



Universidad
Central

INDEPENDENCIA - PLURALISMO - COMPROMISO

UNIVERSIDAD CENTRAL DE CHILE
FACULTAD DE INGENIERÍA
ESCUELA DE INDUSTRIAS

**APRENDIZAJE POR REFUERZO JERÁRQUICO PARA LA
JUSTIFICACIÓN EN TOMA DE DECISIÓN DE UN AGENTE
AUTÓNOMO**

Hugo Sebastián Muñoz Salvatierra

Profesor Guía:
Francisco Cruz

SANTIAGO, CHILE, 2021



Universidad
Central

INDEPENDENCIA - PLURALISMO - COMPROMISO

**UNIVERSIDAD CENTRAL DE CHILE
FACULTAD DE INGENIERÍA
ESCUELA DE INDUSTRIAS**

**APRENDIZAJE POR REFUERZO JERÁRQUICO PARA LA
JUSTIFICACIÓN EN TOMA DE DECISIÓN DE UN AGENTE
AUTÓNOMO**

Hugo Sebastián Muñoz Salvatierra

Memoria para optar al Título
de Ingeniero Civil en Computación e Informática.

Profesor Guía:
Francisco Cruz.

SANTIAGO, CHILE, 2021



Universidad
Central

INDEPENDENCIA - PLURALISMO - COMPROMISO

UNIVERSIDAD CENTRAL DE CHILE
FACULTAD DE INGENIERÍA
ESCUELA DE INDUSTRIAS

APRENDIZAJE POR REFUERZO JERÁRQUICO PARA LA JUSTIFICACIÓN EN TOMA DE DECISIÓN DE UN AGENTE AUTÓNOMO

Memoria preparada bajo la supervisión de
la comisión integrada por los profesores:

CRUZ NARANJO FRANCISCO JAVIER

CANELON OSAL RODOLFO

ROJAS PINO LUIS

Quienes recomiendan que sea aceptada
para completar las exigencias del Título
de Ingeniero Civil en Computación e Informática.

SANTIAGO, CHILE, 2021

DEDICATORIA

AGRADECIMIENTOS

RESUMEN

El principal beneficio para una organización con el uso de inteligencia artificial es que aumenta el rendimiento de los trabajadores y la productividad de la empresa, ya que haciendo uso de ésta se puede enseñar a las máquinas a hacerse cargo de los procesos rutinarios y repetitivos, con esto los trabajadores humanos pueden aprovechar mejor su tiempo, esto simplifica todos los procesos de gestión y control de los datos para organizar la información de una mejor forma. Sin embargo, como todo también tiene un problema y es que los sistemas de inteligencia artificial solo se encargan de ejecutar, y no explica sus acciones lo que genera desconfianza en los trabajadores humanos, por lo que en la actualidad está surgiendo el concepto de “aprendizaje por refuerzo explicable” (explainable reinforcement learning), el cual trata de dotar al agente con la capacidad de justificar sus acciones, un método para lograr esto es el método “Explicable basado en la memoria”, este ya ha sido utilizado en entornos pequeños con un problema simple a resolver. En este trabajo se busca aplicar este método de explicaciones en un entorno más grande y jerárquico, resolviendo una tarea más compleja para verificar si es posible dotar al agente con la capacidad de explicar sus acciones. Hemos realizado experimentos en un escenario simulado, un mundo de cuadrículas limitado de 10x10 en el cual se busca resolver un problema con algunas jerarquías marcadas y obtener la explicación en base a probabilidades. Los resultados obtenidos muestran que es posible resolver una tarea compleja dividiéndola en 3 tareas de alto nivel más sencillas y explicar en base a probabilidad el comportamiento del agente.

Palabras Claves: Aprendizaje por refuerzo, Explicable, Jerárquico.

ABSTRACT

The main benefit for an organization with the use of artificial intelligence is that it increases the performance of workers and the productivity of the company, since by making use of this, machines can be taught to take charge of routine and repetitive processes, with this human workers can make better use of their time, this simplifies all data management and control processes to organize information in a better way. However, like everything else, it also has a problem and that is that artificial intelligence systems are only in charge of executing, and it does not explain their actions, which generates distrust in human workers, which is why the concept of "learning by explainable reinforcement ", which tries to provide the agent with the ability to justify their actions, a method to achieve this is the " explainable memory-based "method, this has already been used in small environments with a simple problem to solve. This work seeks to apply this method of explanations in a larger and more hierarchical environment, solving a more complex task to verify if it is possible to provide the agent with the ability to explain his actions. We have carried out experiments in a simulated scenario, a world of limited 10x10 squares in which we seek to solve a problem with some marked hierarchies and obtain the explanation based on probabilities. The results show that it is possible to solve a complex task by dividing it into 3 simpler high-level tasks and explaining the agent's behavior based on probability.

Key Words: Reinforcement learning, Explainable, Hierarchical..

INDICE

1. Introducción.	3
1.1. Motivación	3
1.2. Descripción del problema	4
1.3. Objetivos	6
1.3.1. Objetivo General	6
1.3.2. Objetivos Específicos	6
1.3.3. Hipótesis	6
1.4. Alcance y limitaciones	7
1.4.1. Alcances	7
1.4.2. Limitaciones	7
1.5. Metodología	8
1.6. Cronograma.	11
2. Marco teórico	12
2.1. Aprendizaje por refuerzo.	12
2.1.1. Que es el aprendizaje por refuerzo	12
2.1.2. Aprendizaje por refuerzo tareas jerárquicas	14
2.2. Inteligencia artificial explicable	15
2.3. Redes neuronales	16
2.3.1. Que es una red neuronal.	16
2.3.2. Función perdida en una red neuronal	16
3. Métodos propuestos.	17
3.1. Método explicable basado en la memoria	17

3.2. Método de escape.	19
4. Escenario experimental	22
4.1. Problema a resolver	22
4.2. Reglas de Entrenamiento:	23
4.3. Medios.	24
4.4. Primera tarea de alto nivel	25
4.5. Segunda tarea de alto nivel	26
4.6. Tercera tarea de alto nivel	26
5. Resultados experimentales.	27
5.1. Cálculo de probabilidad en un entorno 10x10 sin obstáculos	28
5.2. Entrenamiento de la primera tarea de alto nivel	30
5.3. Entrenamiento de la segunda tarea de alto nivel	33
5.4. Entrenamiento de la tercera tarea de alto nivel	36
5.5. Probabilidad General	39
6. Conclusión:	41
6.1. Trabajos futuros	41
7. Bibliografía	43

1. Introducción.

En los últimos años los sistemas de inteligencia artificial han tomado cada vez más fuerza y lo implementan más empresas, estos sistemas de I.A (Inteligencia artificial) facilitan el trabajo repetitivo de una empresa, son entrenados mediante diversos métodos de entrenamiento. Para esto se usa el aprendizaje por refuerzo, estos agentes solo se dedican a resolver la tarea es por esto que pretende dotar a estos agentes con la capacidad de explicar su acción en base a probabilidades, se ubicará un agente en una grilla cerrada de 10x10 con el objetivo de llegar a un estado meta cumpliendo ciertas condiciones esto se hará mediante el método análisis-síntesis, consiste en separar las partes de un todo para estudiarlas de forma individual y la reunión de los elementos para estudiarlos en su totalidad.

1.1. Motivación.

Como ya se mencionó anteriormente en la actualidad la inteligencia artificial está tomando cada vez más fuerza y así aumenta las interacciones entre sistema y usuario, pero el usuario al usar la IA muchas veces no entiende porque la IA hizo eso o actuó de cierta forma lo que puede generar desconfianza en el usuario a pesar de obtener el resultado esperado, es por esto que haciendo uso de las técnicas de aprendizaje por refuerzo se busca entrenar un agente en un entorno de tareas jerárquicos para resolver un problema y además de esto implementar el concepto de Aprendizaje por refuerzo explicable para dotar a este agente con la capacidad de explicar porque realizo esa acción.

1.2. Descripción del problema.

El aprendizaje automático ha tomado mucha fuerza estos últimos años, se trata de una rama de la inteligencia artificial la cual se basa en que las máquinas sean capaces de aprender nuevas tareas que antes solo se destinaban a humanos calificados.

Actualmente esta rama se utiliza para resolver problemas en diferentes áreas tales como estadística, probabilidad, investigaciones profundas de datos o reconocimiento de patrones (*Gramajo, E., García-Martínez, R., Rossi, B., Claverie, E., Britos, P., & Totongi, 1999*).

En un mundo que cada vez hay mayor cantidad de interacciones entre usuarios y sistemas, el área de la inteligencia artificial es cada vez más relevante dado que ayuda a mejorar la interacción entre usuario sistema y haciendo uso de ella se puede facilitar la comunicación entre estos.

Si bien el aprendizaje por refuerzo ya se ha ganado un lugar como un enfoque de aprendizaje eficaz una vez se entrena el agente para lograr el objetivo este solo se dedica a ejecutar las acciones necesarias para lograrlo, es decir no proporciona ninguna explicación sobre porque decidido ejecutar esa acción por sobre las demás (*S. Zepesvári, C, 2010.*).

Esto puede llegar a suponer un problema, ya que al momento de que un usuario no experto en el tema vea trabajar al agente, pero no entiende que está haciendo o porque lo hace esto podría llegar a generar desconfianza del usuario hacia el agente dado que para el humano es más fácil confiar en algo que entiende.

Una posible forma para que un agente explique su comportamiento es en base a los valores Q y recompensas futuras, pero para usuarios promedios esto no tendrá sentido, Para solucionar esto se hará uso aprendizaje por refuerzo explicable. Se propone un aprendizaje por refuerzo explicable basado en la memoria para lograr dar una explicación en términos de probabilidad.

1.3. Objetivos.

1.3.1. Objetivo General.

Entrenar un agente mediante aprendizaje por refuerzo jerárquico que sea capaz de proporcionar una explicación en términos de probabilidad sobre porque elige cierta acción.

1.3.2. Objetivos Específicos.

- Reunir material bibliográfico sobre aprendizaje por refuerzo jerárquico y métodos para explicar las acciones con aprendizaje por refuerzo explicable.
- Definir entorno con tareas de alto nivel en el cual se desenvolverá el agente.
- Entrenar un agente con aprendizaje por refuerzo jerárquico.
- Dotar al agente con la capacidad de dar explicaciones en términos de probabilidad haciendo uso de aprendizaje por refuerzo explicable.

1.3.3. Hipótesis.

La hipótesis a comprobar en este trabajo será verificar si un agente entrenado mediante aprendizaje por refuerzo jerárquico es capaz de explicar sus acciones o por que elige cierta acción en términos de probabilidad para así aumentar la confianza del usuario con el sistema.

La idea de esta hipótesis es entregar confiabilidad al personal que está observando las acciones que toma el agente y porque de tal forma que alguien sin previo conocimiento en aprendizaje automático o aprendizaje por refuerzo sea capaz de entenderlo.

1.4. Alcance y limitaciones.

1.4.1. Alcances.

El entrenamiento encarga únicamente de aprender a resolver las tareas de alto nivel ya establecidas, mediante aprendizaje por refuerzo jerárquico.

El método de explicaciones propuesto se basa únicamente en entregar la explicación con una matriz en base a probabilidad.

1.4.2. Limitaciones.

En esta grilla de 10x10 el agente tiene limitadas las acciones que provoquen que este salga de la grilla.

Al tener 3 tareas de alto nivel el entrenamiento tiene que ser dividido en 3, ya que bajo este enfoque no puede realizarse solo con un entrenamiento.

En el entorno tanto los obstáculos como la meta son estáticos, esto no está diseñado para un entorno cambiante a lo largo del tiempo.

1.5. Metodología.

El método que se emplea en esta investigación es el método análisis-síntesis. Es un método que consiste en la separación de las partes de un todo para estudiarlas en forma individual (Análisis), y la reunión racional de elementos dispersos para estudiarlos en su totalidad. (Síntesis) (*Guevara Giovanni, 2014*).

Las capacidades de análisis y síntesis parten en base a la lectura, investigación, discusiones para intercambiar ideas, practicas etc. Este proceso de análisis-síntesis pasa a ser parte de nuestra vida cotidiana y también.

El análisis-síntesis se puede dividir en 3 fases: una fase previa el preanálisis, análisis y síntesis.

El preanálisis es una fase previa al análisis esta es una fase de recolección de información, ya sea documentos o en el campo, que permite formar una idea de las condiciones que se plantea un proyecto. Pero esta queda como una serie de datos aislados.

El análisis busca analizar la información y relacionarla entre sí, y con ello realizar un diagnóstico de donde sea posible partir hacia una síntesis.

La síntesis es donde se establecen las primeras idea guía para una propuesta, y esta se plasma de una forma concreta, ya sea de forma verbal o grafica.

Existen herramientas que nos facilitan el análisis de una situación como lo son:

- Señalar la idea más importante
- Comparar ideas
- Identificar puntos de controversia
- Identificar relaciones
- Buscar palabras claves
- Señalar distintos aspectos del tema que se han considerado.
- Revisar datos, esquemas, figuras, conceptos
- Gráficos y diagramas que permiten visualizar el análisis de los componentes-

También existen herramientas para la síntesis, que tiene por objetivo captar lo esencial, dar un orden jerárquico a las ideas tales como:

- El subrayado
- El resumen
- Esquemas
- Cuadro comparativo
- Mapas conceptuales
- Fichas de resumen
- Mapas mentales
- Glosarios de términos
- Fórmulas científicas
- Cuadro sinóptico

Las reglas del método de análisis-síntesis son (*Guevara Giovanni, 2014*):

- Observar el comportamiento de un fenómeno partes y componentes.
- Descripción. Identificar todos sus elementos partes y componentes.
- Examen crítico. Revisar rigurosamente cada elemento.
- Descomposición. Análisis del comportamiento y características de cada elemento
- Enumeración. Desintegración de los componentes a fin de identificarlos, registrarlos y establecer sus relaciones con los demás.
- Ordenación. reacomodar las partes del todo descompuesto para regresarlo a su estado original
- Clasificación. Ordenar cada una de las partes por clases, siguiendo el patrón del fenómeno analizado, para conocer su comportamiento y características.
- Conclusión. Analizar los resultados obtenidos, estudiarlos y buscar una explicación del fenómeno observado.

Para comenzar el correcto desarrollo del proyecto lo primero que se debe hacer es observar la situación actual e identificar la problemática, la falta de una explicación o justificación al momento de que un agente entrenado con RL decida tomar una decisión.

Luego se comienza con el desarrollo y diseño del agente buscando que por medio de aprendizaje por refuerzo explicable capaz de entregar una explicación del porque decidió hacer eso.

1.6. Cronograma.

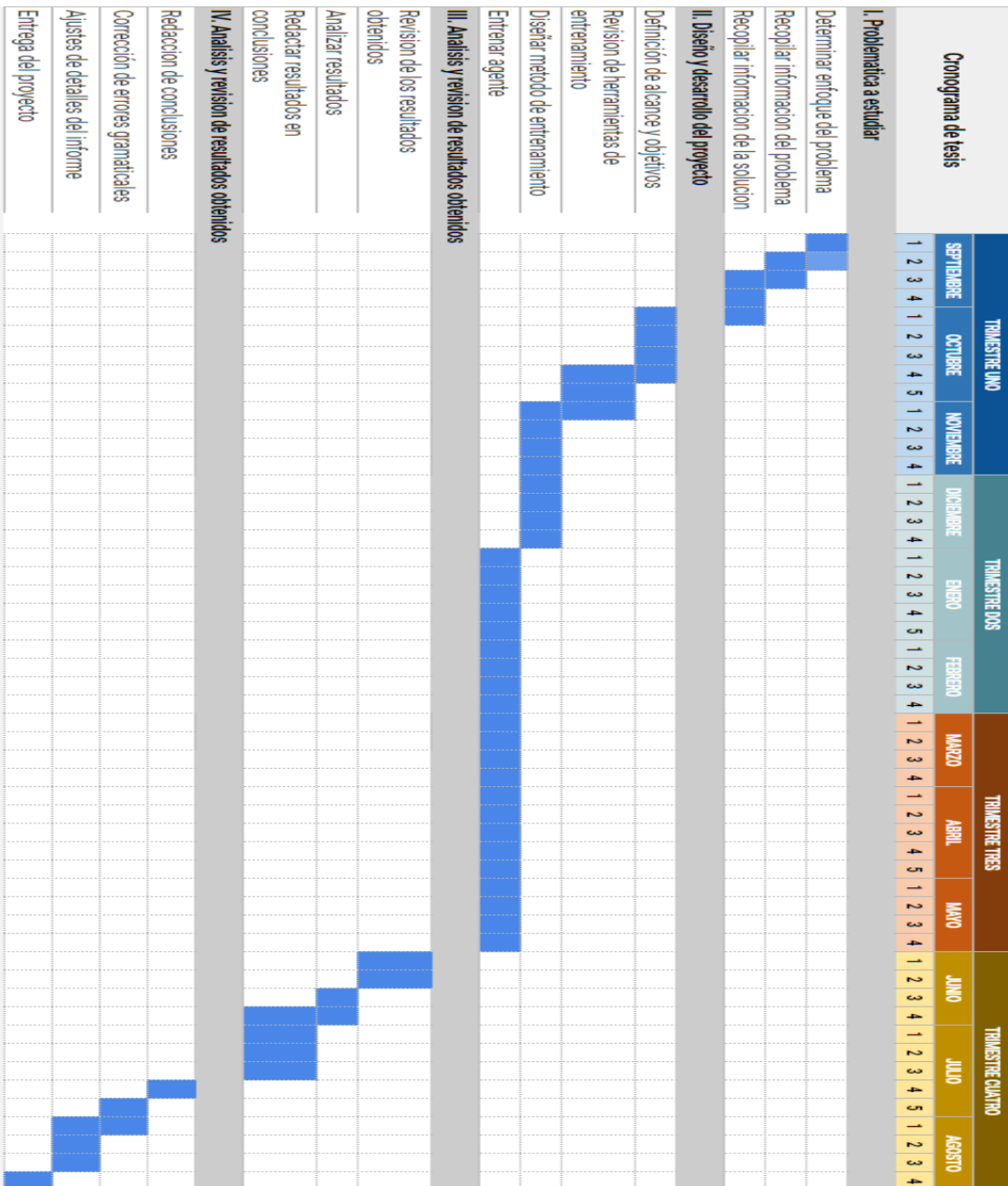


Figura 1: cronograma de trabajo

2. Marco teórico.

2.1. Aprendizaje por refuerzo.

2.1.1. Que es el aprendizaje por refuerzo.

Todo ser vivo exhibe algún tipo de conducta, es decir realizan acciones como respuesta a las señales o estímulos que reciben del entorno en el que se encuentra, algunos incluso llegan a modificar su comportamiento con el paso del tiempo de forma que al recibir señales del entorno frecuentemente van adquiriendo distintos comportamientos, esto se puede asociar a que los seres vivos van aprendiendo de su entorno (*Sancho F.S.C, 2017*).

El campo de aprendizaje por refuerzo (reinforcement learning) es un subcampo del aprendizaje automático el cual se basa en enseñarle a un agente como tomar decisión o elegir que acción realizar en cierto entorno para cumplir con un objetivo. Para simular el aprendizaje de sistemas biológicos reales es necesario tener claro los siguientes conceptos que se relacionan entre sí (*Richard Sutton, Andrew Barto, 2018*). como muestra la Figura 2.

Agente: el programa que entrena, con el objetivo de hacer algún trabajo.

Entorno: el mundo en el que el agente realiza las acciones.

Acción: un movimiento realizado por el agente que provoca un cambio de estado.

Recompensa: la evaluación de una acción esta puede ser positiva o negativa.

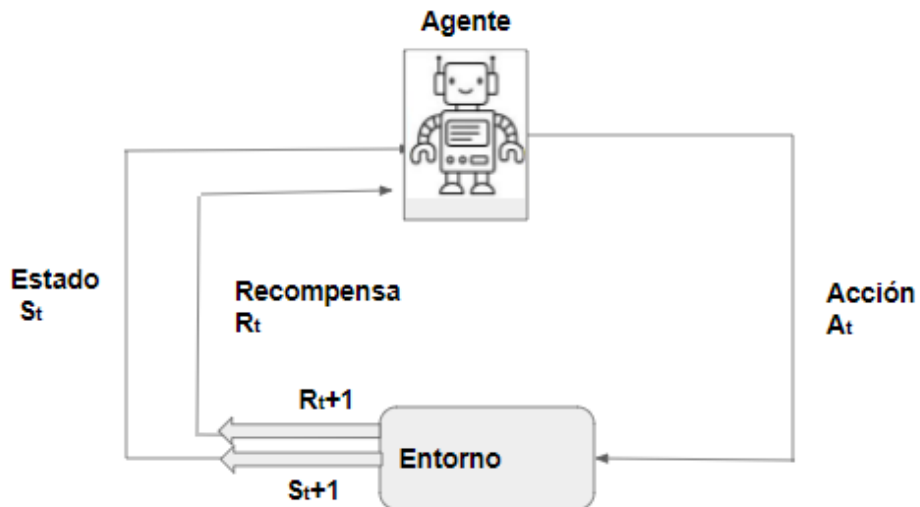


Figura 2: Aprendizaje por refuerzo; Se ubica un agente en un entorno desde un estado 0 para que ejecute una acción en el esto emite una recompensa y un nuevo estado, la recompensa puede ser positiva o negativa, el agente ahora ejecuta la acción en el nuevo estado con lo que obtiene una nueva recompensa y otro nuevo estado. (Adaptado desde el libro de Sutton and Barto).

Por ejemplo, si habláramos del clásico juego “Pacman”

- El **agente** es el mismo Pacman.
- El **estado** es la ubicación dentro del laberinto.
- La **recompensa** puede ser positiva los puntos por comer o negativa al morir
- Las **acciones** son las posibles direcciones en que se desplazara la agente izquierda, derecha, arriba y abajo

2.1.2. Aprendizaje por refuerzo tareas jerárquicas.

Los seres humanos comúnmente son capaces de resolver desafíos complejos o complicados dividiéndolo en pequeñas partes más manejables, también aprenden nuevas tareas rápidamente mediante de la secuencia de los componentes aprendidos, aunque la tarea requiera demasiadas de acciones de bajo nivel (*Frans. Ho. Chen. Abbeel. Schulman, K.F and J.H abd C.A and J.S, 2017*).

Por ejemplo, el hacer panqueques consiste en una serie de acciones de alto nivel como agregar huevos, revolver los huevos etc. Los humanos pueden aprender nuevas tareas de forma rápida clasificando estas partes aprendidas, incluso si esta tarea puede requerir millones de acciones de bajo nivel, como contracciones musculares.

Actualmente el aprendizaje por refuerzo se lleva a cabo por la búsqueda de fuerza bruta para acciones de bajo nivel, este método es ineficiente para tarea que requieren un largo tiempo.

Los agentes representan comportamientos complicados como una secuencia corta de acciones de alto nivel, gracias a esto el agente puede resolver tareas más complejas, si cierta solución llegase a requerir 2000 acciones de bajo nivel gracias a la política jerárquica esto se convierte en una secuencia de 10 acciones de alto nivel (*Frans. Ho. Chen. Abbeel. Schulman, K.F and J.H abd C.A and J.S, 2017*).

2.2. Inteligencia artificial explicable.

La inteligencia artificial explicable, viene siendo un conjunto de procesos y métodos que permite al humano comprender y confiar en los resultados y la información generada por una IA, esta IA explicable se utiliza para describir un modelo de IA (Inteligencia artificial) y su impacto esperado, esta es fundamental en una organización para generar confianza en estos.

Con el tiempo las IA han ido avanzado poco a poco, por esto los humanos son desafiados a comprender como el algoritmo consiguió ese resultado. Todo este proceso de cálculo vendría siendo lo que se conoce como un modelo de “caja negra” así ni un experto en el área logra comprender totalmente lo que está sucediendo dentro de esta “caja negra” y como el algoritmo consiguió resolver la tarea.

El poder comprender como un sistema de IA consigue un resultado específico aporta mucho a los desarrolladores para poder asegurarse que el sistema funcione como es esperado (*Miguel, D. G. V. 2020*).

2.3. Redes neuronales.

2.3.1. Que es una red neuronal.

Las redes neuronales artificiales están inspiradas en el funcionamiento del cerebro humano. Este modelo está formado por un conjunto de nodos llamados neuronas artificiales que están conectadas y transmiten señales entre sí, Estas señales se transmiten desde la entrada hasta genera salida.

Estos modelos de redes neuronales tienen como objetivo principal aprender modificándose a sí mismo de tal forma que podría ser capaz de aprender tareas complejas que no podrían ser realizadas mediante la clásica programación basada en reglas (*Izaurieta, F., & Saavedra, C. 2000*).

2.3.2. Función perdida en una red neuronal.

Una función de pérdida, es aquella función que evalúa la diferencia entre las predicciones realizadas por la red neuronal y los valores obtenidos realmente, cuando menor sea el valor de esta función durante el entrenamiento mejor será la red neuronal, se busca reducir lo más posible la diferencia antes mencionada, para ello se ajustan los distintos pesos de la red neuronal (*Jalil, M. A., & Misas, M. 2007*).

3. Métodos propuestos.

3.1. Método explicable basado en la memoria.

El método Aprendizaje por refuerzo explicable basado en la memoria (memory-Base Explainable Reinforcement Learning).

Los valores Q sirven para explicar el comportamiento de un agente RL, esto busca que el agente sea capaz de entregar explicación en términos que tengan sentido para todo tipo de usuario sea o no sea conocedor de aprendizaje RL. Estas explicaciones pretenden dar la probabilidad de éxito al elegir una acción en cierto estado y el número de transiciones para que el agente logre la tarea. Una vez logrado esto el agente puede entregar la explicación en base a probabilidades porque eligió una acción sobre otra lo que es más comprensible para un usuario común y se le dará una idea al usuario de cuantos pasos serán necesarios para lograr el objetivo.

Se propone un aprendizaje por refuerzo explicable basado en la memoria para calcular la probabilidad de éxito y las transiciones necesarias para lograr el objetivo, consiste en un agente RL con memoria episódica. Al acceder a la memoria del agente se puede comprender el comportamiento del agente en función de su experiencia.

Para esto se implementa una lista de par estado-acción "T" esta lista contiene todas las transiciones que realiza el agente durante su aprendizaje, ahora para poder calcular la probabilidad de éxito también se debe guardar el número total de transiciones que realizó en agente esta será "Tt" y también el número de transiciones en una secuencia de éxito este será "Ts". "Tt" y "Ts" son matrices de par estado acción.

Cada vez que el agente alcanza el estado final se calcula la probabilidad de éxito dividiendo “Tt” en “Ts” (probabilidad de éxito= Ts/Tt) (Francisco Cruz, Richard Dazeley , y Peter Vamplew, 2019).

```

1: Initialize  $Q(s, a), T_t, T_s, P_s, N_t$ 
2: for each episode do
3:   Initialize  $T_{List}[]$ 
4:   Choose an action using  $a_t \leftarrow \text{SELECTACTION}(s_t)$ 
5:   repeat
6:     Take action  $a_t$ 
7:     Save state-action transition  $T_{List}.add(s, a)$ 
8:      $T_t[s][a] \leftarrow T_t[s][a] + 1$ 
9:     Observe reward  $r_{t+1}$  and next state  $s_{t+1}$ 
10:    Choose next action  $a_{t+1}$  using softmax action selection method
11:     $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$ 
12:     $s_t \leftarrow s_{t+1}; a_t \leftarrow a_{t+1}$ 
13:  until  $s$  is terminal (goal or aversive state)
14:  if  $s$  is goal state then
15:    for each  $s, a \in T_{List}$  do
16:       $T_s[s][a] \leftarrow T_s[s][a] + 1$ 
17:    end for
18:  end if
19:  Compute  $P_s \leftarrow T_s/T_t$ 
20:  Compute  $N_t$  for each  $s \in T_{List}$  as  $\text{pos}(s, T_{List}) + 1$ 
21: end for

```

Figura 3: Algoritmo de Método explicable basado en la memoria: Enfoque de aprendizaje por refuerzo explicable basado en la memoria con el método de política SARSA para calcular la probabilidad de éxito y el numero de transiciones al estado objetivo (Francisco Cruz, Richard Dazeley , y Peter Vamplew, 2019).

3.2. Método de escape.

Para resolver este problema se hará uso del aprendizaje por refuerzo de tareas jerárquicas se busca dividir el problema en acciones de alto nivel, las cuales son alcanzar el estado 31, alcanzar el estado 93 y finalmente alcanzar el estado 7. Así el agente aprenderá como resolver esta secuencia de 3 tareas de alto nivel en lugar de aprender miles y miles de acciones de bajo nivel (en este caso el elegir cualquiera de nuestros 4 movimientos permitidos es una tarea de bajo nivel. Se establecen 3 tareas de alto nivel).

Se asigna la entrada "X" a la salida "Q" a través de la ecuación $Q = W(X)$, en donde la X representa el estado actual donde se encuentra el agente, es decir su ubicación en el laberinto y Q las acciones que puede seleccionar (arriba/ abajo/ izquierda/ derecha), Esta representa los valores Q para cada acción, la acción con un valor Q más alto indica una recompensa futura prevista más alta.

La función $Q = W(X)$ intenta predecir la mejor acción Q para el estado actual X de tal manera que el agente logre esquivar los obstáculos y encontrar el mejor camino hacia la salida. Esta función se puede aproximar mediante una red neuronal completamente conectada mediante retro propagación que esta entrenada para optimizar la predicción del valor Q en el estado X (*Yuk-Hoi Yiu, 2018*).

0	1	2	3 -100	4	5	6	7 500	8	9
10	11	12	13 -100	14	15	16	17	18	19
20 -100	21	22 -100	23	24	25	26	27	28	29
30	31 200	32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47	48	49
50	51	52	53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	68	69
70	71	72	73	74	75	76	77	78	79
80	81	82	83	84	85	86	887	88	89
90	91	92	93 200	94	95	96	97	98	99

Figura 4: Mapa de estados y recompensa; Cada ubicación en el laberinto de cuadrícula de 10 x 10 está indexada al número de estado, como se muestra en esta figura. Por ejemplo, la esquina superior izquierda del laberinto está indexada al estado 0, la esquina inferior derecha (la meta) está indexada al estado 99. Cada casilla (o estado) tiene su recompensa, como se muestra en el mapa de estados y recompensas. Por ejemplo, los estados 3, 13, 20 y 22 representan en realidad la ubicación de los obstáculos. Queremos castigar al agente por realizar acciones que conduzcan al estado 3, 13, 20, 22 por lo que recibirá una recompensa negativa (-100) en esos estados. Al llegar al estado 31 cumple con la primera tarea de alto nivel es por que recibirá una recompensa de (+200), al alcanzar el estado 93 cumple la segunda tarea de alto nivel y obtiene una recompensa de (+200) y finalmente en el estado 7 obtiene una recompensa de 500 al terminar la tercera tarea de alto nivel y con ello todo el problema.

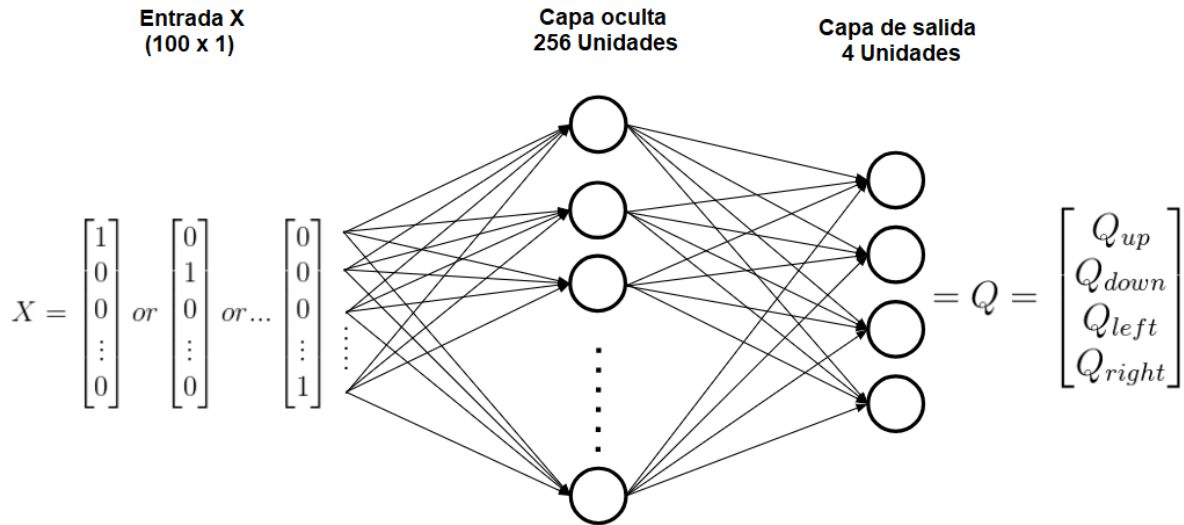


Figura 5: Red neuronal utilizada; Se asignan los estados actuales “X” a las opciones “Q “con La función $W(X)$. La red neuronal toma el vector de entrada (100 x 1), con 1 en la cierta fila que corresponde al cierto estado, y con 0 en las otras filas. La entrada (100 x 1) alimenta a las 256 neuronas de la capa oculta. Finalmente, la red genera las recompensas futuras previstas de las cuatro acciones posible (arriba/ abajo/ izquierda/ derecha) (Yuk-Hoi Yiu, 2018).

4. Escenario experimental.

4.1. Problema a resolver.

Se tiene un astronauta (agente) dentro de una nave en el espacio, el cual debe volver a casa, para lograrlo necesita cruzar un agujero de gusano. Para ello antes debe recoger un escudo, de lo contrario la nave explotará al intentar cruzar este agujero, además hay varios agujeros negros en el entorno, en los cuales si el astronauta cae perderá de inmediato y no podrá volver a casa.








0 	1	2	3 	4	5	6	7 	8	9
10	11	12	13 	14	15	16	17	18	19
20 	21	22 	23	24	25	26	27	28	29
30	31	32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47	48	49
50	51	52	53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	68	69
70	71	72	73	74	75	76	77	78	79
80	81	82	83	84	85	86	887	88	89
90	91	92	93 	94	95	96	97	98	99

Figura 6: Estados; Los estados van desde el 0 hasta el 99. El estado 0 (lugar donde está la nave) representa el estado inicial y el 7 (agujero de gusano) como estado final (meta). Hay agujeros negros en los estados 3, 13 ,20 y 22, el agente pierde al caer en alguno de estos agujeros negros, además para volver a casa se debe pasar sí o sí por el estado 93 a recoger el escudo, de lo contrario a pesar de llegar al estado 7 no se logrará el objetivo.

4.2. Reglas de Entrenamiento:

Este problema será resuelto mediante aprendizaje por refuerzo, estableciendo ciertas reglas:

- Hay 4 acciones (subir / bajar / izquierda / derecha), el agente puede seleccionar cualquier acción, lo que mueve al agente una casilla en la dirección seleccionada.
- El agente no puede caer en un agujero negro o perderá enseguida y será multado con una recompensa negativa de -100.
- El agente no puede salir de esta grilla 10x10, las acciones que lo harían salir están limitadas.
- El agente debe salir de la zona de agujeros negros, esta zona se refleja en la Figura 7. Al lograr salir de dicha zona obtiene una recompensa de 200.
- Para poder escapar el agente tiene que pasar sí o sí por el estado 93 y recoger el escudo, al recoger el escudo obtiene una recompensa de 200.
- El agente obtiene una recompensa de 500 al llegar a la meta (estado 7) habiendo cumplido las reglas de alcanzar el estado 93.
- Cuando el agente llega a la meta, o pierde de alguna forma el episodio de entrenamiento termina y comienza un nuevo episodio.

4.3. Medios.

Para el correcto entrenamiento del agente se necesitan equipos con buenas características ya sea en términos de procesadores, graficar o para cálculo de datos debido a esto usar la computadora personal para desarrollar el proyecto se vuelve una limitante es por esto que se optó por hacer uso de la plataforma de acceso libre creada por Google llamada Colab, esta plataforma permite un entorno de ejecución haciendo uso de jupyter notebooks y programar en Python (lenguaje usado en este proyecto).

Se decidió usar esto debido a la facilidad para correr el proyecto, además de ser una herramienta gratuita cuenta con librerías comunes preinstaladas y la opción de instalar otras que necesitemos, el código es fácil de compartir y al ser una plataforma de Google cuenta con servidores dedicados a este tipo de cálculo (*Bisong, E. 2019*).

4.4. Primera tarea de alto nivel.

Como se puede ver en la Figura 6. El agente parte en el estado 0 y debe llegar al 7 con el escudo para poder escapar.

La primera tarea de alto nivel que debe afrontar nuestro agente es escapar de la esquina superior izquierda, como se logra apreciar. Los agujeros negros rodean a nuestro astronauta en esa esquina por lo que si no se logra escapar de la zona que muestra la Figura 7. Nunca podrá volver a casa (se considera que escapa de la zona de agujeros negros al alcanzar cualquier estado que no esté en la Figura 7).

0 	1	2	3 
10	11	12	13 
20 	21	22 	23

Figura 7: Primera tarea de alto nivel

4.5. Segunda tarea de alto nivel.

Una vez fuera de esa esquina el agente se desenvuelve en la parte del laberinto que muestra la Figura 8. Ahora no puede simplemente ir al estado 7 para escapar antes debe conseguir el escudo que está en el estado 93, por lo que pisar este estado será la segunda tarea de alto nivel.

				4	5	6	7	8	9
				14	15	16	17	18	19
				24	25	26	27	28	29
30	31	32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47	48	49
50	51	52	53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	68	69
70	71	72	73	74	75	76	77	78	79
80	81	82	83	84	85	86	887	88	89
90	91	92	93	94	95	96	97	98	99

Figura 8: Entorno de segunda y tercera tarea de alto nivel.

4.6. Tercera tarea de alto nivel.

Una vez haya escapado de la esquina superior izquierda y recogido el escudo, el agente sigue desenvolviéndose en la parte del laberinto que muestra la Figura 8. puede ejecutar la tercera tarea de alto nivel, la cual es encontrar la salida del laberinto o sea llegar al estado 7 con el escudo.

5. Resultados experimentales.

Para comenzar cada entrenamiento se ubica el agente en la posición inicial (dependiendo de que tarea de alto nivel aprenderá) en cada uno se utilizan los parámetros de $\epsilon = 0,3$ y $\gamma = 0,7$ con lo que queda un 70% de explotación y un 30% exploración por lo que al momento de elegir cierta acción el agente un 70% de las veces elegirá la acción con más posibilidades de cumplir su objetivo mientras que el otro 30% de las ocasiones elegirá una acción de forma aleatoria para explorar posible nuevo camino.

Para conseguir las explicaciones se utilizó el “Método explicable basado en la memoria”.

5.1. Cálculo de probabilidad en un entorno 10x10 sin obstáculos

La primera prueba para conseguir el cálculo de probabilidad fue en una grilla de 10x10 simple sin obstáculos como se puede apreciar en la Figura 9. donde el agente solo debe alcanzar el estado 99 partiendo en el estado 8, así una vez verificado el método de explicación se aplica al entorno establecido anteriormente.



0 	1	2	3	4	5	6	7	8	9
10	11	12	13	14	15	16	17	18	19
20	21	22	23	24	25	26	27	28	29
30	31	32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47	48	49
50	51	52	53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	68	69
70	71	72	73	74	75	76	77	78	79
80	81	82	83	84	85	86	87	88	89
90	91	92	93	94	95	96	97	98	99 

Figura 9: entorno de 10x10 sin obstáculos

Las probabilidades de éxito se ven reflejadas en el mapa de calor que muestra la Figura 10. Donde el eje y representa los 100 estados y el eje x representa las 4 acciones:

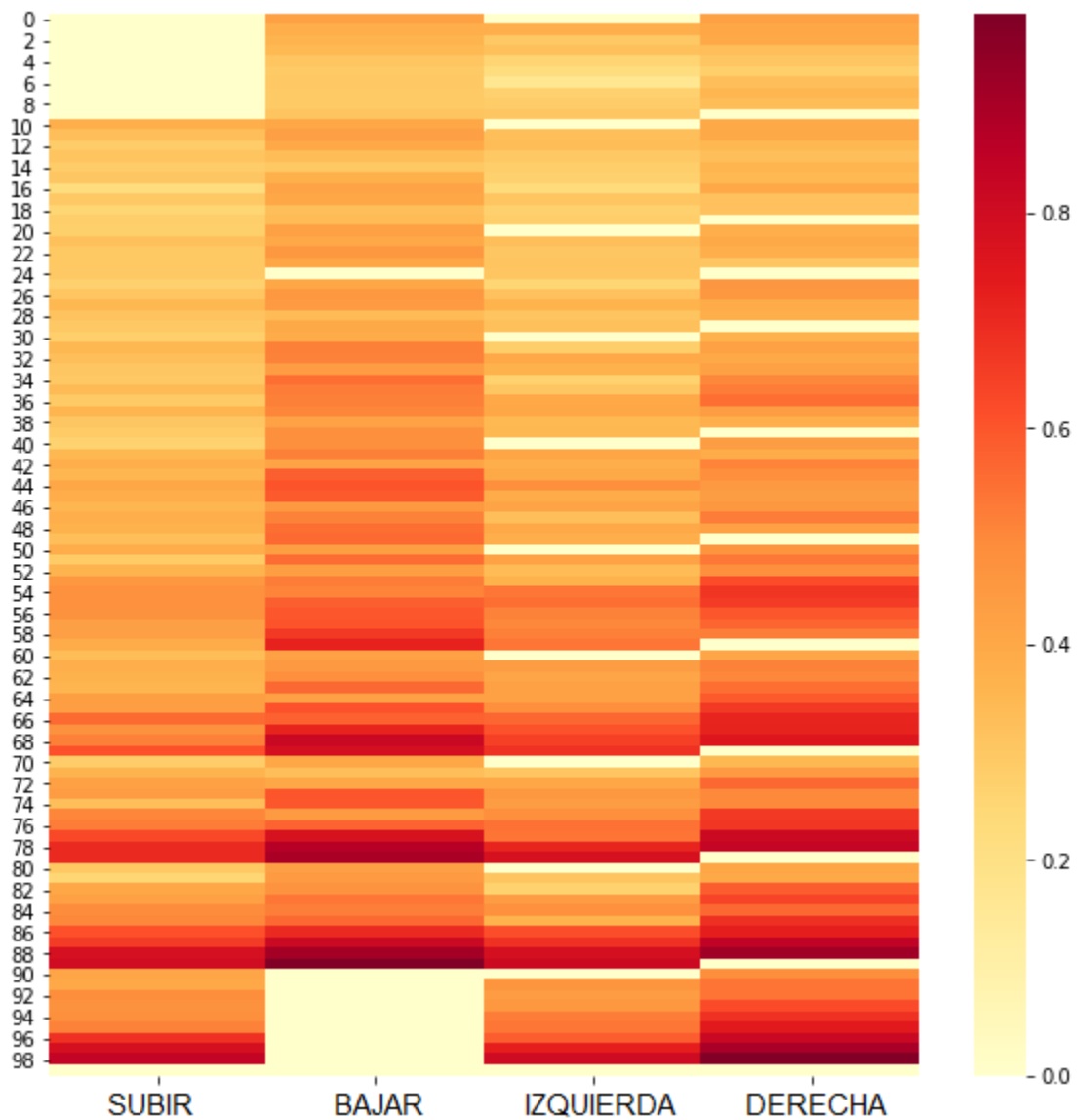


Figura 10: Probabilidades en entorno 10x10 sin obstáculos: Como se logra apreciar en el gráfico, mientras la acción más nos acerque a la meta mejor será la probabilidad, por ejemplo, en estado 98 acción derecha esa casilla tiene un 100% de probabilidades de éxito, el estado 89 acción bajar también tiene un 100% de probabilidades de éxito, ya que siempre que se ejecuten esas acciones en esos estados se entrara al estado meta (99).

5.2. Entrenamiento de la primera tarea de alto nivel

En el primer entrenamiento el agente tiene como punto de partida el estado 0 y como meta el estado 31, aquí solo busca escapar de la zona de agujeros negros sin caer en ellos. Para esto se utilizaron 10.000 episodios con 10 iteraciones en cada uno.

Mejor camino encontrado: 0, 1, 11, 21 ,31.

La Figura 11. muestra los valores de perdida en el tiempo del entrenamiento.

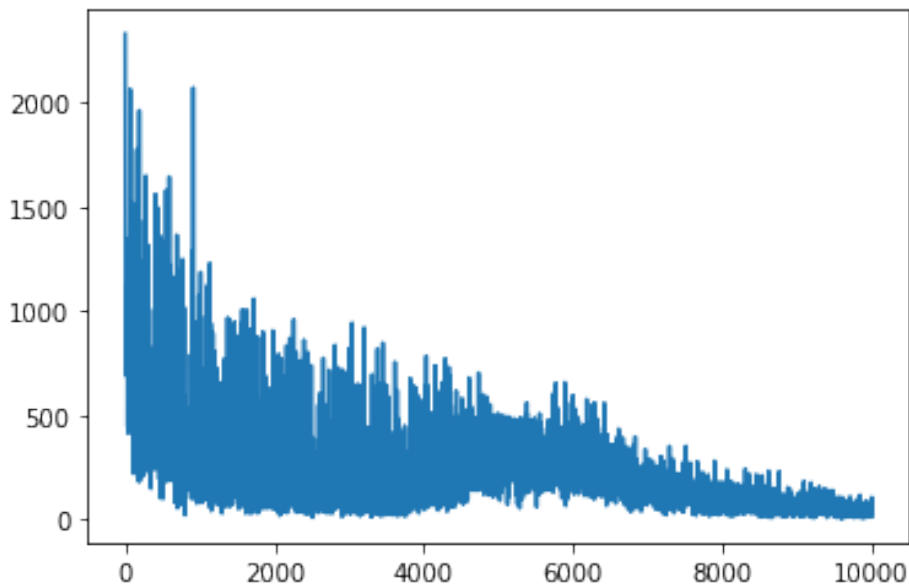


Figura 11: En esta figura se puede apreciar como los valores de perdida comienzan a converger hacia 0 ya en los 10.000 episodios por lo que se determinó que en este entrenamiento 10.000 episodios de entrenamiento son suficientes para lograr el objetivo.

Las probabilidades de éxito se ven reflejadas en el mapa de calor que muestra la Figura 12. Donde el eje y representa los 100 estados y el eje x representa las 4 acciones:

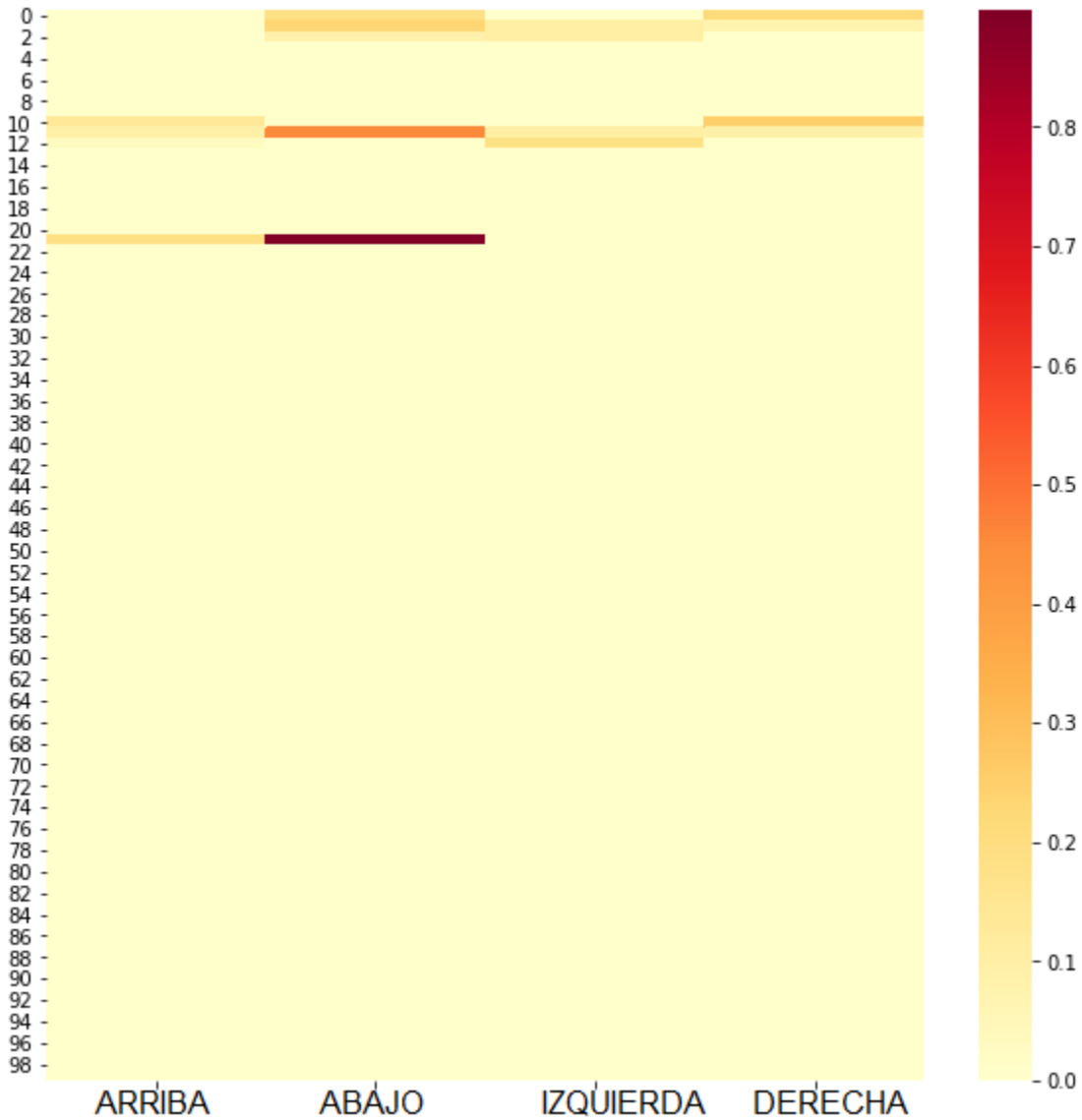


Figura 12: Probabilidades primera tarea de alto nivel en mapa de calor: Como se logra apreciar en el gráfico, los estados 3, 13, 20 y 22 o cualquier acción que haga al agente entrar en alguno de estos estados siempre tiene probabilidad de éxito 0, ya que estos son estados de pérdida, también se observa que mientras la acción más nos acerque a la meta mejor será la probabilidad, por ejemplo en estado 21 acción bajar esa casilla tiene un 100% de probabilidades de éxito ya que siempre que el agente baje desde la casilla 21 escapara de la zona de

agujeros negros completando la primera tarea de alto nivel, ahora en el estado 11 la acción de bajar tiene una alta probabilidad de éxito pero no el 100%, dado que al bajar llegamos a la casilla 21 desde la cual aún se podría perder moviéndose a izquierda o derecha, mientras la acción derecha e izquierda de la casilla 11 tienen una probabilidad más baja pero no nula ya que a pesar de alejarse de la meta aún tiene posibilidades de ganar.

5.3. Entrenamiento de la segunda tarea de alto nivel

En el segundo entrenamiento el agente tiene como punto de partida el estado 31 y como meta el estado 93, aquí busca recoger el escudo. Para esto se utilizaron 14.000 episodios con 100 iteraciones en cada uno.

Mejor camino encontrado: 31, 41, 51, 61, 71, 81, 91, 92, 93.

La Figura 13. muestra los valores de perdida en el tiempo del entrenamiento.

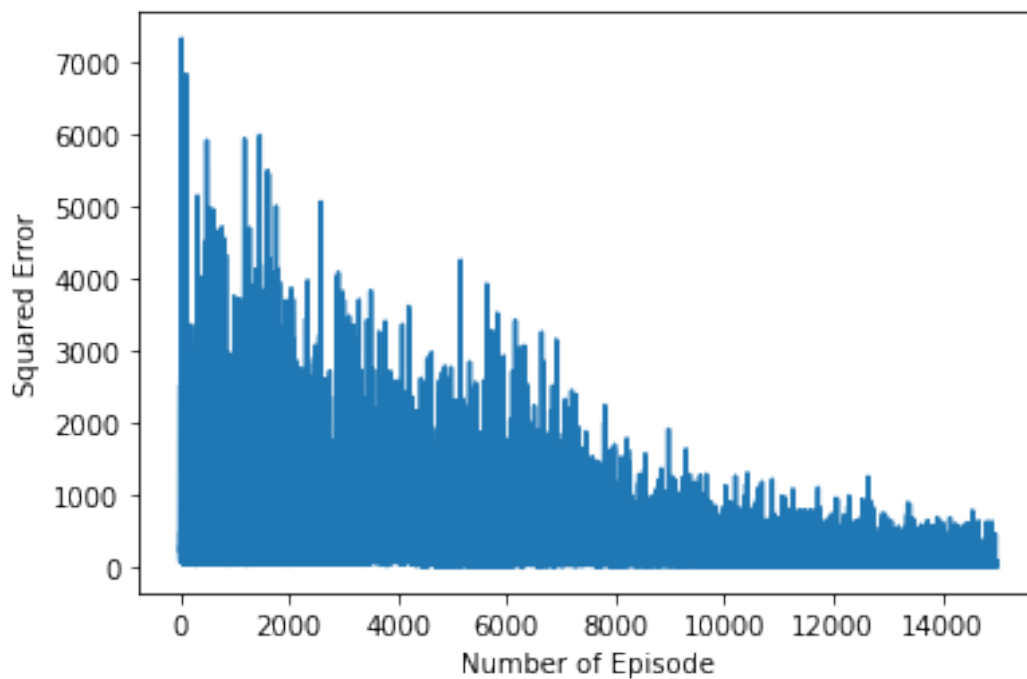


Figura 13: En esta figura se puede apreciar como los valores de perdida comienzan a converger hacia 0 ya en los 14.000 episodios por lo que se determinó que en este entrenamiento 14.000 episodios de entrenamiento son suficientes para lograr el objetivo.

Las probabilidades de éxito se ven reflejadas en el mapa de calor que muestra la Figura 14. Donde el eje y representa los 100 estados y el eje x representa las 4 acciones:

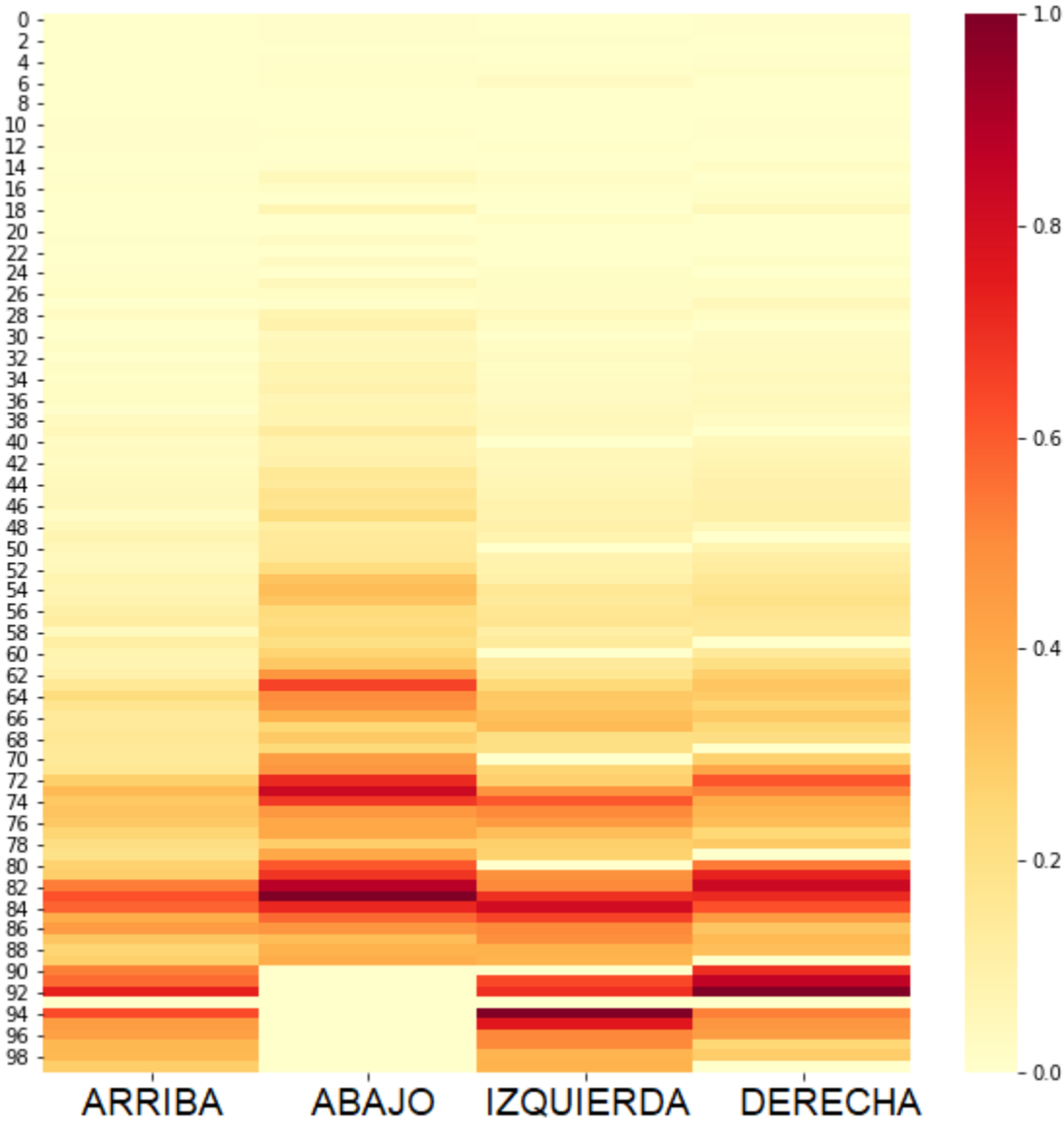


Figura 14: Probabilidades de segunda tarea de alto nivel; Mientras más se acerca el agente a la meta (estado 93) más altas son las probabilidades de éxito, en este caso hay 3 acciones que entregan un 100% de probabilidad de éxito, las 3 acciones con las que se entra al estado 93: del estado 92 hacia la derecha, del

estado 83 hacia abajo y del estado 94 hacia la izquierda, ya que siempre que se ejecute una de estas 3 acciones se entrara en el estado meta, es decir garantizan ganar ahora las demás acciones de estos estados al estar cerca de la meta aún mantienen una probabilidad bastante alta de victoria, por ejemplo en el estado 83 el bajar brinda un 100% de éxito mientras que el subir, derecha e izquierda aún mantienen una alta probabilidad de éxito pero no garantizan la victoria.

5.4. Entrenamiento de la tercera tarea de alto nivel

En el tercer entrenamiento el agente tiene como punto de partida el estado 93 y como meta el estado 7, aquí busca cruzar el agujero de gusano y volver a casa. Para esto se utilizaron 20.000 episodios con 100 iteraciones en cada uno.

Mejor camino encontrado: 93, 94, 95, 96, 97, 87, 77, 67, 57, 47, 37, 27, 17, 7.

La Figura 15. muestra los valores de perdida en el tiempo del entrenamiento.

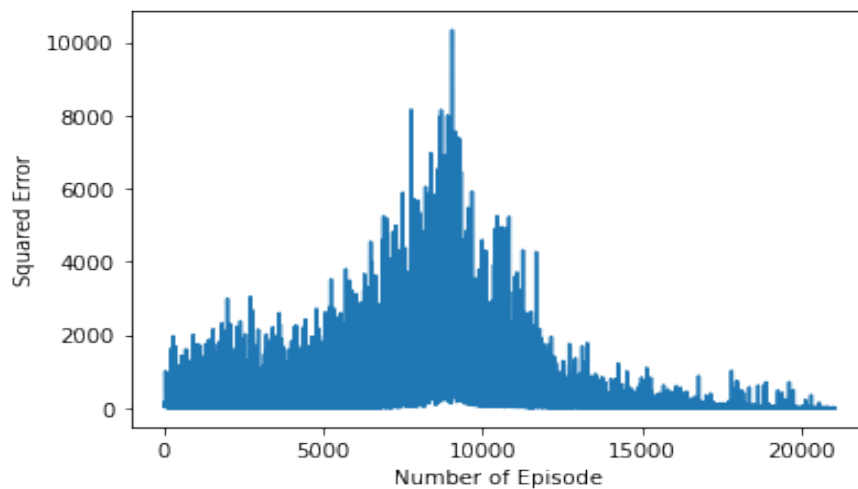


Figura 15: En esta figura se puede apreciar como los valores de perdida comienzan a converger hacia 0 ya en los 20.000 episodios por lo que se determinó que en este entrenamiento 20.000 episodios de entrenamiento son suficientes para lograr el objetivo.

Las probabilidades de éxito se ven reflejadas en el mapa de calor que muestra la Figura 16. Donde el eje y representa los 100 estados y el eje x representa las 4 acciones:

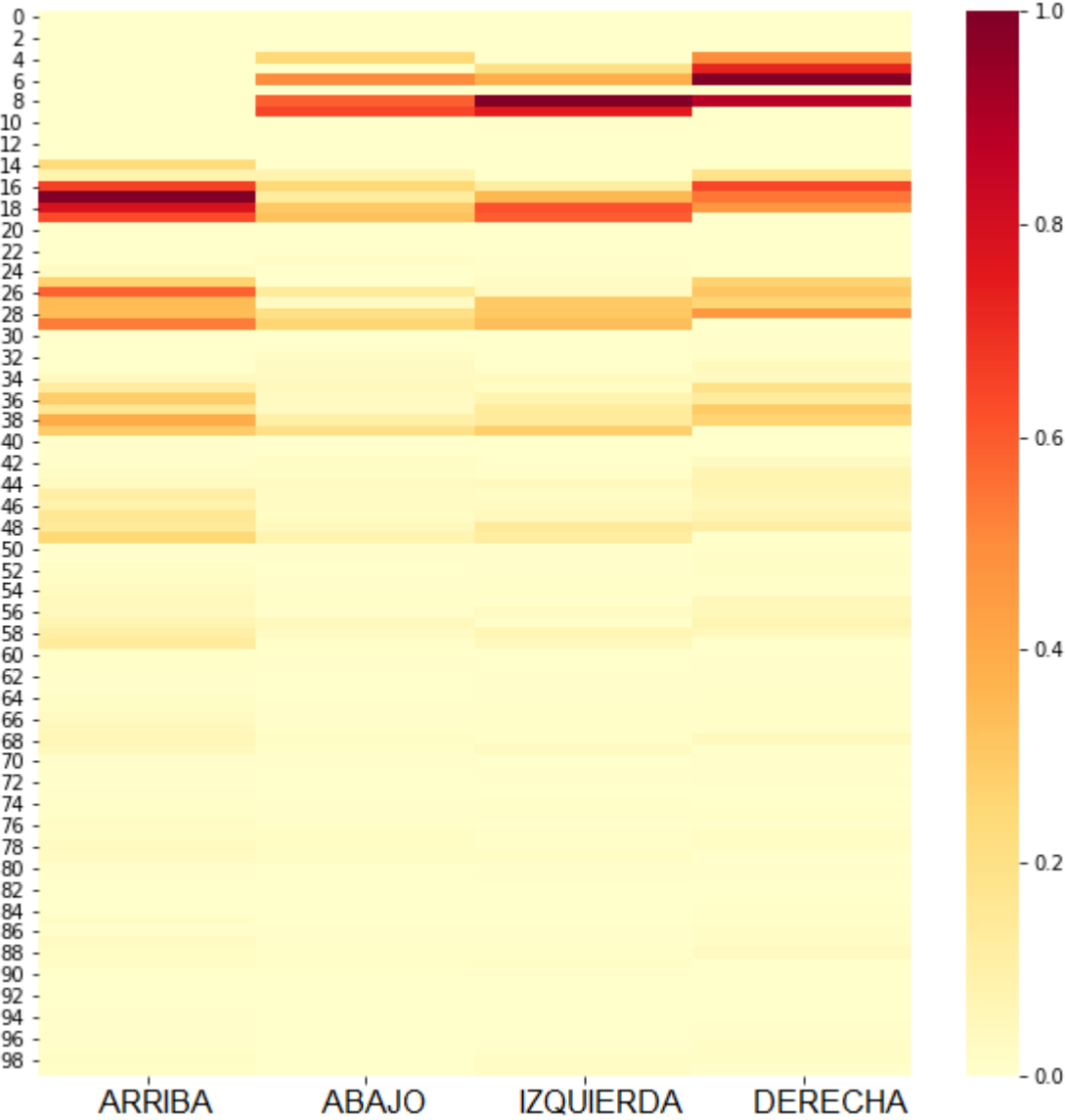


Figura 16: Probabilidades de tercera tarea de alto nivel; Mientras más se acerca a la meta (estado 7) más altas son las probabilidades de éxito, en este caso hay 3 acciones que entregan un 100% de probabilidad de éxito (ya que al haber completado previamente la segunda tarea de alto nivel se cuenta con el escudo), están vendrían siendo las 3 acciones con las que se entra al estado 93:

del estado 92 hacia la derecha, del estado 83 hacia abajo y del estado 94 hacia la izquierda, siempre que se ejecute una de estas 3 acciones se entrara en el estado meta, es decir garantizan ganar, ahora las demás acciones de estos estados al estar cerca de la meta aún mantienen una probabilidad bastante alta de victoria, por ejemplo en el estado 83 el bajar brinda un 100% de éxito mientras que el subir, derecha e izquierda aún mantienen una alta probabilidad de éxito pero no garantizan la victoria.

5.5. Probabilidad General

Como se pudo apreciar en el ítem anterior se obtuvo un cálculo de probabilidad para cada una de las tareas de alto nivel en las que se dividió el problema, pero eso no nos permite tener una idea general de las probabilidades en términos globales del problema sino solo de una tarea de alto nivel a la vez, es por esto que se busca establecer una matriz de probabilidades global, para esto se calcula el promedio de las 3 matrices de probabilidades obtenidas anteriormente, esta matriz se puede observar en la Figura 17.

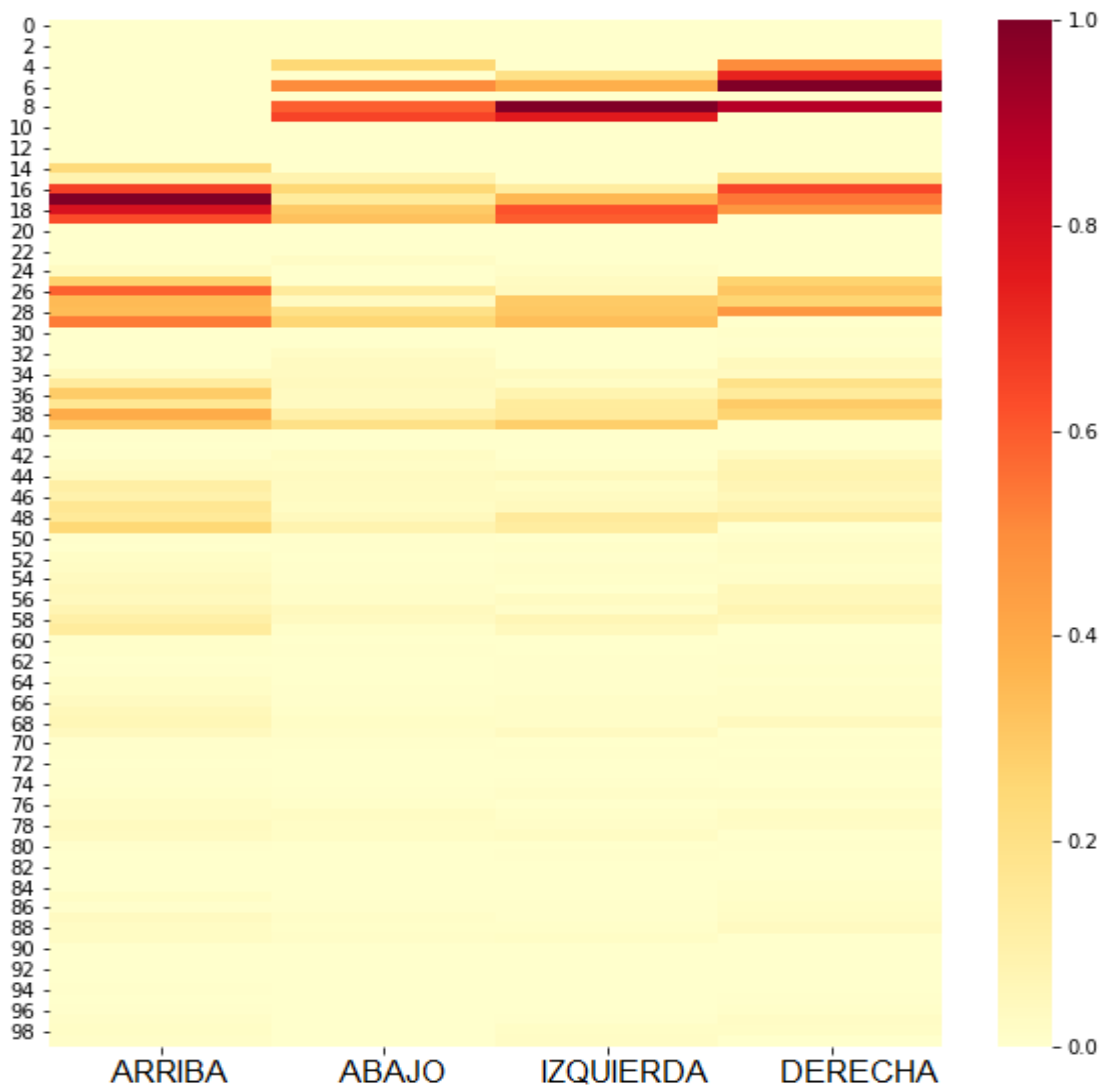


Figura 17: Matriz global; A diferencia de las figuras las otras 3 matrices esta matriz no representa solo una tarea de alto nivel, sino que engloba las 3 para así

obtener las probabilidades generales del problema, Al representar la probabilidad de manera global se puede apreciar de mejor manera el comportamiento del agente, por ejemplo en los estados 0, 1, 2, 3, 4, 5, 6, 7, 8 y 9 la acción subir siempre será 0, en los estados 0,10, 20, 30, 40, 50, 60, 70, 80, 90 la acción izquierda siempre será 0, en el estado 9, 19, 29, 39, 49, 59, 69, 79, 89 y 99 la acción derecha siempre será 0 y en los estados 90, 91, 92, 93, 94, 95, 96, 97, 98 y 99 la acción bajar siempre será 0 ya que el agente tiene prohibido salir de esta grilla 10x10, para esto en la programación de este se limitaron estas acciones y no pueden ser ejecutadas. Los estados 3, 13, 20, 22 y cualquier acción que entre a uno de estos tiene probabilidad 0, ya que estos son estados de perdida. En esta matriz a diferencia de las anteriores ninguna acción nos brinda un 100% de probabilidad de éxito porque ninguna acción garantiza la victoria. El entrar a la casilla 7 (meta) no garantiza la victoria ya que se reduce a la pregunta ¿al entrar en esta casilla tiene el escudo?

Si no lo tiene perderá en lugar de ganar, lo mismo con las acciones que hagan al agente entrar al estado 93 no tienen un 100% de victoria al recoger el escudo porque no garantiza la victoria aún debe cruzar el agujero de gusano y en el camino podría caer en un agujero negro con lo que se pierde inmediatamente.

Las probabilidades de éxito más altas se obtienen al completar las tareas de alto nivel sin embargo ninguna llega a ser de 100% dado que en este problema ninguna acción puede garantizar la victoria.

6. Conclusión:

En el presente trabajo queda en evidencia que el método Explicable basado en la memoria es aplicable a un algoritmo “Maze Runner” en un entorno de grilla 10x10, este entrega las probabilidades sobre ejecutar una acción en cierto estado en forma de una matriz 100x4 (estados x acción), además para comprobar la efectividad de este método poco a poco se fueron agregando obstáculos a esa grilla de 10x10 hasta construir el escenario jerárquico con el que se trabajó, y ver cómo se comporta el método Explicable basado en la memoria en esta situación, al ser un escenario jerárquico el entrenamiento se dividió en completar 3 tareas de alto nivel, este método fue capaz de brindar buenas matrices de probabilidad para explicar cada una de las tareas de alto nivel (una matriz por cada tarea), pero esto solo otorga la explicación sobre cómo cumplir cada tarea de alto nivel, por lo que para poder ver las probabilidades generales del problema se consiguió una matriz final promediando las 3 matrices obtenidas de las 3 tareas, esta matriz final también fue capaz de brindar buenas explicaciones sobre las probabilidades de victoria que tendrá el agente al ejecutar una acción en cierto estado, esta vez para resolver el problema de manera general, por lo que se concluye que el método explicable basado en la memoria es aplicable a un problema de aprendizaje por refuerzo en un entorno de tareas jerárquicas.

6.1. Trabajos futuros

Como trabajo futuro se propone explorar la opción de tener un entorno cambiante, es decir que tanto los obstáculos como la meta se muevan con el paso del tiempo, y a este aplicar el “método explicable basado en la memoria” para verificar su efectividad de explicaciones en un entorno aún más complejo.

7. Bibliografía.

1. Sancho F.S.C, marzo del 2017. Aprendizaje por refuerzo: algoritmo Q Learning.
2. Sutton and Barto, R. S. S. and A. G. B, enero del 2018. Reinforcement Learning: An Introduction.
3. Frans. Ho. Chen. Abbeel. Schulman, K.F and J.H abd C.A and J.S, Octubre del 2017. Learning a Hierarchy
4. Juozapaitis, Z., Koul, A., Fern, A., Erwig, M., & Doshi-Velez, F, Enero del 2019. Explainable reinforcement learning via reward decomposition. In *IJCAI/ECAI Workshop on Explainable Artificial Intelligence*.
5. Gramajo, E., García-Martínez, R., Rossi, B., Claverie, E., Britos, P., & Totongi, A, 1999 . Una visión global del aprendizaje automático. *Rev. Inst. Tecnológico B*.
6. S. Zepesvári, C. (2010). Algorithms for reinforcement learning. Synthesis lectures on artificial intelligence and machine learning
7. Guevara Giovanni, 2014. Síntesis y análisis.
8. Bisong, E. (2019). Google colabratory. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform* (pp. 59-64). Apress, Berkeley, CA.
9. Francisco Cruz, Richard Dazeley, y Peter Vamplew, 2019, memory-Base Explainable Reinforcement Learning.

10. Miguel, D. G. V. (2020). Aproximaciones a la Explicación de Decisiones Algorítmicas: Inteligencia Artificial Explicable.

11. Izaurieta, F., & Saavedra, C. (2000). Redes neuronales artificiales. *Departamento de Física, Universidad de Concepción Chile*.

12. Jalil, M. A., & Misas, M. (2007). Evaluación de pronósticos del tipo de cambio utilizando redes neuronales y funciones de pérdida asimétricas. *Revista Colombiana de Estadística*.

13. Yuk-Hoi Yiu, 2018, RL-maze-runner.

