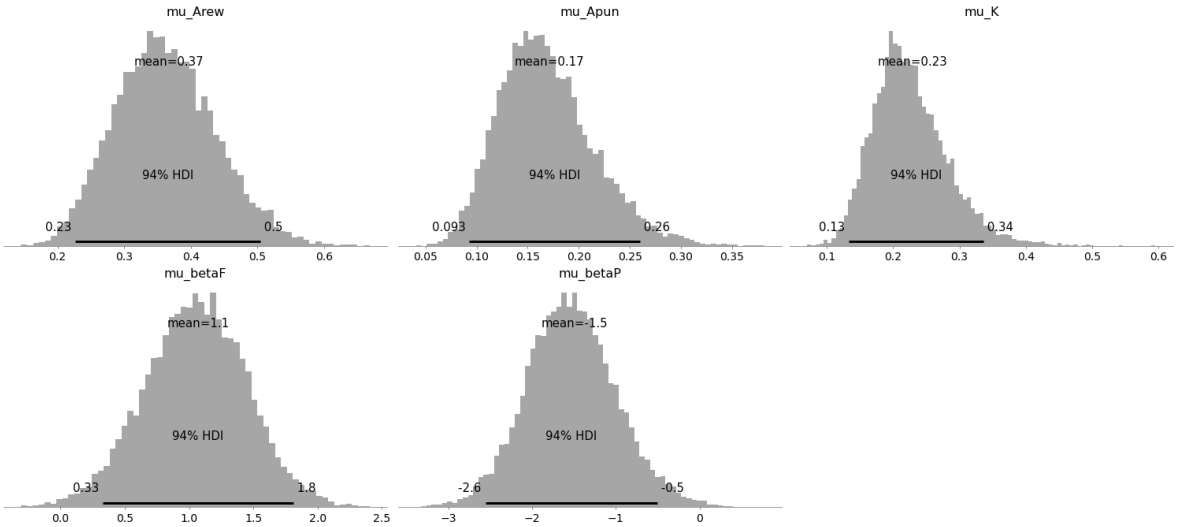


Figura 4.11: Valor promedio de los parámetros para los adolescentes tardíos

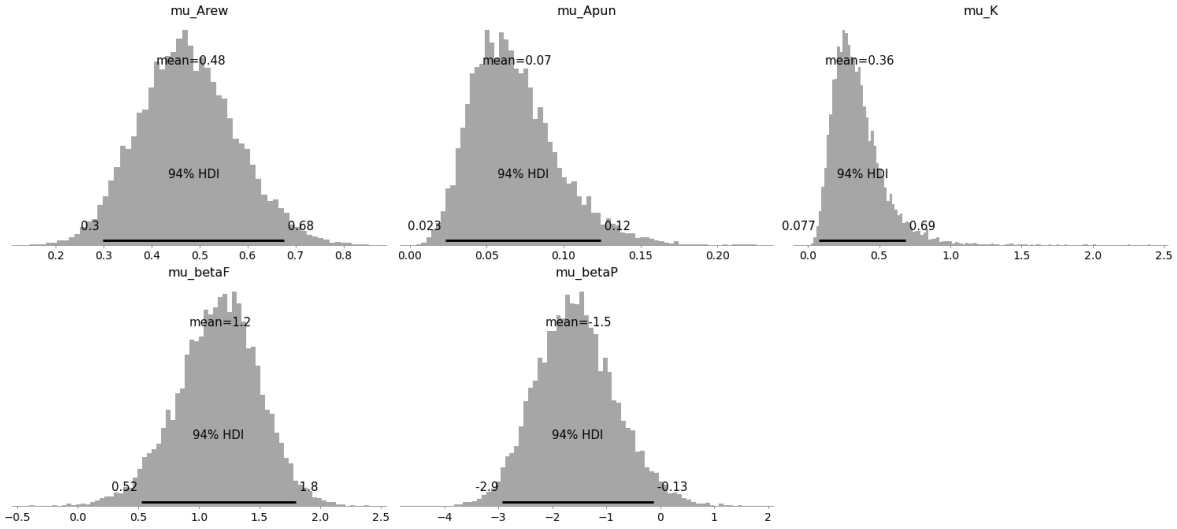


Figura 4.12: Valor promedio de los parámetros para los adultos jóvenes

Capítulo 5

DISCUSIÓN

En esta investigación se comparó el desempeño de dos tipos de algoritmos para estimar los parámetros de modelos de toma de decisiones humanas. Para esta comparación se utilizó un modelo de aprendizaje por refuerzo cuántico y los cuatro modelos de aprendizaje por refuerzo clásico que reportan mejor desempeño [34], se encontró que el modelo QRL obtuvo mejor desempeño que el mejor modelo de CRL. De igual forma se comprobó que las predicciones de las decisiones tomadas por los participantes se ajustan mejor con el modelo de QRL, estas predicciones no utilizan información de la prueba IGT, únicamente los parámetros ajustados al modelo.

El proceso de aprendizaje en la toma de decisiones basadas en el valor, se basa en ajustar los valores (o pesos) de cada acción (por ejemplo, preferencia por una opción en el IGT), en los modelos de CRL los valores se actualizan por medio de la función de valor, mientras que, en los modelos de QRL los valores de cada acción se representan como las probabilidades de ocurrencia de cada acción [8]. Al actualizar las probabilidades de ocurrencia, se está tomando preferencia por alguna acción o convirtiendo más aleatoria la decisión. A diferencia de la computación clásica, la computación cuántica permite actualizar simultáneamente todas las posibles opciones en cada ronda del IGT, también, permite a los estados estar en superposición (estados indefinidos) antes de tomar una acción. La simulación de un algoritmo cuántico realizada para esta investigación, no obtuvo los beneficios en velocidad de aprendizaje o paralelismo cuántico que se esperarían de una implementación física del algoritmo y aun así obtuvo mejor desempeño que los modelos de CRL, lo cual, favorece la idea de que los postulados de la teoría cuántica aportan herramientas útiles en el estudio de la cognición y toma de decisiones, por ejemplo, se han utilizado las propiedades de entrelazamiento cuántico y decoherencia para estudiar y modelar el procesamiento de información en el cerebro [42] [43].

El portador físico de un qubit puede ser cualquier sistema cuántico de dos estados, como una partícula con spin-$1/2$ o un fotón polarizado. Aunque la implementación física de algoritmos cuánticos esté alejada del alcance de esta investigación, su simulación resultó en un modelo de toma de decisiones que puede predecir la toma real de decisiones basadas en el valor para participantes del IGT. Debido a que los modelos de toma de decisiones y sus parámetros representan características individuales de los participantes, es de interés probar el desempeño de los modelos utilizados en esta investigación con otras pruebas de toma de decisiones.

Aprovechando la implementación de los modelos de toma de decisiones se realizó el protocolo de adquisición para dos grupos de diferentes edades. Debido al reporte de estudios anteriores que concluyen el aumento de decisiones de alto riesgo principalmente en la adolescencia (12 a 18 años) y en un periodo denominado adolescencia tardía (hasta los 25 años) [9] [10], en este estudio también se compararon los resultados obtenidos en la prueba IGT para adolescentes tardíos y adultos jóvenes. Los resultados del IGT demostraron que ambos grupos aumentan el desempeño en la actividad con el paso de las rondas (Figura 4.1 A), sin embargo, también se evidencia que los adultos jóvenes abarcaron más rondas para aprender de la actividad, mientras que la adolescencia tardía muestra un

comportamiento de aprendizaje continuo en la prueba. Aunque se observa tendencia de los adolescentes tardíos por tomar decisiones ventajosas, la prueba T no mostró significancia estadística de los datos, por lo que, no se puede concluir que ningún grupo tiene preferencia por decisiones de alto o bajo riesgo. Contrario a la hipótesis inicial de encontrar mayor cantidad de decisiones riesgosas en el grupo de adolescencia tardía. Se encontraron otros factores que pueden afectar los resultados de la prueba como el impacto de experiencias de riesgo en el pasado y fumar frecuentemente (los mismos participantes se tenían que definir como fumadores frecuentes o no) [26] [46]. Debido a que el IGT se diseñó para analizar la toma de decisiones en pacientes con daño en la corteza prefrontal (déficit de toma de decisiones), no se deberían esperar resultados similares en participantes sanos con diferentes edades, si bien, la toma de decisiones cambia con la edad, estos cambios no significan déficits en la toma de decisiones para los rangos de edades utilizados en esta investigación.

Para la simulación del algoritmo de QRL en el entorno tipo laberinto se utilizó la misma regla de actualización de diferencia temporal que para el modelo de CRL. En general esta comparación demuestra que el QRL debe ser probado en diferentes tipos de entornos y también es un enfoque prometedor para nuevas aplicaciones. Además, los resultados obtenidos concuerdan con una implementación anterior de un algoritmo de QRL en el mismo tipo de entorno [4].

Los modelos de CRL para el estudio de la toma de decisiones usualmente relacionan los parámetros de los modelos con características de los participantes [29] [49]. En este caso, la estimación de los parámetros de los modelos mostró una diferencia en el parámetro que representa la sensibilidad a la pérdida, en promedio este parámetro fue menor para la adolescencia tardía, esto sugiere que los participantes de este grupo tomaron decisiones riesgosas, pero igualmente esto ayudó a la exploración de todas las opciones posibles. Hay que recordar que es importante la exploración para poder encontrar las políticas óptimas. Otro resultado importante fue el desempeño de los algoritmos en los dos grupos, todos los modelos mostraron mejor desempeño en el grupo de adultos jóvenes, esto puede deberse a que este grupo contiene menos participantes o a que es más difícil predecir las decisiones para personas en el rango de edad de la adolescencia tardía.

Capítulo 6

RECOMENDACIONES Y TRABAJOS FUTUROS

Existen otro tipo de pruebas de toma de decisiones, no necesariamente enfocadas a pacientes con daño en la corteza prefrontal. Sería de interés probar los modelos en pruebas diferentes al IGT y verificar la similitud de las respuestas, por un lado, las respuestas del IGT de los diferentes grupos y por otro lado, verificar la consistencia de los parámetros ajustados con una prueba e intentar predecir las decisiones tomadas por el mismo participante en otra prueba [29] [34].

Se demostró el buen desempeño del modelo de QRL en la prueba IGT, pero es necesaria una prueba de toma de decisiones basadas en el valor diseñada para explorar las diferencias entre CRL y QRL. También, se recomienda analizar una base de datos con más participantes y equilibrada para los grupos que se estén analizando.

Se espera que se realicen más estudios en diferentes aplicaciones de los algoritmos de QRL, y, sobre todo, en las razones por las cuales se han encontrado mejores desempeños en algoritmos de QRL que en los de CRL, ¿puede ser que la teoría cuántica logre describir mejor los procesos de cognición y toma de decisiones que los modelos clásicos?

Capítulo 7

CONCLUSIONES

Los modelos de aprendizaje por refuerzo cuántico pueden ser utilizados en la neurociencia y son un competidor potencial de los modelos de aprendizaje por refuerzo clásico. Además, se logró simular un algoritmo de QRL y obtener un desempeño comparable al de los modelos clásicos de CRL.

Los resultados de la prueba IGT y los parámetros obtenidos de los modelos demostraron que no hay diferencias significativas en las decisiones tomadas por ambos grupos, pero la adolescencia tardía tomó mejores decisiones en las últimas rondas, y esto se puede deber a que la tendencia a tomar decisiones riesgosas moderadas ayuda a la exploración de las opciones. El cambio de los parámetros en los dos grupos significa que la edad cambió características en la toma de decisiones de las personas, como estos parámetros influyen en la forma en la que se toman las decisiones, entonces la maduración es un proceso en el que se alteran estos parámetros, por ejemplo, se aumenta la sensibilidad a la pérdida.

El desempeño de los algoritmos cuánticos implementados en esta investigación y otras investigaciones [4] [6] [8], demuestra que la teoría cuántica aporta información al estudio del proceso de aprendizaje humano, y que el entendimiento de ese proceso es responsable de avances de la inteligencia artificial.

Capítulo 8

REFERENCIAS

[1] Silver, D, et. al. Mastering the game of Go with deep neural networks and tree search Nature 529, 484–503 (2016).

[2] W. D. Smart and L. P. Kaelbling, “Effective reinforcement learning for mobile robots,” in Proc. IEEE Int. Conf. Robot. Autom., 2002, pp. 3404–3410.

[3] T. Kondo and K. Ito, “A reinforcement learning with evolutionary state recruitment strategy for autonomous mobile robots control,” Robot. Auton. Syst., vol. 46, no. 2, pp. 111–124, Feb. 2004.

[4] Daoyi Dong, Chunlin Chen, Hanxiong Li, & Tzyh-Jong Tarn. (2008). Quantum Reinforcement Learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 38(5), 1207–1220. doi:10.1109/tsmcb.2008.925743

[5] Niv, Y. Reinforcement learning in the brain. J. Math. Psychol. 53, 139–154 (2009).

[6] Busemeyer, J., & Bruza, P. (2012). *Quantum Models of Cognition and Decision*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511997716

[7] Yukalov, V. I., & Sornette, D. (2015). *Quantum probability and quantum decision-making. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences.*

[8] Li, J.-A., Dong, D., Wei, Z., Liu, Y., Pan, Y., Nori, F., & Zhang, X. (2020). *Quantum reinforcement learning during human decision-making. Nature Human Behaviour.*

[9] Duell, N., Steinberg, L., Icenogle, G., Chein, J., Chaudhary, N., Di Giunta, L., … Chang, L. (2017). Age Patterns in Risk Taking Across the World. Journal of Youth and Adolescence.

[10] Spear, L. P. (2000). *The adolescent brain and age-related behavioral manifestations. Neuroscience & Biobehavioral Reviews.*

[11] Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

[12] Kenji Doya. 2000*. Reinforcement Learning in Continuous Time and Space*. Neural Comput. 12, 1 (January 2000), 219–245*.*

[13] Sutton, R. S. (1988). *Learning to predict by the methods of temporal differences. Machine Learning, 3(1), 9–44.*

[14] Lee, D., Seo, H., & Jung, M. W. (2012). *Neural Basis of Reinforcement Learning and Decision Making. Annual Review of Neuroscience, 35(1), 287–308.*

[15] Byrne, K. A., Norris, D. D., & Worthy, D. A. (2015). Dopamine, depressive symptoms, and decision-making: the relationship between spontaneous eye blink rate and depressive symptoms predicts Iowa Gambling Task performance. Cognitive, Affective, & Behavioral Neuroscience, 16(1), 23–36.

[16] Ahn, W.-Y., Vasilev, G., Lee, S.-H., Busemeyer, J. R., Kruschke, J. K., Bechara, A., & Vassileva, J. (2014). Decision-making in stimulant and opiate addicts in protracted abstinence: evidence from computational modeling with pure users. Frontiers in Psychology, 5.

[17] Gold, J. I., & Shadlen, M. N. (2007). The Neural Basis of Decision Making. Annual Review of Neuroscience, 30(1), 535–574.

[18] Kawagoe, R., Takikawa, Y., & Hikosaka, O. (2004). Reward-Predicting Activity of Dopamine and Caudate Neurons—A Possible Mechanism of Motivational Control of Saccadic Eye Movement. Journal of Neurophysiology, 91(2), 1013–1024.

[19] Kringelbach, M. L. (2005). The human orbitofrontal cortex: linking reward to hedonic experience. Nature Reviews Neuroscience, 6(9), 691–702.

[20] Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. Cognition, 50(1-3), 7–15.

[21] Redgrave, P., Prescott, T. J., & Gurney, K. (1999). *The basal ganglia: a vertebrate solution to the selection problem? Neuroscience, 89(4), 1009–1023.*

[22] Busemeyer, J. R. & Johnson, J. G. in Handbook of Judgment and Decision Making (eds Koehler, D. & Narvey, N.) 133–154 (Blackwell Publishing Co., New York, 2004)

[23] Lee, D., Seo, H., & Jung, M. W. (2012). *Neural Basis of Reinforcement Learning and Decision Making. Annual Review of Neuroscience, 35(1), 287–308.* doi:10.1146/annurev-neuro-062111-150512

[24] Rangel, A., Camerer, C., & Montague, P. R. (2008). *A framework for studying the neurobiology of value-based decision making. Nature Reviews Neuroscience, 9(7), 545–556.* doi:10.1038/nrn2357

[25] Worthy, D. A., Pang, B., & Byrne, K. A. (2013). *Decomposing the roles of perseveration and expected value representation in models of the Iowa gambling task. Frontiers in Psychology, 4.*

[26] Buelow, M. T. & Suhr, J. A. Risky decision making in smoking and nonsmoking college students: examination of Iowa Gambling Task performance by deck type selections. Appl. Neuropsychol. Child 3, 38–44 (2014).

[27] Worthy, D. A., & Maddox, W. T. (2012). *Age-Based Differences in Strategy Use in Choice Tasks. Frontiers in Neuroscience, 5.*

[28] Ahn, W. Y., Dai, J., Vassileva, J., Busemeyer, J. R., & Stout, J. C. (2016). Computational modeling for addiction medicine. Neuroscience for Addiction Medicine: From Prevention to Rehabilitation - Methods and Interventions, 53–65.

[29] Ahn, W.-Y., Busemeyer, J., Wagenmakers, E.-J., & Stout, J. (2008). *Comparison of Decision Learning Models Using the Generalization Criterion Method. Cognitive Science: A Multidisciplinary Journal, 32(8), 1376–1402.*

[30] Busemeyer, J. R., & Stout, J. C. (2002). *A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. Psychological Assessment, 14(3), 253–262.*

[31] Erev, I., & Barron, G. (2005). *On Adaptation, Maximization, and Reinforcement Learning Among Cognitive Strategies. Psychological Review, 112(4), 912–931.*

[32] Tversky, A., Kahneman, D. Advances in prospect theory: Cumulative representation of uncertainty. *J Risk Uncertainty* **5,** 297–323 (1992).

[33] Haines, N., Vassileva, J., & Ahn, W.-Y. (2018). *The Outcome-Representation Learning Model: A Novel Reinforcement Learning Model of the Iowa Gambling Task. Cognitive Science.*

[34] Haines, N., Vassileva, J., & Ahn, W.-Y. (2018). The Outcome-Representation Learning Model: A Novel Reinforcement Learning Model of the Iowa Gambling Task. Cognitive Science.

[35] Rescorla, R. & Wagner, Allan. (1972). A theory of Pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement. Classical Conditioning: Current Research and Theory.

[36] Yechiam, E., Busemeyer, J.R. Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychonomic Bulletin & Review* **12,** 387–402 (2005).

[37] Erev, Ido and Roth, Alvin, (1998), Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria, American Economic Review, 88, issue 4, p. 848-81.

[38] Kaelbling, L.P., Littman, M.L. and Moore, A.W. (1996) Reinforcement Learning: A Survey. Journal of Artificial Intelligence Research, 4, 237-285.

[39] Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 69*(5), 379–399.

[40] Payne, J. W., Bettman, J. R., & Johnson, E. J. (1992). Behavioral decision research: A constructive processing perspective. *Annual review of psychology*, *43*(1), 87-131.

[41] Shankar, R. (2012). *Principles of quantum mechanics*. Springer Science & Business Media.

[42] Khrennikov, A., & Asano, M. (2020). *A Quantum-Like Model of Information Processing in the Brain. Applied Sciences, 10(2), 707.*

[43] Asano, M., Ohya, M., Tanaka, Y., Basieva, I., & Khrennikov, A. (2011). *Quantum-like model of brain's functioning: Decision making from decoherence. Journal of Theoretical Biology, 281(1), 56–64.*

[44] Yukalov, V. I., & Sornette, D. (2015). *Quantum probability and quantum decision-making. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2058).*

[45] Preskill, J. (1998). Reliable quantum computers. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 454(1969), 385–410.

[46] Xue, G., Lu, Z., Levin, I. P., & Bechara, A. (2010). *The impact of prior risk experiences on subsequent risky decision-making: The role of the insula. NeuroImage, 50(2), 709–716.*

[47] Stoet, G. (2010). PsyToolkit - A software package for programming psychological experiments using Linux. *Behavior Research Methods, 42(4)*, 1096-1104.

[48] Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology, 44(1)*, 24-31.

[49] Ahn, Woo-Young; Haines, Nathaniel; Zhang, Lei (2017). Revealing Neurocomputational Mechanisms of Reinforcement Learning and Decision-Making With the hBayesDM Package. Computational Psychiatry, 1(), 24–57.

[50] Chen, C.-L., & Dong, D.-Y. (2008). Superposition-Inspired Reinforcement Learning and Quantum Reinforcement Learning. Reinforcement Learning.

[51] L. K. Grover, "Quantum mechanics helps in searching for a needle in a haystack," Phys. Rev. Lett., vol. 79, no. 2, pp. 325–327, Jul. 1997

Capítulo 9

ANEXOS

El diagrama de Gantt se muestra a continuación:

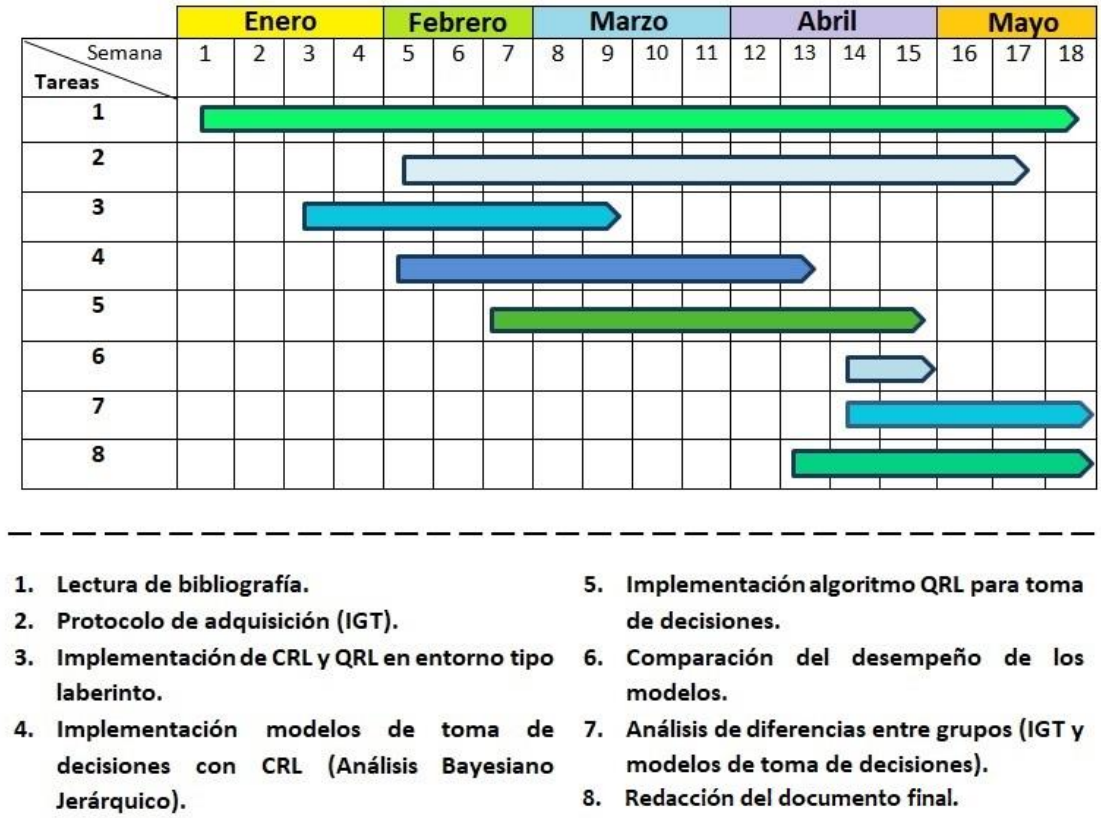


Figura 9.1: Diagrama de Gantt

La pieza de divulgación utilizada para invitar a participar de la toma de datos es la siguiente:

Figura 9.2: Pieza de divulgación