Universidad Nacional Autónoma de México, Facultad de Psicología

Reinforcement Learning y Functional Experience Weighted Attraction en juegos de negociación y sistemas Agent Based que implican diferentes niveles de justicia

Hugo Gómez Cárdenas

Tesis

Director: Arturo Bouzas Riaño

Asesor: German Palafox Palafox

Ciudad de México, Marzo del 2025

Índice

1. Resumen **5**

2. Introducción **6**

2.1. Problema 6
2.2. Justificación 7
2.3. Literatura 9
2.4. Objetivos 12

3. Marco **14**

3.1. Aprendizaje por Refuerzo 14
3.2. Exploración - Explotación 17
3.3. Functional Experience Weighted Attraction 18
3.4. Teoría de juegos 22
3.5. Juego de Negociación 23
3.6. Aversión a la Inequidad 25
3.7. Agent Based 26

4. Métodos **28**

4.1. Consideraciones previas 28
4.2. Preguntas de investigación 28
4.2.1. Reinforcement Learning Q-Learning 28
4.2.2. Functional Experience Weighted Attraction 29
4.2.3. Reinforcement Learning Q-Learning y Functional Experience Weighted Attraction 29
4.2.4. Agent Based 30
4.3. Variables 31
4.4. Diseño 32
4.5. Muestra 32
4.6. Contexto 34
4.7. Procedimiento 34
4.7.1. Fase de aprendizaje 34
4.7.2. Fase de prueba de tamaños de población 43
4.7.3. Fase de prueba de distribuciones de población 48
4.8. Instrumentos 49

5. Cronograma **50**

6. Análisis **51**
6.1. Reinforcement Learning Q-Learning 51
6.2. Functional Experience Weighted Attraction 53
6.3. Reinforcement Learning Q-Learning y Functional Experience Weighted Attraction 56
6.4. Agent Based 70

7. Conclusiones **100**

8. Referencias **102**

9. Anexos **107**
9.1. Gráficos de frecuencias de selección por rival 107
9.1.1. Valores de epsilon para rival con preferencia baja por resultados ventajosos 107
9.1.2. Valores de epsilon para rival con preferencia media por resultados justos 109
9.1.3. Valores de epsilon para rival con preferencia alta por resultados ventajosos 110
9.2. Gráficos de valor que asignaron RL y FEWA por rival 112
9.2.1. Valores de epsilon para rival con preferencia baja por resultados ventajosos 112
9.2.2. Valores de epsilon para rival con preferencia media por resultados justos 113
9.2.3. Valores de epsilon para rival con preferencia alta por resultados ventajosos 115
9.3. Gráficos de acumulación de ganancias de RL y FEWA por rival y por tamaño de muestra 117
9.3.1. Valores de epsilon para rival con preferencia baja por resultados ventajosos 117
9.3.2. Valores de epsilon para rival con preferencia media por resultados justos 119
9.3.3. Valores de epsilon para rival con preferencia alta por resultados ventajosos 120
9.4. Aprendizaje de RL y FEWA por rival, por epsilon y por ronda 122
9.4.1. Valores de epsilon para rival con preferencia baja por resultados ventajosos en ronda 1 122
9.4.2. Valores de epsilon para rival con preferencia media por resultados justos en ronda 1 124
9.4.3. Valores de epsilon para rival con preferencia alta por resultados ventajosos en ronda 1 126
9.4.4. Epsilon promediado para rival con preferencia baja por resultados ventajosos en ronda 2 128

9.4.5. Epsilon promediado para rival con preferencia media por resultados justos en ronda 2 129
9.4.6. Epsilon promediado para rival con preferencia alta por resultados ventajosos en ronda 2 130
9.4.7. Epsilon promediado para rival con preferencia baja por resultados ventajosos en ronda 3 131
9.4.8. Epsilon promediado para rival con preferencia media por resultados justos en ronda 3 132
9.4.9. Epsilon promediado para rival con preferencia alta por resultados ventajosos en ronda 3 133
9.5. Gráficos de Recall y F1-score de RL y FEWA 134
9.5.1. Recall 134
9.5.2. F1-Score 138
9.6. Gráficos de desempeño de RL y FEWA en el sistema AB por distribución y por tamaño 142
9.6.1. Distribuciones para tamaño de muestra de 300 agentes 142
9.6.2. Distribuciones para tamaño de muestra de 50 agentes 170
9.7. Gráficos de Segregación 198
9.7.1. Distribuciones para tamaño de muestra de 150 agentes 198
9.7.2. Distribuciones para tamaño de muestra de 300 agentes 202
9.7.3. Distribuciones para tamaño de muestra de 50 agentes 205
9.8. Gráficos de Equilibrio 209
9.8.1. Distribuciones para tamaño de muestra de 300 agentes 209
9.8.2. Distribuciones para tamaño de muestra de 50 agentes 216

1. Resumen

Resumen

Los modelos de aprendizaje se han utilizado ampliamente para explicar resultados conductuales basados en principios cognitivos. En este trabajo, se implementó Reinforcement Learning Q-learning (RL) y Functional Experience Weighted Attraction (FEWA) dentro de un ambiente virtual que simula interacciones sociales estratégicas, como un juego de la negociación basado en Teoría de Juegos. Los modelos de aprendizaje compitieron contra tres tipos de agentes, cuyas preferencias están definidas por los parámetros del modelo de Aversión a la Inequidad, con el objetivo de comparar los resultados de cada modelo de aprendizaje según los parámetros rival. Finalmente, se analizará cómo los agentes RL y FEWA aplican las estrategias aprendidas en un ambiente con diferentes distribuciones heterogéneas de tipos de rivales en un sistema de Agent Based (AB). Los resultados muestran que FEWA se desempeña ligeramente mejor que RL contra los rivales que aprendieron previamente. Además, en los escenarios en los que FEWA es capaz de identificar rivales que RL y FEWA no aprendieron previamente, FEWA se desempeña notoriamente mejor.

Palabras clave: Modelos de aprendizaje, Teoría de Juegos, Agent Based.

2. Introducción

2.1. Problema

Los diseños experimentales de la Teoría de Juegos permiten estudiar la habilidad de tomar decisiones estratégicas. Sin embargo, hay gran cantidad de juegos, o variaciones de juegos, donde los agentes se desvían de los equilibrios derivados matemáticamente a partir de los supuestos y de las condiciones iniciales del juego. Creemos que la importancia de los modelos psicológicos que modifican la utilidad de las ganancias asociadas a las estrategias de los jugadores reside en mostrar cómo se aprenden creencias que desencadenan desviaciones de los equilibrios. Por lo tanto, estos modelos aportan respuestas acerca de cuáles condiciones de las reglas de un juego son necesarias para que un agente aprenda desviaciones que impliquen preferencias más allá de maximizar sus ganancias, por ejemplo, justicia, confianza o codicia.

Por otra parte, la inteligencia artificial ha avanzado a gran velocidad en su aplicación para la clasificación y generación de datos. Sin embargo, su uso para explicar o clasificar decisiones en situaciones que impliquen diferentes preferencias sociales también debería ser estudiada para la posible integración de interacciones humano-máquina, o retroalimentación al humano por parte de modelos de aprendizaje de máquina con el fin de mejorar la toma de decisiones en interacciones sociales estratégicas. Asimismo, creemos que es relevante utilizar estas herramientas para impulsar el progreso de las ciencias cognitivas y del comportamiento, llevando a cabo experimentos simulados que permitan poner a prueba estos modelos, e identificando sus fortalezas y debilidades en diferentes escenarios.

Finalmente, aunque logremos diseccionar cómo se aprenden y se toman decisiones considerando las preferencias del rival, no estamos contemplando escenarios más cercanos a la vida cotidiana. Las ciencias sociales constantemente se enfrentan a argumentos de que sus mediciones no capturan elementos más complejos de la realidad social. Estos escenarios implican que a lo largo del tiempo un agente llegará a interactuar con varios rivales con preferencias heterogéneas, también podría encontrarse en escenarios que impliquen interactuar con distintos grupos de rivales simultáneamente. Las simulaciones Basadas en Agentes se encargan de mostrar la complejidad que conlleva incrementar el número de agentes en un ambiente, estos sistemas identifican cuáles son las condiciones iniciales que se necesitan para que surjan patrones en una población a partir de reglas simples que aplica individualmente cada agente de la población. Dado que se busca estudiar interacciones sociales, es importante considerar entornos con la interacción de varios agentes con preferencias heterogéneas.

2.2. Justificación

El análisis de las interacciones sociales desde el marco de Teoría de Juegos tiene como objetivo de formalizar la toma de decisiones y sus preferencias asociadas, con el fin de desarrollar políticas públicas bien fundamentadas y asegurar su efectividad. Aportar a la literatura de teoría de juegos promueve una mejor comprensión del proceso de elección, lo que ayudará a las personas a tomar decisiones más informadas. Además, a niveles más elevados, ayuda a entender problemas sociales y darles solución.

Asimismo, las negociaciones son un proceso cotidiano e importante, con estos acordamos reparticiones de recursos o resolvemos conflictos. Los modelos formales de negociaciones son estudiados por diversos campos, como inteligencia artificial, economía, ciencias sociales y, evidentemente, teoría de juegos.

Los modelos psicológicos de teoría de juegos son importantes porque ofrecen explicaciones a resultados que se desvían de las soluciones matemáticas esperadas en un juego. Usar el modelo de Aversión a la Inequidad permite demostrar cómo y porqué la utilidad de las estrategias cambia dependiendo de las preferencias individuales y, por lo tanto, permite dar cuenta de por qué estrategias no dominantes son seleccionadas. Las aportaciones de un modelo psicológico en teoría de juegos tiene aplicaciones en interacciones sociales, en política y en economía, por ejemplo, para diseñar mecanismos y políticas que promuevan comportamientos cooperativos y decisiones justas.

Existen varios modelos que intentan explicar lo que las personas consideran como resultados justos. Sin embargo, algunos modelos imponen restricciones que pueden generar resultados poco plausibles cuando se busca simular elecciones y no se busca ajustar el modelo a un conjunto de datos. Por ejemplo, en el modelo de Justicia de Rabin (1993) que devuelve *equilibrios justos*, se deben asumir creencias a priori de un jugador en dos niveles de racionalidad si se quiere simular resultados, además, el modelo original se limita a juegos simultáneos de dos jugadores. Otro ejemplo es el modelo de Reciprocidad, también de Charness y Rabin (2002), el cual llega a un *equilibrio recíproco-justo*, sin embargo, necesita ajustar una gran cantidad de parámetros y un par de ellos adquieren valores diferentes cuando el rival se *portó mal* (*misbehaved*), lo cual es difícil de ajustar en simulaciones con agentes no sesgados. Por su parte, el modelo de Aversión a la Inequidad puede aplicarse a juegos de n jugadores y con n estrategias, por ello ha sido exitosamente usado para explicar conducta que se desvía de equilibrios en juegos económicos, como el juego del ultimátum, el juego del dictador, el juego de intercambio de regalos, el juego de mercados y el juego de bienes

públicos (Fehr & Schmidt, 1999). Además, el modelo cuenta únicamente con dos parámetros y solo se necesitan las estrategias y las ganancias de los jugadores para ser estimados, lo que lo hace viable para simular resultados.

El uso de simulaciones para probar modelos de aprendizaje es un método de investigación que proporciona una perspectiva diferente para estudiar conducta. Este campo brinda herramientas que facilitan la predicción de comportamiento bajo diferentes contextos, gracias a que las simulaciones de estos modelos buscan emular las condiciones de un laboratorio, lo que les permite generalizar sus resultados y corroborar la robustez de los modelos. Los modelos de aprendizaje pueden revelar qué elementos iniciales de los agentes y del ambiente se requieren para replicar y generalizar la conducta. Además, nos ayudan a entender cómo se aprenden políticas óptimas que maximizan el valor asociado a una acción, esto desde la perspectiva de un agente no sesgado.

La combinación de agentes artificiales que aprenden y de la teoría de juegos permite pensar qué necesita un agente para que tome decisiones óptimas en entornos que representan interacciones sociales, es decir, en entornos donde sus acciones dependen y afectan a otros agentes.

Existe una gran variedad de modelos de aprendizaje de máquina, sin embargo, el Aprendizaje por Refuerzo Q-learning es un modelo de aprendizaje que surge de principios psicológicos fundamentales, lo cual es importante por su valor en interpretabilidad y base teórica. Este modelo ha sido ampliamente utilizado en diferentes tareas por la facilidad de adaptabilidad a varios entornos virtuales y sus resultados eficientes.

Por su parte, Experience Weighted Attraction también es un modelo que surge de principios psicológicos de aprendizaje, y toma conceptos cognitivos para interpretar sus parámetros, sin embargo, no suele ser tan usado como Aprendizaje por Refuerzo por la cantidad de parámetros involucrados. A pesar de esto, la relevancia de Experience Weighted Attraction radica en combinar Aprendizaje por Refuerzo y Aprendizaje por Creencias.

La relevancia de los sistemas basados en agentes radica en que nos ayudan a entender los elementos con los que emergen fenómenos sociales, este método es útil incluso si se reducen los elementos involucrados en la interacción para simplificar su simulación, o si se lleva a cabo un mapeo más abstracto de los rasgos del fenómeno. De cualquier modo, los modelos formales de fenómenos sociales implican buscar argumentos matemáticos y/o computacionales de las características que componen un sistema, y así demostrar cómo ciertos supuestos llevan a

ciertas conclusiones. Dichas simulaciones de sistemas pueden, inicialmente, ser usadas para describir patrones en los resultados, sin embargo, se espera que estas descripciones eventualmente ayuden a formular un modelo, con parámetros que representen características de un sistema y que se relacionan con ciertos resultados. Por ejemplo, varios modelos iniciaron como simulaciones de relaciones lógicas y después fueron desarrollados para calibrar modelos que den cuenta de epidemias, evolución de elecciones sociales, formación de una estructura de redes sociales, etc (Smaldino, 2023).

2.3. Literatura

Tomando en cuenta la importancia de las herramientas mencionadas, creemos que tanto los modelos de aprendizaje como Aprendizaje por Refuerzo (*Reinforcement Learning*; RL) y Atracción Ponderada por Experiencia (*Experience Weighted Attraction*; EWA) son capaces de responder de forma óptima en diversas negociaciones que impliquen diferentes niveles de justicia, de acuerdo al modelo de Aversión a la Inequidad (*Inequity Aversion*; IA).

Anteriormente se han propuesto modelos de negociación autónoma diseñados con el objetivo de aplicarlos en e-commerce, como el modelo que proponen Cao y Kiang (2012), que considera creencias (información disponible), deseos (objetivos) e intenciones (la selección de una estrategia). Asimismo, Brzostowski y Kowalczyk (2006) usan regresión no lineal paramétrica de datos conductuales para proponer un agente de negociación autónoma. J. Zhang et al. (2015) crearon un agente de negociación bayesiano, donde en cada ronda actualiza la utilidad que el rival le asigna a cada oferta para determinar la distribución de probabilidad de las preferencias del rival, además, el agente toma en cuenta el decaimiento de la utilidad de las ganancias a lo largo del tiempo, por lo que tiene un número de rondas para hacer una oferta exitosa. Bazaar es un modelo conocido de Zeng y Sycara (1998), este responde bajo un marco de aprendizaje bayesiano, el agente actualiza la probabilidad de las preferencias del rival con la retroalimentación de la negociación, es aplicado a diferentes problemas de negociación y en ambientes multi-agente, sin embargo puede tener dificultades para obtener resultados óptimos y asume supuestos como que la distribución de probabilidad es normal. Baarslag et al. (2014) comparan los resultados de los modelos de negociación autónoma que fueron ganadores de la Competencia de Agentes de Negociación Autónomos (ANAC, por sus siglas en inglés), los evalúan tomando en cuenta sus condiciones de aceptación y bajo diferentes escenarios de negociación, los agentes fueron “Agent K”, “Yushu”, “Nozomi”, “Agent Smith”, de los cuales, su condición de aceptación está basado en la utilidad de las ganancias y en el tiempo de la negociación, y “IAMHaggler”, “FSEGA”, “IAMcrazyHaggler”, para

los que su condición de aceptación está únicamente basado en la utilidad de las ganancias. Además, Chen y Weiss (2015) también prueba los agentes ganadores del ANAC 2010-12 con su propio agente llamado OMAC*, con el cuál buscan modelar en tiempo real las preferencias de sus oponentes a través de discretización y regresión no lineal, el cuál, en comparación a los anteriores modelos, mostró resultados sobresalientes. Finalmente, Huang et al. (2010), presentan otro modelo para comercio virtual, una arquitectura de negociación basado en agentes que contiene compradores y vendedores, su simulación demostró que sus agentes lograron optimizar las ganancias tanto para los compradores como para los vendedores.

Como se puede observar, se han propuesto modelos específicamente diseñados para e-commerce (Baarslag et al., 2014; Brzostowski & Kowalczyk, 2006; Cao & Kiang, 2012; Chen & Weiss, 2015; Huang et al., 2010), es decir, modelos pre-entrenados para obtener los mejores resultados en negociaciones en línea, donde se prueban tácticas de negociación en escenarios bajo diversas variables, pero al fin y al cabo diseñados exclusivamente para negociar en estos entornos. Sin embargo, a nosotros nos interesa comparar el aprendizaje de modelos con bases teóricas conductuales y cognitivas en una interacción estratégica, comparar las dinámicas de elección (prueba y error) conforme progresa el aprendizaje de los modelos, y determinar qué sustentos teóricos favorecen los resultados a los que llegan en escenarios que implican diferentes niveles de equidad. También, aunque algunos de los anteriores trabajos sostienen que su agente estuvo en un sistema basado en agentes (Brzostowski & Kowalczyk, 2006; Huang et al., 2010), ninguno implica un análisis de segregación, evolución o contagio. Además, el trabajo de J. Zhang et al. (2015), como muchos que implican agentes bayesianos dependen de datos previos, el de Brzostowski y Kowalczyk (2006) aunque no sea bayesiano, también lo necesita.

Por otra parte, recopilando lo que se ha encontrado en la literatura respecto al uso del modelo de Aversión a la Inequidad (IA) con modelos de aprendizaje tenemos la investigación de Hughes et al. (2018), este consiste en un sistema Multi Agente con agentes de RL tipo Actor-Critic que han sido probados contra bots que responde de acuerdo al modelo de IA en una variación del juego de bienes comunes y en una variación del juego del esfuerzo, los autores muestran cómo la aversión a la inequidad promueve el aprendizaje de cooperación a lo largo del tiempo. En la investigación de S. Zhang et al. (2024) se explora cómo se evita inequidad al observar a otros, se usaron los datos de un juego del ultimátum para inferir valores de los parámetros de IA, para que estos luego se usen contra un agente de RL. También se ha probado RL tipo Q-learning en una variación espacial del juego de la batalla de los sexos, pero a la función de ganancias del agente se integra aversión a la desventaja con el modelo

de IA, esta integración la clasificaron como aversión a la pérdida, lo autores encontraron que la cooperación se acelera bajo estas condiciones (Gasparrini & Sánchez-Fibla, 2019). En una variación de Deep RL, los autores integran aversión a la culpa a partir del modelo de IA y creencias tipo Teoría de la Mente (*Theory of Mind*; ToM) para poner a prueba la capacidad de este nuevo agente para coordinarse con sus rivales en un juego de Stag Hunt, su desempeño es comparado con el de un agente que solo usa ToM, otro que solo usa aversión a la culpa con el modelo de IA y un último agente Deep RL, en su trabajo demostraron que su nuevo modelo de aprendizaje pueden aprender comportamiento cooperativo (Nguyen et al., 2020). En otra investigación se combina la función de utilidad de Homo Egualis con Continuous Action Learning Automata en un sistema Multi-Agent de un juego del ultimatum para aprender los valores de los parámetros de IA que acercan los resultados a ganancias equitativas, los autores demostraron que su agente llega a resultados compatibles con lo que se espera de resultados justos en humanos (Jong et al., 2008). A pesar de estos hallazgos, en nuestra búsqueda no hemos encontrado trabajos que de alguna manera combinen de forma directa y explícita el modelo de aprendizaje de Atracción Ponderada por Experiencia (EWA) e IA, o Atracción Ponderada por Experiencia Funcional (FEWA) e IA.

Además, dado que los sistemas basados en agentes (AB) parecen una buena alternativa para probar el aprendizaje de los agentes RL y EWA en entornos más heterogéneos de niveles de justicia, nos dimos a la tarea de buscar lo que se ha probado con AB y IA. El trabajo de Ahmed y Karlapalem (2014) ejecuta simulaciones de un sistema AB donde los agentes tienen preferencias heterogéneas de IA, los valores del modelo determinan si cada agente es cooperador o desertor en un juego del dilema del prisionero, terminan por surgir agrupamientos que crecen, ya que los agentes pueden cambiar sus preferencias (valores en sus parámetros de IA), la aportación de este artículo consiste en que en una población variada puede emerger cooperatividad entre agentes que inicialmente no son similares. Más cercano a lo que buscamos en esta investigación tenemos el trabajo de Kuperman y Risau-Gusman (2008), donde crearon un ambiente espacial con agentes con diversas preferencias de acuerdo a IA, los agentes juegan ultimátum con los agentes vecinos sin moverse de su posición, en este trabajo se demostró que los resultados de los agentes se ve influenciado por la topología, es decir, por la distribución inicial y el número de agentes. En un caso similar tenemos el trabajo de Xianyu (2010), donde agentes que actúan de acuerdo al modelo de IA negocian en un juego del ultimátum con sus vecinos en un sistema AB, este trabajo no incluye dinámicas espaciales ya que lo importante de este trabajo es en qué valor convergen los valores de los parámetros del modelo de IA para explicar la evolución de resultados justos. En otra investigación se toma como base el modelo de IA para modelar confianza y basarse en este

nuevo modelo para simular decisiones en un juego de bienes comunes en un sistema AB, en este trabajo no hay dinámicas espaciales como en el anterior ya que, una vez más, lo importante de este trabajo es cómo cambian los parámetros de cada agente (M. A. Janssen et al., 2022). En el artículo de M. Janssen (2014) se usaron datos obtenidos de un experimento de bienes públicos para ajustar un modelo que formaliza interacciones de agentes en un sistema AB, los agentes toman decisiones de acuerdo al modelo de IA y los resultados contestan a cómo los agentes resuelven la distribución de recursos de acuerdo a IA. También se han probado poblaciones con preferencias heterogéneas que provengan del modelo IA para determinar la habilidad y el esfuerzo de aportación en un juego de bienes públicos, pero sin estar en un sistema AB (Kölle et al., 2016). Una investigación que incluye un agente de aprendizaje relacionado a nuestros intereses es el trabajo de McKee et al. (2020), donde un agente Actor-Critic RL aprende de una población con valores diversos provenientes de Social Value Orientation (SVO), el artículo no usa el modelo de IA, o ningún juego de Teoría de Juegos en un sistema AB, los autores sugieren que la heterogeneidad lleva al agente RL a comportamientos más variados y complejos, que le ayudan a generalizar más en comparación a los que solo aprendieron de poblaciones homogéneas. Nuestra búsqueda en la literatura parece indicar que se han probado poblaciones con preferencias heterogéneas que provengan del modelo IA en diversos sistemas basados en agentes, sin embargo no hemos encontrado que se agreguen dinámicas de movimiento espacial en los agentes del sistema, ni que integren directamente o explícitamente agentes pre-entrenados que usen los modelos de aprendizaje de interés.

2.4. Objetivos

Los propósitos principales que guiaron la realización del trabajo presente son:

1. Entender la toma de decisiones estratégicas en situaciones homogéneas y heterogéneas del modelo de Aversión a la Inequidad, utilizando métodos de aprendizaje y simulaciones Basadas en Agentes.

2. Simular aprendizaje y toma de decisiones en escenarios con diversos niveles de justicia utilizando modelos de aprendizaje:

    a) Determinar qué necesitan los modelos de aprendizaje del entorno para aprender y adaptarse a una interacción social estratégica de Teoría de Juegos donde varían las preferencias de los rivales hacia estrategias justas o injustas:

1) Creación de un entorno virtual de negociación con las características de interés.
2) Implementación y adaptación de algoritmos RL y FEWA en el entorno virtual.

b) Analizar la evolución de la política óptima aprendida contra diferentes rivales que prioricen estrategias justas o injustas.

3. Simular interacciones y dinámicas en un sistema AB que incluya a los modelos de aprendizaje pre-entrenados y una población heterogénea de agentes con preferencias que provengan del modelo de Aversión a la Inequidad:

a) Determinar qué necesita un sistema AB para que emerjan dinámicas entre los agentes a partir de reglas simples:

1) Creación de un entorno virtual de negociación en un sistema AB con las características de interés.
2) Integración de algoritmos RL, FEWA y de población de agentes generada a partir de una distribución determinada en el entorno virtual.

b) Evaluar la clasificación de rivales conocidos y desconocidos por parte de los agentes RL y FEWA.

c) Evaluar las elecciones de políticas óptimas y el desempeño total de los agentes RL y FEWA.

d) Observar las dinámicas espaciales y de ganancias que emerjan del movimiento y de las interacciones de los agentes, para determinar si existen patrones de agrupamiento, aleatoriedad, cooperación, competición o explotación de agentes.

3. Marco

3.1. Aprendizaje por Refuerzo

Las bases de la Inteligencia Artificial (*Artificial Intelligence*; AI) fueron cimentadas por Turing (1936) al introducir la idea de una máquina de cómputo universal, capaz de llegar a resultados únicamente con su configuración inicial. Hoy en día, la AI es una herramienta en constante crecimiento y que ha cambiado drásticamente nuestra vida cotidiana. Los avances relevantes de esta área están asociados con *Machine Learning* (ML) (Samuel, 1959), la rama de Inteligencia Artificial que se ocupa de permitir que las máquinas aprendan a partir de datos. Una perspectiva más reciente de AI es *Deep Learning* (DL) (Brooks et al., 2012; LeCun et al., 2015; Saxe et al., 2021), una técnica basada en capas de redes neuronales artificiales generadas para aprender patrones más complejos de datos. Los autores del área de Deep Learning hace mucho énfasis en que los algoritmos de Inteligencia Artificial deben emular elementos biológicos del cerebro, sin embargo, nosotros tenemos más interés por herramientas con una base teórica sólida del comportamiento humano y cognición, como *Reinforcement Learning* (RL) (Silver et al., 2021), un área de estudio central en el campo de AI, y es la noción de que un agente aprende interactuando con el mundo real y es recompensado o penalizado por sus acciones, además, debido a su amplia aplicabilidad también ha sido integrado con redes neuronales (Mnih, 2013; Mnih et al., 2015).

Aprendizaje por Refuerzo es un modelo derivado de la *Ley del Efecto* de Thorndike (Thorndike, 1898, 1931), la cual estipula que si la respuesta de un agente a un estímulo del ambiente es recompensada, es probable que dicha respuesta sea repetida, asimismo, si es castigada será evitada, por lo que el aprendizaje de los agentes está basado en experiencia previa.

Así, Aprendizaje por Refuerzo, en la literatura de Inteligencia Artificial, es un campo de modelos que explican el aprendizaje como una interacción del agente con el ambiente, planteado como un *Proceso de Decisión Markoviano*. En RL la meta del agente es aprender la política óptima, es decir, generar un mapeo que asocie acciones para cada estado del ambiente. El algoritmo consta de tres elementos: la política (π), que puede entenderse como el conjunto de acciones que sigue un agente; la recompensa (R), que es es aquello que busca maximizar el agente; y la función de valor, que es la serie de valores que se le asigna a los estados (S) del ambiente, o conjuntos estados y acciones ($Q(S, A)$) del ambiente. Entonces, el agente debe seguir una política, que es determinada por una función de valor aprendida, para maximizar la recompensa (Sutton & Barto, 2018).

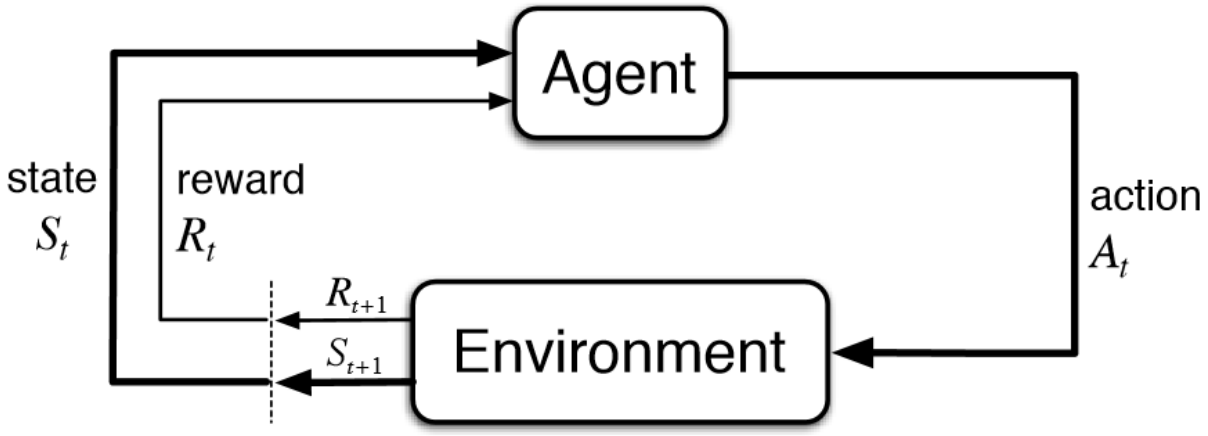

Figura 1: Proceso de Decisión Markoviano

No obstante, la política que ejecuta un agente en un inicio será diferente a la política final, ya que la política inicial promueve el aprendizaje por medio de la exploración y la política final promueve la maximización de recompensa por medio de explotación, por lo que a la política final se le conoce como política óptima, y para llegar a esta los algoritmos de aprendizaje por refuerzo pasan por tres pasos (Sutton & Barto, 2018).

El primero se le conoce como evaluación de la política que consiste en medir las recompensas que obtuvo el agente dada una cierta política, el segundo es mejora de la política y es donde se elige la política que obtuvo mejores resultados, el último es iteración de la política que busca resolver problemas como mínimos locales por medio de repetir los pasos de evaluación y mejora hasta llegar a la política óptima. Este proceso se repite hasta que ya no hay modificaciones significativas en la función de valor, por lo que se considera que ese sería el resultado óptimo, a este método se le llama *Iteración de Política Generalizada* y es la más común en los algoritmos de RL (Sutton & Barto, 2018).

Aprendizaje por Diferencia Temporal es el campo de RL al que pertenecen *Q-learning* (Watkins & Dayan, 1992) y su variante *SARSA* (Rummery & Niranjan, 1994), este campo combina los métodos de *Programación Dinámica* y *Monte Carlo*. La gran diferencia entre ambos algoritmos es que SARSA usa un método de selección de acciones conocido como *on-policy*, que implica que todas las acciones seleccionadas deben ser derivadas de la matriz de valor aprendida. Mientras que Q-learning usa un método llamado *off-policy*, que implica que en la etapa de exploración del agente se le permite seleccionar acciones con base en una política independiente de la matriz de valor aprendida, esto con el objetivo de no estancarse

en un máximo local, y durante la etapa de explotación selecciona una política que maximice la recompensa según la matriz de valor. La gran ventaja del método off-policy es que aventura al agente a explorar opciones desconocidas o que inicialmente tienen poco valor para evitar que el agente se estanque (Sutton & Barto, 2018).

Cuando el agente se encuentra explotando, una acción (a) del estado (s) tienen una probabilidad (p) de ser seleccionada proporcional a su recompensa asociada (r), en comparación con las recompensas asociadas a las otras acciones disponibles. Este heurístico proviene de la *Ley de Igualación* propuesta por el psicólogo Herrnstein (1970), en donde la idea principal es que la probabilidad de elegir una acción incrementa con la recompensa que esa acción ha generado en el pasado. Así, las probabilidades de cada acción inician de forma arbitraria y el proceso de aprendizaje se da a través de una actualización iterativa de los valores de los conjuntos estado-acción en una matriz de valor llamada *Q-value*, la cual determinará las futuras probabilidades.

La regla de actualización de Q-learning es la siguiente:

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha \left[R_{t+1} + \lambda \cdot \underset{a}{max} Q(s_{t+1}, a) - Q(S_t, A_t) \right]$$

Leída literalmente se entiende como el valor aprendido del conjunto estado-accción al que se llegó ($Q(S_t, A_t)$) más el error de predicción (TD) ponderado por el parámetro de aprendizaje (α). El error de predicción está comprendido como la diferencia entre el valor esperado del conjunto estado-acción al que se llegó ($Q(S_t, A_t)$) y la recompensa obtenida (R_{t+1}), a la que se le añade el valor máximo esperado del siguiente estado ($\underset{a}{max} Q(s_{t+1}, a)$) pero descontado temporalmente (λ). En otras palabras, el valor del conjunto estado-acción aumentará si se obtiene una recompensa desconocida o si la secuencia de estados inmediatamente siguientes aumentan de valor, y el valor de ese conjunto disminuye de valor si ya no se obtiene una recompensa conocida o si la secuencia de estados inmediatamente siguientes disminuyen de valor. El que la regla de actualización tome en cuenta el valor máximo del siguiente estado, en lugar de esperar a que acabe el episodio para actualizar toda la secuencia de selección como el método Monte Carlo, vuelve a Q-learning un algoritmo más eficiente.

3.2. Exploración - Explotación

Existen métodos para evitar que los agentes se estanquen en mínimos locales por falta de exploración de su entorno y para evitar que tengan un desempeño total subóptimo por falta de explotación a lo largo de los ensayos. Una de los algoritmos más conocidos para solucionar el dilema de exploración-explotación se le conoce como *epsilon-greedy*, donde epsilon (ϵ) determina la probabilidad con la que el agente explora aleatoriamente de entre sus opciones, epsilon usualmente comienza como una probabilidad alta pero que va disminuyendo poco a poco a lo largo de los ensayos, esto para que conforme pasa el tiempo el agente use cada vez más su política óptima aprendida para explotar, o para que sea "codicioso" (greedy) (Sutton & Barto, 2018).

$$A_t = \begin{cases} argmax Q(s_t, a) & \text{con probabilidad} = \epsilon \\ a \sim U(A) & \text{con probabilidad} = 1 - \epsilon \end{cases}$$

Otro algoritmo que busca solucionar este dilema es *Upper Confidence Bound* (UCB), este método en lugar de explorar aleatoriamente inicia explorando a partir de la incertidumbre asociada a las opciones disponibles, lo que evita el exceso de repeticiones y la sub-selección de opciones que puede implicar aleatorizar. Ya que UCB es la suma de la matriz de valor y la ponderación de la incertidumbre asociada a cada opción disponible, conforme el agente aprende, y el valor intrínseco de la función de valor aumenta por encima de la incertidumbre, el agente comienza a explotar más. Este algoritmo le asigna valor a cada acción en un determinado estado de acuerdo a la siguiente ecuación (Sutton & Barto, 2018)
:

$$A_t = argmax \left[Q(s_t, a) + c\sqrt{\frac{ln(t)}{N_t(a)}} \right]$$

Como se puede notar por la regla de elección, la incertidumbre de cada acción depende del número de veces que dicha acción ha sido elegida ($N_t(a)$) y del ensayo en el que se encuentre la simulación ($ln(t)$). Este resultado es ponderado por el parámetro de confianza (c) y finalmente se le adhiere el valor aprendido de la acción en cuestión ($Q(s_t, a)$), que es extraído de la matriz de valor, así es como UCB le asigna valor a cada acción y la selección de una depende de la que tenga mayor valor asociado.

3.3. Functional Experience Weighted Attraction

El Aprendizaje Basado en Creencias (*Belief Learning*; BL) el agente observa las acciones pasadas de su oponente y usa esta información para formar creencias sobre lo que cree que hará, es decir, determina la probabilidad con la que seleccionará cada una de sus acciones disponibles basado en su historial. Como en RL, el aprendizaje es un proceso de actualización iterativa, sin embargo, lo que se actualiza es el modelo del oponente, así anticipa su siguiente movimiento y determina la mejor respuesta para ello. Sin embargo, RL no forma creencias del rival debido a que solo hace un conteo de sus recompensas, mientras que BL no contempla sus éxitos pasados ya que se centra en hacer un conteo de las acciones del rival (Durlauf & Blume, 2009).

Las Dinámicas de la Mejor Respuesta de Cournot (1960) es uno de los modelos iniciales de BL, y consiste en que el agente asume que su oponente repetirá en la ronda actual la estrategia que usó en la ronda pasada. Por otra parte, Juego Ficticio (*Fictitious Play*; Brown, 1951) es uno de los modelos más estudiados de BL, la idea principal es que cada agente predice que la probabilidad de que su oponente elija una estrategia en el periodo actual depende de una suma ponderada para cada estrategia, de su probabilidad inicial y la frecuencia con la que ha sido elegida hasta el momento. En otras palabras, Cournot solo contempla la acción anterior y Juego Ficticio el historial de acciones (Durlauf & Blume, 2009).

Por otra parte, C. Camerer y Ho (1999) introdujeron Atracción Ponderada por Experiencia (*Experience-Weighted Attraction*; EWA), un modelo híbrido de aprendizaje que incluye RL y BL. La manera en la que EWA contempla ambos modelos es a través su matriz de valor aprendida, que refleja el reforzamiento de la estrategia que fue seleccionada y las ganancias hipotéticas que el agente pudo haber obtenido con sus otras estrategias. Así, la matriz de valores aprendidos para cada conjunto estado-acción se le llama matriz de Atracciones, y la probabilidad de elegir una estrategia depende de esta. Es importante resaltar que, en casos especiales, con uno de los parámetros de EWA (δ), el modelo puede dejar de ser híbrido y alternar entre RL ($\delta = 0$) y BL ($\delta = 1$), y en otro de los casos especiales del modelo, con de otro de los parámetros (ϕ), BL puede alternar entre Cournot ($\phi = 0$), Juego Ficticio ($\phi = 1$) y diferentes ponderaciones de Juego Ficticio. El modelo EWA fue extendido de varias maneras, como una versión de aprendizaje sofisticado que usa niveles de razonamiento para juegos secuenciales (C. F. Camerer et al., 2002), además, Realpe-Gómez et al. (2018) introdujo una versión con inspiración cognitiva creada específicamente para que el modelo

diera cuenta de la cooperación en juegos llamada *Experience Weighted Attraction with Norm Psychology* (EWAN).

Sin embargo, la versión que usamos en este trabajo corresponde a Atracción Ponderada por Experiencia Funcional (*Functional Experience Weighted Attraction*; FEWA), donde algunos de los parámetros pueden ser fijados a priori y otros se autoajustan, fue creado por Ho et al. (2007) con la intención de dar respuesta a las críticas de la cantidad de parámetros a estimar en la versión original de EWA, además, esta versión del modelo le da relevancia a la representación cognitiva de cada parámetro. Como la versión del modelo inicial, asigna atracción a cada una de las estrategias disponibles para el jugador, estas atracciones son usadas para determinar la probabilidad con la que un jugador elige cada estrategia. Las atracciones pueden tener valores a priori que representen la experiencia pre-juego por análisis, introspección o transferencia de aprendizaje.

La regla de actualización de atracciones de acuerdo a C. Camerer y Ho (1999) se ve de la siguiente manera:

$$A_i^j(t) = \frac{\phi \cdot N(t-1) \cdot A_i^j(t-1) + [\delta + (1-\delta) \cdot I(s_i^j, s_i(t)) \cdot \pi_i(s_i^j, s_{-i}(t))]}{N(t)}$$

La forma en la que EWA usa BL y RL es ponderando de forma contínua los resultados de las estrategias no seleccionadas con el parámetro δ, el cual representa la atención a las ganancias perdidas. Por otra parte, el autoajuste de este parámetro en FEWA se lleva a cabo de la siguiente manera (Ho et al., 2007):

$$\delta_{ij}(t) = \begin{cases} 1 & \text{if} \quad \pi_i(s_i^j, s_{-i}(t)) \geq \pi_i(t) \\ 0 & \text{otherwise} \end{cases}$$

De acuerdo a FEWA, el valor de δ depende de si la estrategia seleccionada es la mejor respuesta disponible, si el agente está seleccionando la mejor estrategia entonces δ=0 para no prestarle atención a otras opciones, si el agente no está eligiendo la mejor estrategia posible entonces δ=1 para forzarlo a acercarse probabilísticamente a la mejor respuesta disponible (Ho et al., 2007).

Otro parámetro que fue modificado en FEWA para que se autoajuste es ϕ, este parámetro puede desvanecer drásticamente el valor de las atracciones para "olvidar" cuando la regla de respuesta del rival ha cambiado. El parámetro ϕ es considerado una función de detección de cambio que decae la experiencia del jugador cuando se encuentra con un cambio sorpresivo. La manera en la que este ajusta su valor es al comparar un historial de frecuencias de las estrategias que usó el rival con un historial de unos y ceros de estrategias recientemente usadas de la siguiente manera (Ho et al., 2007):

$$S_i(t) = \sum_{k=1}^{m_{-i}} (h_i^k(t) - r_i^k(t))^2$$

El resultado es el índice de sorpresa ($S_i(t)$) y está comprendido como la suma de las diferencias cuadráticas entre el vector de historia acumulada ($h_i^k(t)$) y el vector de historia reciente ($r_i^k(t)$). Cuando el índice se acerca a cero quiere decir que el ambiente es constante, mientras que si se acerca a dos el ambiente es turbulento. Dado que el índice siempre está variando entre cero y dos, su valor debe ser dividido para que ϕ esté entre cero y uno. Debido a que $\phi = 1$ cuando mantiene el valor de las atracciones (memoria) y $\phi = 0$ cuando decrementa las atracciones (olvido), el valor del índice de sorpresa debe ser invertido para que el agente olvide el aprendizaje cuando la sorpresa sea alta de la siguiente manera (Ho et al., 2007):

$$\phi(t) = 1 - \frac{1}{2} S_i(t)$$

Por otra parte, FEWA contiene parámetros de EWA que no fueron modificados, por ejemplo N, que es derivado de los modelos de BL, y representa la experiencia de los jugadores, por lo que pondera las atracciones previas. Este parámetro, así como las atracciones (A), puede tener un valor a priori, el cuál puede ser interpretado como la importancia que se le da a la introspección pre-juego. El peso de la experiencia aumenta débilmente conforme aumentan los ensayos de la siguiente manera (Ho et al., 2007):

$$N(t) = \phi \cdot N(t-1) + 1$$

Finalmente, otro parámetro que no fue modificado en FEWA es el de sensibilidad (λ), que pondera las atracciones de la función softmax. Una función softmax es comúnmente usada para asignar probabilidad a las diferentes opciones dependiendo de su valor, y así un agente elige probabilísticamente de su conjunto de acciones disponibles. Este método es comúnmente usado en la literatura de toma de elección bajo incertidumbre, y se le conoce como *Regla de Luce (1959)*, donde la probabilidad de que se elija una opción es igual a su intensidad relativa, en la literatura de aprendizaje por refuerzo se usa en el algoritmo de *Bandido con Gradiente* para atribuirle probabilidad a preferencias (Sutton & Barto, 2018), en literatura de Teoría de Juegos se ha usado en modelos de *Quantal Response Equilibrium* para atribuirle probabilidad a cada estrategia dependiendo de su utilidad esperada (Carpenter & Robbett, 2022). La distribución softmax, también conocida como función logística en el caso particular de que la elección sea binaria, es adaptada de la siguiente manera en FEWA:

$$P_i^j(t+1) = \frac{e^{\lambda \cdot A_i^j(t)}}{\sum_{k=1}^{m_i} e^{\lambda \cdot A_i^j(t)}}$$

La función presente tiene dos características relevantes, la primera es el parámetro de sensibilidad (λ) que va de 0 a 1, entre más se acerque a cero las probabilidades se harán cada vez más uniformes, y entre más se acerque a 1 las probabilidades corresponderán cada vez más al valor de las opciones. La segunda característica es la transformación exponencial que le aplica a los valores de la matriz de atracciones ($A_i^j(t)$), la cual sirve para asignar probabilidades incluso a valores negativos, además amplifica la diferencia entre valores, lo cual puede ser contraproducente si los valores ya tienen mucha separación entre ellos. Además, el autor advierte que otras funciones de probabilidad pueden ser usadas para la elección, lo relevante es que la elección esté basada en las atracciones (Ho et al., 2007).

3.4. Teoría de juegos

La *Teoría de Juegos* y la *Economía Conductual* son dos áreas cercanas, ya que a ambas les interesa la elección racional en términos de *utilidad*, entendida como una medida de satisfacción que las personas le otorgan a algo. Además, ambas tienen un alcance enorme para estudiar procesos sociales tales como decisiones ambientales, políticas y económicas. Sin embargo, Economía Conductual tiene mayor enfoque en la influencia de la racionalidad limitada, el riesgo, el entorno, los sesgos y heurísticos sobre la elección individual (Kahneman & Tversky, 1979; Simon, 1955; Thaler & Sunstein, 2008). Mientras que Teoría de Juegos (von Neumann & Morgenstern, 1944) puede entenderse como una metodología para formalizar y analizar toma de decisiones estratégicas que involucre a más de un agente, es decir, aquellas situaciones en las que la elección de un agente se ve afectada por la elección del otro.

Una interacción social es estudiada a través de un juego, es decir, de una representación simplificada del fenómeno que se busca estudiar, en este se define la cantidad de jugadores, las estrategias disponibles por jugador (π_i^j) y la ganancia asociada a cada conjunto de acciones seleccionadas por los jugadores (p_j), también debe declararse cuántas rondas tendrá el juego y cuánta información tiene cada jugador sobre el juego y su rival o rivales (Carpenter & Robbett, 2022).

	Jugador 1	
	Estrategia 1 (π_1^1)	Estrategia 2 (π_2^1)
Jugador 2		
Estrategia 1 (π_1^2)	(p_1, p_2)	(p_1, p_2)
Estrategia 2 (π_2^2)	(p_1, p_2)	(p_1, p_2)

Figura 2: Representación de un juego simultáneo como *Matriz de Ganancias*

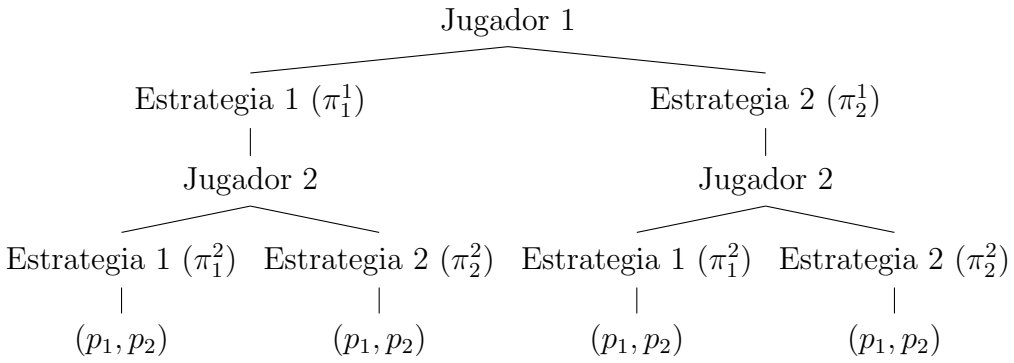

Figura 3: Representación de un juego secuencial como *Árbol de Decisión*

Comúnmente la solución de un juego se le conoce como *Equilibrio de Nash* (J. F. Nash, 1950), la noción de equilibrio en elecciones simultáneas implica que, ambos jugadores ya están usando su mejor estrategia posible dadas las ganancias y el conjunto de estrategias del propio jugador y del rival, y cualquier desviación unilateral solo les perjudica, por lo que no tienen razones para cambiar sus estrategias. Por su parte, en elecciones secuenciales se llega a un *Equilibrio de Nash Perfecto en Subjuego*, que es cuando ambos jugadores usan su mejor estrategia en cada ronda de elección. En una negociación, el equilibrio significa que las ganancias de ambas partes constituyen la utilidad total que se puede obtener, considerando las preferencias de cada uno (Carpenter & Robbett, 2022).

3.5. Juego de Negociación

El juego de negociación representa cualquier situación en la que dos jugadores o más deben acordar cómo repartirse algo que aporta utilidad a todas las partes involucradas, y si no son capaces de acordar una partición ninguna parte recibirá utilidad (Carpenter & Robbett, 2022).

Nash describió en qué consistía una negociación y propuso dos soluciones. La primera solución consiste en axiomas (J. F. J. Nash, 1950) y consiste en que la solución a la que llegan agentes racionales en cualquier problema de negociación requiere de los principios de invarianza a la escala, simetría, pareto eficiencia e independencia de alternativas irrelevantes. Mientras que la segunda solución (J. Nash, 1953) es estratégica, y se extiende a negociaciones no cooperativas, esta especifica el comportamiento de los agentes involucrados en una negociación cuando se toma en cuenta el tiempo, información, herramientas de compromiso, opiniones externas y amenazas.

Más tarde, Rubinstein (1982) llegaría a la solución de una negociación sin límite de rondas, el equilibrio en esta situación es que el jugador ofrezca desde el inicio lo que él crea que su rival le aceptaría, ya que hacer la misma oferta más a futuro solo haría que la utilidad de las ganancias sean descontadas temporalmente. Ese mismo año, Güth et al. (1982) introducirían el juego del ultimátum, que es una negociación de una sola ronda, el equilibrio consiste en ofrecer la menor cantidad posible del presupuesto, y el rival debe aceptar cualquier oferta que sea mayor a cero, o mantenerse indiferente a aceptar o rechazar si le ofrecen cero. Sin embargo, la relevancia de este juego radica en que los jugadores deben pensar en cuánto ofrecer para no ser rechazados y recibir una recompensa.

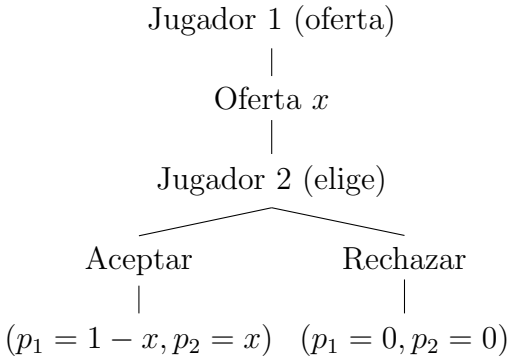

Figura 4: Árbol de decisión del Juego del Ultimátum

Por supuesto, una versión común del juego de negociación implica un límite de rondas para contraofertar, es decir, el jugador debe ofertar en la primera ronda, y si el rival declina la oferta del jugador, ahora el rival puede hacer una contraoferta en la siguiente ronda, y ahora el ciclo de ofertas se repite con el jugador teniendo que decidir si acepta o declina la contraoferta del rival, si no se llega a un acuerdo una vez que se termine el número de rondas disponibles para ofertar ningún jugador recibe nada del presupuesto total. Por lo descrito, este equilibrio también indica que el rival acepte cualquier oferta mayor a cero que le haga el jugador desde la primera ronda, sin embargo, en esta versión el jugador también debe pensar en cuánto aceptaría el rival para recibir una recompensa, y tiene un número de oportunidades equivalente al número de rondas para revelar las preferencias del rival (Carpenter & Robbett, 2022).

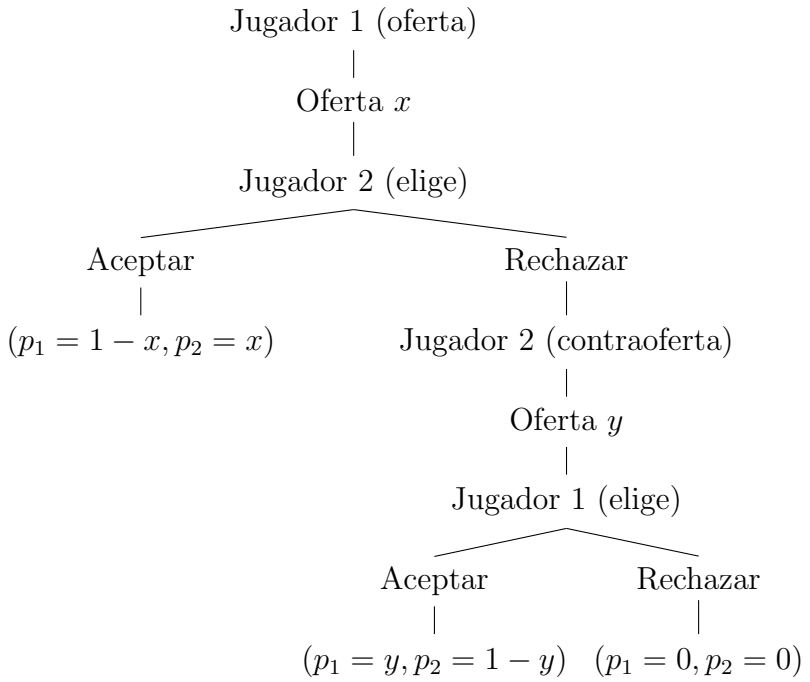

Figura 5: Árbol de decisión de un juego de negociación con una ronda de contraoferta

3.6. Aversión a la Inequidad

La respuesta que da Teoría de Juegos a elecciones de jugadores que se desvían de las soluciones de los juegos, es que los jugadores llevan consigo preferencias que transforman la utilidad psicológica a las ganancias (Carpenter & Robbett, 2022). Por lo anterior, el modelo de Aversión a la Inequidad (*Inequity Aversion*; IA) de Fehr y Schmidt (1999) considera que hay personas que les importa sus ganancias y las ganancias de otros jugadores, ya que una recompensa es percibida diferente dependiendo del contexto en el que se obtuvo, y por ello se desvían de las soluciones de los juegos derivadas lógica o matemáticamente.

El modelo de Aversión a la Inequidad es conocido por explicar conducta justa, competitiva o cooperativa al definir las preferencias de los jugadores una vez que se calculan sus parámetros. El modelo solo contiene dos parámetros que ponderan la utilidad total de dos sujetos en un juego, donde α puede ser interpretado estratégicamente como ventaja y β como desventaja, o psicológicamente α podría interpretarse como egoísmo y β como culpa (C. F. Camerer, 2011). Dependiendo de las ganancias finales de ambos jugadores se calculan los valores de los parámetros del jugador para determinar si tiene aversión a la ventaja, aversión a la desventaja o aversión a la inequidad como se muestra a continuación Fehr y Schmidt (1999):

$$u_i(x_i, x_j) = x_i - \alpha_i max\left\{x_j - x_i, 0\right\} - \beta_i max\left\{x_i - x_j, 0\right\}$$

Como se puede leer por la ecuación, gracias a los operadores se usa solo uno de los dos parámetros, dependiendo de si las ganancias finales le otorgan al jugador ventaja se usa α y si el jugador está en desventaja se usa β, si el jugador no está en ventaja o desventaja ambos parámetros son igual a cero.

3.7. Agent Based

Los modelos Basados en Agentes (*Agent Based*; AB) son modelos donde cada individuo es representado como un agente que puede actuar localmente en una población. von Neumann (1966) estableció los principios teóricos de este campo, al proponer autómatas autorreproductores que crean sistemas complejos a partir del comportamiento colectivo, más tarde Schelling (1971) con su trabajo sobre segregación y dinámicas sociales, demostró cómo surgen patrones complejos a partir de reglas simples, y Axelrod y Hamilton (1981) usaron estas ideas para aplicarlas a teoría de juegos y demostrar la emergencia de cooperación a través de la estrategia *Tit for Tat*. Actualmente los sistemas AB son una clase de modelos computacionales que pueden proveer representaciones de segregación, contagio de enfermedades u opiniones, influencia social hacia una ideología u opinión, evolución biológica o de comportamiento social, y redes de conexiones biológicas o sociales (Smaldino, 2023).

Estos modelos dan cuenta, principalmente, de cómo pueden emerger resultados complejos bajo la idea de que si cada individuo en una población usa comportamientos simples para responder, como heurísticos por ejemplo, entonces pueden emerger patrones complejos en dicha población. Una simulación de este tipo requiere de reglas y dinámicas preestablecidas, las reglas corresponden a qué hace cada agente, mientras que las dinámicas son simplificaciones de los elementos del fenómeno que se quiere medir, por ejemplo, movimiento, contagio, reproducción, orden de ejecución de reglas, etc. El cómo se decide que un agente ejecute su regla puede deberse a un cálculo o puede ser a través de instrucciones lógicas (Smaldino, 2023).

Comúnmente los agentes son situados en un plano bidimensional donde los agentes pueden permanecer en su posición inicial o pueden moverse aleatoriamente en el entorno. El

espacio bidimensional suele tener límites periódicos, es decir, si un agente sobrepasa el límite izquierdo saldría por el límite derecho, si sobrepasa el límite de arriba saldría por el límite de abajo, y viceversa (Smaldino, 2023).

Los modelos AB suelen capturar sistemas heterogéneos que se pueden encontrar en el mundo real, pero al precio de estocacidad e incremento de incertidumbre. Se puede implementar estocacidad en la inicialización de la población y en cómo se ejecutan las dinámicas de la tarea. Idealmente se espera que las variaciones sean igual a la cantidad de inicializaciones y de dinámicas consideradas, sin embargo, si existen demasiadas combinaciones, se espera que las variaciones sean representativas, es decir, que cada variación capturen los efectos de un conjunto de combinaciones y que, a su vez, le permita diferenciarse de variaciones con rangos distintos de combinaciones, esto con la intención de encontrar y explicar los resultados más probables dadas las variaciones de las simulaciones (Smaldino, 2023).

Una simulación Agent Based puede terminar cuando se llega a un equilibrio, a un punto estable, donde cada agente ya no tiene la necesidad de aplicar su regla individual y por lo tanto se queda estático. En otros casos puede que el sistema nunca llegue a un equilibrio, por lo que las simulación se termina después de que la simulación haya llegado a un patrón reconocible, por ejemplo, que cada agente está aplicando la misma conducta cada vez que ejecuta su regla individual y por lo tanto se queda constante (Smaldino, 2023).

4. Métodos

4.1. Consideraciones previas

El trabajo presente es una investigación deductiva, ya que partimos de teorías para analizar las conclusiones de sus modelos. No obstante, es importante destacar que los resultados de este trabajo son generativos, ya que la verificación empírica fue a través de datos generados en simulaciones. Estos datos se obtuvieron al simular de manera iterativa los resultados de los modelos matemáticos y lógicos asociados a las teorías de interés, dentro de un entorno virtual controlado.

Las investigaciones que usan modelos matemáticos de conducta o procesos cognitivos para simular resultados tienen el objetivo de evaluar la consistencia interna de los modelos, al analizar sus predicciones bajo diversas variaciones en los parámetros; explorar sus resultados bajo diferentes condiciones o supuestos, para explicar qué variaciones provocan los diferentes resultados; y comparar las predicciones de diferentes modelos teóricos. Además, estas investigaciones también tiene la intención de probar modificaciones en los modelos para que se adapten a entornos virtuales que representan simplificaciones de la realidad.

4.2. Preguntas de investigación

4.2.1. Reinforcement Learning Q-Learning

1. ¿Cómo afectan los parámetros del modelo de Aversión a la Inequidad a las estrategias de negociación óptimas aprendidas por un agente RL Q-Learning?

    a) El agente RL Q-Learning aprenderá a hacer ofertas más equitativas a medida que los parámetros de aversión a la inequidad del rival aumentan.

2. ¿La política óptima aprendida por el agente RL Q-learning se acerca o se aleja del Equilibrio Perfecto en Subjuegos de Nash?

    a) El Equilibrio Perfecto en Subjuegos de Nash (SPNE) será diferente para cada rival, y la política óptima aprendida por el agente será la misma que el mismo SPNE que aplique para ese rival.

4.2.2. Functional Experience Weighted Attraction

1. ¿Cómo afectan los parámetros del modelo de Aversión a la Inequidad a las estrategias de negociación óptimas aprendidas por un agente FEWA?

    a) El agente FEWA aprenderá a hacer ofertas más equitativas a medida que los parámetros de aversión a la inequidad del rival aumentan.

2. ¿La política óptima aprendida por el agente FEWA se acerca o se aleja del Equilibrio Perfecto en Subjuegos de Nash?

    a) El Equilibrio Perfecto en Subjuegos de Nash (SPNE) será diferente para cada rival, y la política óptima aprendida por el agente será la misma que el mismo SPNE que aplique para ese rival.

4.2.3. Reinforcement Learning Q-Learning y Functional Experience Weighted Attraction

1. ¿En qué se diferenciarán las estrategias óptimas aprendidas por los modelos?

    a) Ambos modelos aprenderán la política óptima que sea igual al SPNE para cada rival.

2. ¿En qué se diferenciará el desempeño de los modelos?

    a) Ambos modelos obtendrán resultados similares debido a que usan la misma política de exploración y explotación.

3. ¿En qué se diferenciará el aprendizaje de los modelos?

    a) El valor de las matrices aprendidas por los agentes serán diferentes, la matriz del agente FEWA tendrá valores más altos en general y la diferencia proporcional en los valores de las estrategias será más notoria en la matriz del agente FEWA.

4. ¿Ambos agentes serán capaces de clasificar correctamente a un rival con un rango de ofertas aceptables conocido y de clasificar correctamente a un rival con un rango de ofertas aceptables desconocido?

    a) Tanto RL como FEWA serán capaces de clasificar correctamente rivales con un rango de ofertas aceptables que sea diferenciable y que hayan aprendido previamente. Sin embargo, solo el agente FEWA será capaz de identificar y aprender rivales con un rango de ofertas aceptables desconocido y diferenciable, debido a su parámetro de detección de cambio.

4.2.4. Agent Based

1. ¿En qué se diferenciará el desempeño de los modelos de aprendizaje en el sistema AB si en la población solo hay rivales aprendidos previamente?

    a) Entre más similares sean las estrategias óptimas aprendidas por los agentes RL y FEWA serán más similares las ganancias finales de ambos agentes y viceversa.

2. ¿En qué se diferenciará el desempeño de los modelos si en la población se incluye a un rival que no fue aprendido previamente por los modelos, que ofrece más ganancias y tiene un rango de ofertas aceptables diferenciable?

    a) FEWA será el único capaz de reconocer a los rivales que no fueron aprendidos previamente y, debido a que este nuevo rival ofrece más ganancias, FEWA obtendrá más ganancias que RL al terminar la simulación.

3. ¿En qué medida las variaciones en la distribución de rivales resultan en segregación?

    a) Entre mayor sea la probabilidad de aparición de agentes con preferencia alta a resultados ventajosos y de agentes con preferencia por resultados eficientes en la distribución de población habrá más desacuerdos y, por lo tanto, más movimineto, lo que se traduce en menos formación de grupos y menos segregación.

4. ¿En qué medida las variaciones en el tamaño de la población resultan en segregación?

    a) Sin importar el tamaño de la muestra, si la probabilidad de aparición de agentes con preferencia alta a resultados ventajosos y de agentes con preferencia por resultados eficientes es alta en la distribución de población seguirá habiendo desacuerdos, y por lo tanto, menos segregación.

5. ¿Algún estado del modelo provoca el surgimiento de un patrón reconocible o equilibrio?

    a) En la simulación los agentes siempre deben ejecutar su regla individual, por lo que nunca se llegará a ningún punto estático. Sin embargo, se espera llegar a patrones constantes y reconocibles en los movimientos y en las ganancias, las cuales dependerán, una vez más, de la probabilidad de aparición de agentes con preferencia alta a resultados ventajosos y de agentes con preferencia por resultados eficientes, si su probabilidad es alta, habrá más constancia en las pérdidas de la población y por lo tanto en los movimientos, si su probabilidad es baja habrá más ganancias y menos movimientos, esto sin importar el tamaño de la muestra.

4.3. Variables

Tipo de agente: Se probaron dos tipos de agentes que aprenden con diferentes métodos de aprendizaje,

1. Reinforcement Learning Q-learning.

2. Functional Experience Weighted Attraction.

Preferencias de rivales: Se probaron cuatro tipos de agentes que no aprenden y que responden de acuerdo al modelo de Aversión a la Inequidad, cada uno con variaciones en los valores de los parámetros,

1. Agente con preferencia baja a resultados ventajosos: ponderación baja en alpha, α=0.1 y β=0, representando situaciones donde un agente tiene únicamente poca aversión a la desventaja,

2. Agente con preferencia media a resultados justos: ponderación media en ambos parámetros, α=2 y β=3, representando situaciones donde un agente tiene aversión a la inequidad,

3. Agente con preferencia alta a resultados ventajosos: ponderación alta en alpha, α=25 y β=0, representando situaciones donde un agente tiene únicamente mucha aversión a la desventaja,

4. Agente con preferencia por resultados eficientes: ponderación nula en ambos parámetros, α=0 y β=0, representando situaciones donde un agente sin aversión a la inequidad y que se le considera eficiente por ofrecer la cantidad mínima posible y aceptar cualquier oferta mayor a cero.

Tamaño de la muestra: Se probaron tres cantidades diferentes de rivales en la población,

1. Una población pequeña de 30 agentes,

2. Una población mediana de 150 agentes,

3. Una población grande de 300 agentes.

Proporción de rivales con diferentes preferencias en la muestra: se probaron siete proporciones diferentes,

1. Distribución de rivales conocidos para RL y FEWA donde todos tienen una proporción uniforme.

2. Distribución de rivales conocidos para RL y FEWA donde los rivales con preferencia baja a resultados ventajosos tienen mayoría.

3. Distribución de rivales conocidos para RL y FEWA donde los rivales con preferencia media a resultados justos tienen mayoría.

4. Distribución de rivales conocidos para RL y FEWA donde los rivales con preferencia alta a resultados ventajosos tienen mayoría.

5. Distribución de rivales conocidos para RL y FEWA y un rival desconocido donde todos tienen una proporción uniforme.

6. Distribución de rivales conocidos para RL y FEWA donde el rival desconocido tiene minoría.

7. Distribución de rivales conocidos para RL y FEWA donde el rival desconocido tiene mayoría.

4.4. Diseño

El experimento tiene el diseño mixto de 2 (agentes: RL vs FEWA) $\times$ 4 (preferencias: baja vs media vs alta vs eficiente) $\times$ 3 (tamaños: chica vs media vs grande) $\times$ 7 (proporciones: uniforme de tres vs mayoría de preferencia baja vs mayoría de preferencia media vs mayoría de preferencia alta vs uniforme de cuatro vs minoría de preferencia eficiente vs mayoría de preferencia eficiente)

4.5. Muestra

La muestra de la **fase de aprendizaje** no es probabilística, fue seleccionada para comparar dos modelos de aprendizaje en escenarios específicos del modelo de Aversión a la Inequidad. La muestra consiste en dos agentes no sesgados que aprenden bajo los modelos de Reinforcement Learning Q-learning y Functional Experience Weighted Attraction. Cada uno de los anteriores agentes juega contra tres rivales en orden, sus rivales son agentes que no aprenden y que responden de acuerdo al modelo de Aversión a la Inequidad, el primer rival es el agente con preferencia baja a resultados ventajosos, el segundo rival es el agente con preferencia media a resultados justos, y el tercer rival es el agente con preferencia alta a resultados ventajosos.

La muestra de la **fase de prueba de tamaños**, consiste en poblaciones probabilísticas de tres diferentes tamaños, primero de 30 agentes que no aprenden y que responden de acuerdo al modelo de Aversión a la Inequidad, luego 150 y finalmente de 300. A cada una de estas poblaciones se les agregó un agente RL Q-Learning y un agente FEWA que aprendió previemente de los rivales mencionados en la fase de aprendizaje.

La muestra en la **fase de prueba de distribuciones**, consiste en una población probabilística de 7 distribuciones diferentes:

1. Distribución de rivales conocidos para RL y FEWA donde todos tienen una proporción uniforme:
   a) Consta de 0.33 de agentes con preferencia baja a resultados ventajosos,
   b) 0.33 de agentes con preferencia media a resultados justos y
   c) 0.33 de agentes con preferencia alta a resultados ventajosos.
2. Distribución de rivales conocidos para RL y FEWA donde los rivales con preferencia baja a resultados ventajosos tienen mayoría:
   a) Consta de 0.5 de agentes con preferencia baja a resultados ventajosos,
   b) 0.25 de agentes con preferencia media a resultados justos y
   c) 0.25 de agentes con preferencia alta a resultados ventajosos.
3. Distribución de rivales conocidos para RL y FEWA donde los rivales con preferencia media a resultados justos tienen mayoría:
   a) Consta de 0.25 de agentes con preferencia baja a resultados ventajosos,
   b) 0.5 de agentes con preferencia media a resultados justos y
   c) 0.25 de agentes con preferencia alta a resultados ventajosos.
4. Distribución de rivales conocidos para RL y FEWA donde los rivales con preferencia alta a resultados ventajosos tienen mayoría:
   a) Consta de 0.25 de agentes con preferencia baja a resultados ventajosos,
   b) 0.25 de agentes con preferencia media a resultados justos y
   c) 0.5 de agentes con preferencia alta a resultados ventajosos.
5. Distribución de rivales conocidos para RL y FEWA y un rival desconocido donde todos tienen una proporción uniforme:

a) Consta de 0.25 de agentes con preferencia baja a resultados ventajosos,

b) 0.25 de agentes con preferencia media a resultados justos,

c) 0.25 de agentes con preferencia alta a resultados ventajosos,

d) 0.25 de agentes con preferencia a resultados eficientes.

6. Distribución de rivales conocidos para RL y FEWA donde el rival desconocido tiene minoría:

   a) Consta de 0.3 de agentes con preferencia baja a resultados ventajosos,

   b) 0.3 de agentes con preferencia media a resultados justos,

   c) 0.3 de agentes con preferencia alta a resultados ventajosos,

   d) 0.1 de agentes con preferencia a resultados eficientes.

7. Distribución de rivales conocidos para RL y FEWA donde el rival desconocido tiene mayoría:

   a) Consta de 0.2 de agentes con preferencia baja a resultados ventajosos,

   b) 0.2 de agentes con preferencia media a resultados justos,

   c) 0.2 de agentes con preferencia alta a resultados ventajosos,

   d) 0.4 de agentes con preferencia a resultados eficientes.

A cada una de estas poblaciones se les agregó un agente RL Q-Learning y un agente FEWA que aprendió previemente de los rivales mencionados en la fase de aprendizaje.

4.6. Contexto

La programación de las simulaciones del experimento se llevaron a cabo en el **Laboratorio de Conducta Adaptable**, Facultad de Psicología, UNAM. El mobiliario en el que se llevó a cabo es un cubículo con escritorio, silla y computadora con conexión a internet.

4.7. Procedimiento

4.7.1. Fase de aprendizaje

Experimento simulado de dos grupos. En el primer grupo el agente RL juega contra el rival con preferencia baja a resultados ventajosos, luego contra el rival con preferencia media

a resultados justos y finalmente contra el rival con preferencia alta a resultados ventajosos. Contra cada rival se simularon 2500 ensayos, cada ensayo simula un juego de negociación entre el agente RL y el rival contra el que está compitiendo. En el segundo grupo simplemente se cambia al agente RL por el agente FEWA. Para ambos se toman medidas repetidas de las estrategias seleccionadas, del valor que se le asigna a las estrategias y de las ganancias para cada agente a lo largo del proceso de aprendizaje.

Tarea

El escenario de la fase de aprendizaje es un ambiente simulado programado para emular una tarea de negociación de máximo tres rondas en donde se debe dividir un presupuesto de 100 puntos. Las acciones son ofertar cualquier número entero al rival, aceptar o declinar la oferta del rival y contraofertar.

El agente RL o FEWA comienza ofertando una partición del presupuesto total, si la oferta es rechazada le toca al rival contraofertar, si esta contraoferta es rechazada le toca ofertar una última vez al agente RL o FEWA, si esta es rechazada el ensayo de la negociación termina y ninguno obtiene ganancias, si en alguna de las tres rondas la oferta es aceptada el ensayo de la negociación termina con la división de ganancias acordada.

Selección de estrategia de rivales conforme a modelo de Aversión a la Inequidad (IA)

Los tres diferentes rivales a los que se enfrentan RL y FEWA responden de acuerdo al modelo de IA. El primer rival tiene valores en los parámetros α=0.1, β=0, el segundo tiene valores α=2, β=3 y el tercero tiene valores α=25, β=0. Los rivales solo aceptan ofertas que tengan una utilidad mayor a cero $u > 0$. La intención es que el primero acepte casi cualquier oferta del conjunto, el segundo acepta solo ofertas cercanas a la mitad del conjunto de estrategias y el último solo acepta ofertas que estén por encima de la mitad del conjunto de ofertas.

Si la oferta del agente RL o FEWA no es aceptada, ahora es turno del rival contra el que estén jugando de hacer una oferta de su conjunto de ofertas con utilidad positiva. El primero le asigna probabilidad uniforme a su conjunto de ofertas ($a \sim \mathcal{U}$ para a con $u(a) > 0$), el segundo le asigna una probabilidad normal a su conjunto de ofertas con media en la mitad del conjunto de ofertas ($a \sim \mathcal{N}(\mu, \sigma^2)$ para a con $u(a) > 0$), y el tercero le asigna probabilidad exponencial a su conjunto de ofertas ($a \sim \text{Exp}(\lambda)$ para a con $u(a) > 0$), siendo las más bajas para el agente RL o FEWA a las que les asigna más probabilidad. La intención

es que el primero ofrezca aleatoriamente, que el segundo ofrezca cercano a lo equitativo para ambos y que el tercero ofrezca en su mayoría muy favorecedor para él y desfavorecedor para el agente RL o FEWA. Si la contraoferta del rival fue rechazada por el agente RL o FEWA llegan a la tercera y última ronda, en esta los agentes basados en modelos de aprendizaje deben hacer su última oferta, si es rechazada ambos terminan esa simulación del juego con 0 de recompensa.

Selección de estrategia de los agentes RL y FEWA

Ambos agentes fueron programados para ejecutar Iteración de Política Generalizada en orden de aprender la política óptima. Asimismo, ambos siguen el mismo método off-policy para resolver el dilema exploración-explotación. En un inicio, sus métodos predeterminados arrojaban estrategias subóptimas, para RL, epsilon-greedy no terminaba de probar todas sus opciones apropiadamente por la aleatorización, y para FEWA, la función softmax se estancaba frecuentemente en políticas poco plausibles, consideramos que ambos problemas se debían a la cantidad de acciones disponibles en la tarea.

Después probamos Upper Confidence Bound (UCB), sin embargo, aunque UCB parecía prometedor también mostraba errores. La regla de elección completa de UCB incluye añadir el valor del conjunto estado-acción a la incertidumbre y elegir la opción con mayor valor.

Por lo anterior, si una opción adquiere un valor por encima de la incertidumbre temprano en los ensayos, dicha opción comenzará a ser elegida constantemente corriendo el riesgo de no volver a salir de una opción subóptima, este riesgo es particularmente notorio cuando la matriz de valor es acumulativa como en FEWA. Por otra parte, si se intenta mitigar ese problema aumentando considerablemente el valor del parámetro de confianza para que el agente explore durante más tiempo, el agente seguirá explorando constantemente incluso en ensayos avanzados, ya que los valores de la función de valor se verían opacados por la incertidumbre, este riesgo es particularmente notorio cuando la matriz de valor no puede adquirir valores tan altos por límites como el error de predicción de RL.

Por los problemas presentados se optó por usar los elementos que funcionaban de los anteriores métodos para el algoritmo de exploración-explotación de ambos agentes. La forma en la que ambos agentes deciden si exploran o explotan es, esencialmente, con epsilon-greedy. El uso inicial de epsilon-greedy tiene la intención de evitar cambios súbitos, ya que el agente empieza a explotar gradualmente después de un largo periodo de exploración.

No obstante, si epsilon determina que el agente debe explorar, la elección de una acción no es completamente aleatoria, sino que primero le asigna valor a cada opción disponible de acuerdo al método de incertidumbre de UCB, pero sin agregar el valor de los conjuntos estado-acción. La regla de incertidumbre de UCB consiste en que si una estrategia a es seleccionada su incertidumbre disminuye y, por consiguiente, la incertidumbre de las estrategias que no fueron seleccionadas en ese ensayo aumentan. Así, cuando el agente explora, la probabilidad con la que se selecciona una acción depende únicamente del número de veces que esa acción no ha sido seleccionada a lo largo de los ensayos.

Además, la regla de elección de UCB también indica que se seleccione aquella opción con mayor valor, pero en lugar de eso ahora se selecciona una acción de acuerdo a la probabilidad asignada con una función softmax que depende de la incertidumbre de cada acción:

$$a(t) = \left\{ \begin{array}{cc} P_a(t+1) = \frac{e^{a(t)}}{\sum e^{a(t)}} & \text{con probabilidad} = \epsilon \\ a(t) = c\sqrt{\frac{ln(t)}{N_t(a)}} & \text{con probabilidad} = 1 - \epsilon \end{array} \right\}$$

Se usó únicamente la incertidumbre de UCB durante la exploración porque se buscaba asignarle más probabilidad a las opciones que no han sido seleccionadas, y así evitar que el agente infravalore algunas opciones por haber sobre-seleccionado otras. Gracias a este cambio ya no se tiene que aumentar considerablemente el número de simulaciones para que el agente pruebe más uniformemente todas las opciones disponibles.

El uso de la función softmax durante la explotación como sustituto del operador de maximización de UCB y de epsilon-greedy tiene la intención principal de agregar variabilidad de selección a estrategias adyacentes en valor a la mejor estrategia disponible, esto para que los agentes prueben estrategias con potencial, es decir, estrategias que posiblemente todavía no han adquirido todo su valor. Además, cuando el aprendizaje de ambos agentes está avanzado el efecto de la incertidumbre de UCB termina por volverse minúsculo.

Finalmente, si no es turno del agente RL o FEWA de elegir una oferta, pero sí de aceptar o rechazar una oferta, se usa únicamente el método de explotación descrito anteriormente para formular una estrategia que se espera del rival. Inicialmente el agente RL o FEWA debe elegir una estrategia óptima por el método de explotación, después, al presupuesto total (x)

le sustrae la estrategia óptima seleccionada y el resultado es la estrategia que espera del rival ($E(t) = x - a(t)$). De acuerdo a lo anterior, la regla para aceptar o rechazar consiste en que si el rival le hace una oferta peor que su oferta esperada ($oferta < E$) entonces el agente la rechaza, si el rival le hace una oferta igual o mejor que la oferta esperada ($oferta \geq E$) entonces el agente acepta la oferta del rival.

Actualización de Matriz de Valor Q y Matriz de Atracciones

Por la descripción de la tarea, tanto el agente RL como el agente FEWA, empiezan con una matriz de matrices que contiene un número de matrices equivalente al número de rivales a los que se enfrentará (r), la matriz de un solo rival contiene una cantidad de vectores equivalente a la cantidad de rondas del juego (s) y cada vector contiene la secuencia de elementos equivalente al número de particiones del presupuesto que el agente puede ofrecer (a), por lo que la matriz de RL tiene la froma $Q \in \mathbb{R}^{r \times s \times a}$ y la de FEWA tiene la forma $A \in \mathbb{R}^{r \times s \times a}$.

Si la estrategia que fue elegida por el agente RL o FEWA fue aceptada por el rival, entonces el agente recibe una recompensa equivalente a su parte de la ganancia ($R(t) = x - a(t)$), esta se usa para actualizar el valor de ese conjunto estado-acción y ahí terminaría la simulación de ese juego. Si la primera oferta fue rechazada pero el agente RL o FEWA acepta la contraoferta del rival, el agente usa la recompensa obtenida para actualizar la estrategia correspondiente de la partición aceptada ($R(t) =$ oferta del rival). Finalmente, si la contraoferta de la segunda ronda es rechazada, porque el agente está explorando o porque la oferta no cumplió con las expectativas del agente, entonces RL o FEWA hace una última oferta, sin importar el resultado la acción de este estado siempre es actualizada, sin embargo, si la oferta es aceptada se usa la recompensa como actualización de la estrategia seleccionada y, si la oferta es rechazada, la estrategia es actualizada con 0 de recompensa.

En el caso de RL, la acción seleccionada en el estado en el que se encuentra es actualizada de acuerdo a la predicción de error, que es determinada por la recompensa recibida y a la acción con mayor valor del siguiente estado, como se muestra en su regla de actualización. En el caso de la regla de FEWA, la acción seleccionada es actualizada de acuerdo a la acumulación de la recompensa recibida sin ponderar, y todas las otras acciones no seleccionadas de ese estado son actualizadas con la recompensa que se espera de acuerdo a las creencias que se tiene del rival, ponderadas por un parámetro de atención.

Parámetros de FEWA

Como se mencionó en el marco teórico, FEWA tiene 4 parámetros y dos de ellos se autoajustan. El primer parámetro que se autoajusta es δ, el cual representa la atención a las ganancias perdidas de estrategias no seleccionadas y es el que nivela el valor de los resultados de RL y BL.

Originalmente, el δ vale 0 si el agente está seleccionando la mejor estrategia disponible para no prestarle atención a otras opciones, y vale 1 si el agente no está eligiendo la mejor estrategia posible para forzarlo a acercarse a estrategias con mayores ganancias asociadas.

Sin embargo, este método tiene un problema en las simulaciones que corrimos, el principal fue que, debido a que el rival acepta, rechaza y ofrece de acuerdo a los valores de sus parámetros del modelo de aversión a la inequidad, ninguno de los tres rivales le acepta al agente FEWA su mejor estrategia posible, es decir, una partición del 100 para el agente y 0 para el rival. Por lo tanto, δ siempre es igual a 1, FEWA siempre usa BL, y por ello la mejor estrategia posible, aunque es inviable, es la que sigue adquiriendo más valor a lo largo de los ensayos por tener las mayores ganancias perdidas.

Por esto se implementaron dos cambios, el primero fue hacer que las ganancias perdidas de cada estrategia fueran igual a la *ganancia esperada* ($\mathbb{E}u_i = \sum p_i \cdot u_i$) de cada estrategia. De esta manera, el valor de cada estrategia del agente FEWA es igual a las ganancias que obtendría si el rival acepta dicha oferta por la frecuencia con la que el rival ha aceptado dicha oferta, más las ganancias que obtendría si el rival rechaza la oferta por la frecuencia con la que el rival ha rechazado la oferta en cuestión.

De esta manera, estrategias que nunca han sido aceptadas ya no son consideradas como la mejor estrategia que el agente FEWA podría jugar. Sin embargo, el hecho de que el algoritmo use $\delta = 1$ o $\delta = 0$, quiere decir que FEWA usa BL o usa RL, solo usa uno u otro, cuando creemos que el punto central de FEWA es que sea un híbrido de ambos, es decir, que de cierta manera use ambos métodos de aprendizaje. Por lo anterior, la regla de autoajuste del parámetro δ fue cambiada para que empiece con un valor de 1, de esta manera el agente empieza otorgándole el mismo nivel de atención a todas sus estrategias, y conforme pasan los ensayos y se actualizan las creencias de las estrategias del rival, el valor de δ decae lentamente hasta un límite de 0,5, así, la atención que le asigna a estrategias no seleccionadas decae conforme el agente empieza a explotar.

El segundo parámetro que se autoajusta es ϕ, y representa el olvido de lo aprendido ante la sorpresa, ya que este parámetro es usado como un detector de cambio. La funcionalidad del parámetro es igual a la planteada originalmente por los autores durante la fase de aprendizaje, es decir, compara un historial de frecuencias de las estrategias que usó el rival con un historial de unos y ceros de estrategias recientemente usadas.

Sin embargo, ahora también se usa ϕ para cuando el agente debe categorizar a un nuevo rival, de esta manera, se aprovecha la funcionalidad del parámetro autoajustable y la base teórica que conlleva. Además, debido a que cuando el agente se está enfrentando a un rival que siempre responde de acuerdo al modelo de IA, sus estrategias no cambian en lo absoluto, por lo mismo, el parámetro no tiene un uso significativo en la fase de aprendizaje. Esta nueva función consiste en agregar un umbral en el valor mínimo que puede adquirir este parámetro ($\phi > 0{,}2$). Así, cuando FEWA compara las estrategias de un rival desconocido con las de los rivales que FEWA se ha enfrentado anteriormente, y ninguno de los rivales aprendidos sobrepasa este umbral, se asume que se está enfrentando a un nuevo rival y, en lugar de olvidar el aprendizaje previo, se abre una nueva matriz para aprender de él. Esta implementación favorece la detección y adaptación a nuevos rivales, a diferencia de RL, además, la función original y explicación teórica del parámetro se mantienen. La adaptación de este parámetro es particularmente importante en la fase de clasificación de rivales en el sistema AB.

En cuanto a los parámetros que no se autoajustan, el parámetro de sensibilidad λ fue fijado en 1 para que la probabilidad asignada correspondiera completamente al valor de las atracciones. Asimismo, el parámetro N, que representa la experiencia de los jugadores, sigue la misma regla planteada por los autores de incremento débil a lo largo de los ensayos.

Finalmente, se pueden asignar valores a priori para las atracciones y para el parámetro N. El valor inicial de N se fijó en 1, esto para que no se vea afectado el valor inicial de las atracciones que fue asignado por introspección pre-juego. Con esto en cuenta, el valor inicial de cada estrategia es igual a la ganancia esperada de dicha estrategia cuando no se tiene información de la frecuencia de respuestas del rival, es decir, asumiendo que el que el rival elige entre aceptar o rechazar con la misma probabilidad en un inicio. Además, el valor de una estrategia del agente FEWA está determinada por la recompensa que se obtendría si el rival elige aceptar dicha partición por su probabilidad inicial, más la recompensa que se obtendría si el rival elige rechazar la partición por su probabilidad inicial. Lo anterior determina el valor inicial de las atracciones de cada estrategia, en cada ronda, para cada rival.

Categorización de rival desconocido

Este algoritmo para categorizar rivales fue implementado en RL como en FEWA para que ambos agentes sean capaces de distinguir contra quién están jugando en el sistema AB. Esta nueva función consiste en usar las políticas óptimas finales de la Matriz de Valor Q aprendida en el caso de RL y la Matriz de Atracciones aprendida en el caso de FEWA, para hacerle al rival desconocido la mejor oferta (o secuencia de ofertas) de cada uno de sus rivales aprendidos. En otras palabras, si ambos agentes aprendieron con anterioridad a tres rivales diferentes, en un ensayo de prueba, tanto RL como FEWA, le hacen tres ofertas diferentes a su rival desconocido, una oferta por cada rival que aprendieron. Por el momento se fijó que esta función ejecute 10 ensayos de prueba, por lo tanto RL y FEWA hacen un total de 30 ofertas al rival desconocido y obtienen sus respectivas respuestas.

Además, tanto el agente RL como el agente FEWA aprendieron un historial de estrategias aceptadas por cada rival enfrentaron, este historial es una matriz con una cantidad de vectores equivalente al número de rivales que enfrentaron, y un vector contiene una cantidad de valores equivalente al número de estrategias disponibles, cada vector contiene la frecuencia de las estrategias que fueron aceptadas y de las que fueron rechazadas.

Para identificar contra qué rival se enfrentó el agente RL o FEWA, se compara el historial de estrategias de cada uno de los tres rivales aprendidos con las respuestas que el rival desconocido dio a todas las ofertas en todos los ensayos de prueba, esto quiere decir que se comparan todas las aceptaciones y rechazos en las que el rival desconocido coincidió con cada rival aprendido, y aquellas en las que coinciden fueron contadas como aciertos. Por lo tanto, el rival aprendido con el que coincida en más respuestas será el tipo del rival más probable con el que se asociará al rival desconocido. Sin embargo, la manera en la que se determinó la probabilidad de que el rival desconocido se asemeje a uno de los rivales aprendidos, fue usando el número de aciertos (respuestas que coinciden) que el rival desconocido tiene con cada rival aprendido para determinar la densidad de probabilidad de una distribución Dirichlet, que es una generalización multinomial de una distribución beta. Debido a esto fue que también se contaron las coincidencias en rechazos, simplemente para distinguir con menos incertidumbre contra quien se está jugando.

Sin embargo, únicamente la función de categorización de rival desconocido del agente FEWA se le añadieron tres funciones diferentes basándose en los supuestos de los parámetros del modelo.

El primero es la capacidad de iniciar probando sus mejores estrategias posibles basándose en su conocimiento pre-juego. En este caso se está aprovechando la capacidad de FEWA de implementar valores a priori en las atracciones y en el parámetro N para probar al rival desconocido, como si en un inicio se estuviera basando en la introspección pre-juego y no en su experiencia previa con otros rivales. Si en un ensayo de prueba de la función de categorización de rival desconocido del agente RL se hace una oferta por rival aprendido, entonces FEWA hace una oferta de una matriz de atracciones a priori y después hace una oferta por rival aprendido. Esto se debe a que, si no hay consecuencias, probar tu mejor estrategia posible no está de más cuando te enfrentas a un rival que desconoces.

El segundo agregado consiste en una función para determinar si FEWA nunca se ha enfrentado con anterioridad al rival desconocido actual. Como se mencionó anteriormente, para lograr esto re-adaptamos la función del parámetro ϕ, ahora en lugar de usar el historial de frecuencias de estrategias de un rival, usa la suma de los historiales de frecuencias de estrategias de todos los rivales aprendidos, esto se hace para evitar falsas alarmas, ya que de esta manera si sale un valor de ϕ bajo quiere decir que ninguno de los tres agentes solía aceptar (o rechazar) una estrategia que acaba de ser aceptada o rechazada por el nuevo rival desconocido en la etapa de prueba. Entonces, se compara la suma de los historiales de frecuencias de estrategias de todos los rivales aprendidos con el historial de unos y ceros de estrategias recientemente usadas por el rival desconocido, y así extraer el índice sorpresa para calcular ϕ. Otro agregado fue un umbral de 0.2, que es un valor suficientemente bajo como para sugerir que hay un cambio verdadero, por lo que si el valor de ϕ no sobrepasa este umbral significa que este rival desconocido ha respondido algo que ninguno de los rivales aprendidos ha respondido anteriormente.

Finalmente, en tercer lugar se implementó la opción de aprender a un rival desconocido si FEWA nunca se ha enfrentado a él con anterioridad. Para esto quisimos aprovechar la función original de ϕ que implica olvidar el conocimiento previo para facilitar nuevo aprendizaje una vez que se detecta un cambio. Sin embargo preferimos que no olvide el conocimiento previo, pero que si tuviera un nuevo inicio para facilitar el aprendizaje. Si el valor de ϕ de las respuestas del rival desconocido no sobrepasa el umbral de 0.2, el agente decide aprender de este nuevo rival sin sesgos de su experiencia previa, sin embargo, no borra la Matriz de Atracciones de los anteriores rivales. El método más efectivo fue ejecutar una nueva fase de aprendizaje contra este rival, es decir, FEWA concatena una nueva matriz en la Matriz de Atracciones y vuelve a simular ensayos del juego de la negociación contra este rival descono-

cido para llenar la nueva matriz concatenada. En este escenario la Matriz de Valor Q y la Matriz de Atracciones terminarían siendo más diferentes.

4.7.2. Fase de prueba de tamaños de población

Experimento simulado de tres grupos. El primer grupo tiene una población de 30 agentes que responden de acuerdo al modelo de Aversión a la Inequidad, a estos les llamaremos agentes que no aprenden, el segundo es una población de 150 agentes que no aprenden y el tercero es de 300 agentes que no aprenden. A las tres poblaciones se les agregaron los agentes FEWA y RL, ambos serán referenciados como agentes que aprenden. Se tomaron medidas repetidas de los agentes que aprenden como categorización del rival enfrentado, posiciones, estrategias seleccionadas y ganancias de cada agente que aprende a lo largo del proceso de prueba. También se toman medidas repetidas de los agentes que no aprenden sobre la estrategia seleccionada, posiciones, ganancias totales de un grupo del mismo tipo y ganancias por agente a lo largo del proceso de prueba.

Tarea

Las interacciones están programadas para simular una tarea de negociación de una sola ronda, es decir, un juego del ultimátum. Se cambió la cantidad de rondas por dos razones, la primera es que, sin importar cuánto se aumenten las rondas, el resultado es el mismo para los agentes que no aprenden, es decir, si en la primera ronda no hubo un acuerdo lo más probable es que no habrá un acuerdo en una ronda 3 o en una ronda 10. Además, los agentes que aprenden, cuando se enfrentan a un nuevo rival ejecutan su función de categorización de rival y, después de ejecutar dicha función, desde la primera ronda hacen una oferta que consideran óptima, por lo que los agentes que aprenden no necesitan de más rondas para llegar a un acuerdo. La segunda razón es que entre más rondas haya en la tarea más capacidad computacional requiere la simulación de un solo ensayo.

Espacio

El escenario de prueba es un ambiente espacial que inicialmente asigna una posición aleatoria en el entorno a los agentes que aprenden y a cada agente que no aprende de la población. El tamaño del plano donde se moverán los agentes depende del tamaño de la población, por lo que el tamaño del ambiente se define como:

$$T = \lceil \sqrt{n+2} \rceil$$

donde n representa el número de agentes y $\lceil \cdot \rceil$ es la función techo. Así, la matriz del ambiente se representa como $A \in \mathbb{R}^{T \times T}$ donde A es una matriz de ceros de tamaño $T \times T$.

Los únicos obstáculos del ambiente son los cuatro cuadrantes que delimitan el entorno, además ningún agente puede tomar una posición que ya ha sido tomada por otro agente, más allá de eso los agentes son libres de moverse sin restricciones en el ambiente. El tamaño del espacio es importante debido a que la selección de los agentes que participaban en la tarea de negociación depende de la distancia entre ellos. Específicamente, si dos o más agentes están a menos de una unidad de distancia en el entorno entonces tienen que ejecutar un juego de negociación por cada uno, es importante mencionar que ejecutar un juego tiene el costo de 25 puntos para cada agente involucrado.

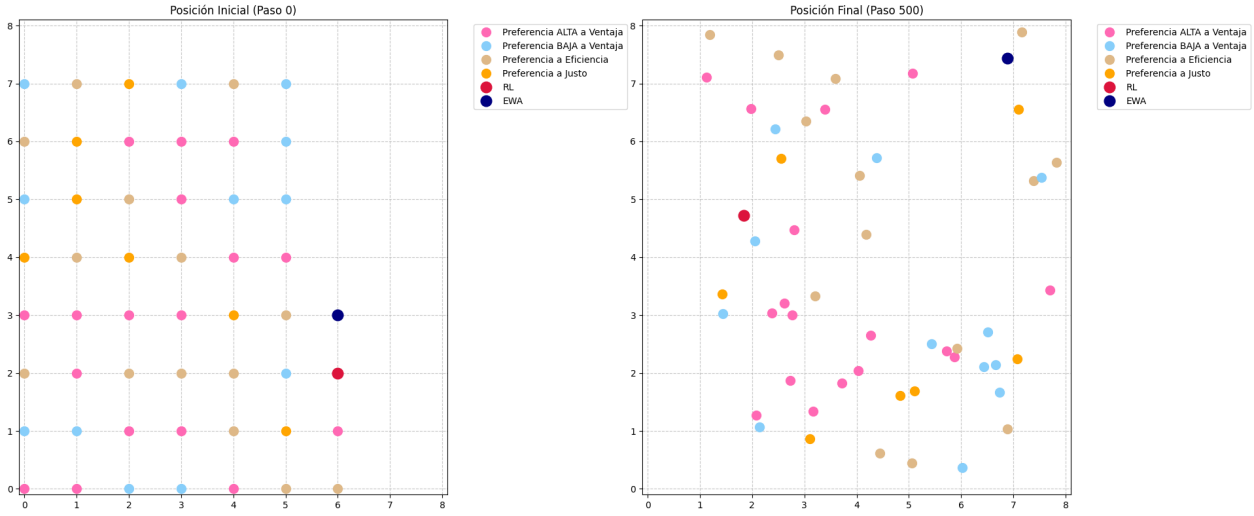

Figura 6: Ejemplo del espacio cuando la población es de 50 agentes.

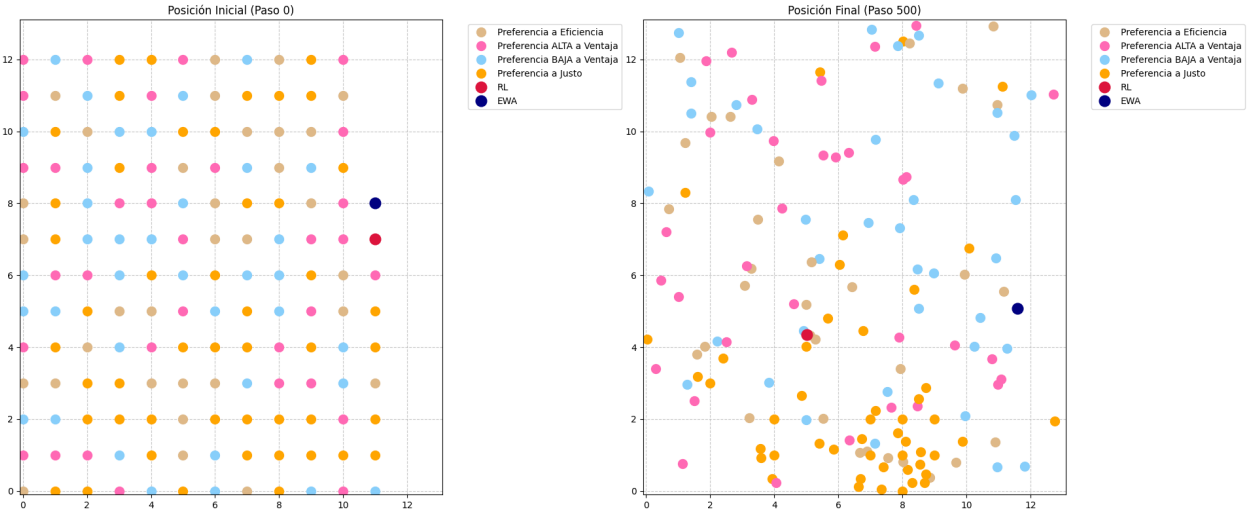

Figura 7: Ejemplo del espacio cuando la población es de 150 agentes.

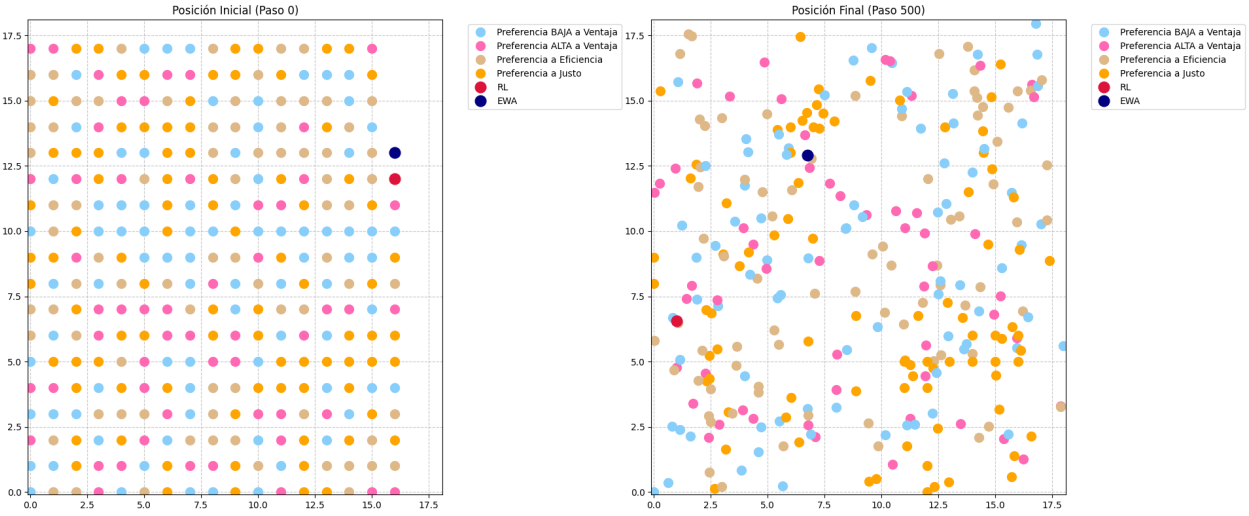

Figura 8: Ejemplo del espacio cuando la población es de 300 agentes.

Movimiento de agentes que no aprenden

El movimiento de cada agente que no aprende depende de la siguiente regla: si estás aumentando tus ganancias conserva tu posición, si estás perdiendo ganancias muévete. Si un agente que no aprende determina que debe moverse por las pérdidas, entonces se le selecciona una nueva posición aleatoria en un rango de -2 y 2 unidades para x (un movimiento hacia la izquierda o la derecha) y -2 y 2 unidades para y (un movimiento hacia arriba o hacia abajo).

Movimiento de agentes que aprenden

El movimiento de los agentes que aprenden es diferente, ya que se les asignó comportamiento de exploración y explotación. Esto se debe a que ambos agentes que aprenden son capaces de identificar a su rival para hacerles la mejor oferta aprendida, por lo que en realidad es raro que cualquiera de los dos agentes que aprenden lleguen a desacuerdos y pierdan ganancias. Así que, si los agentes que aprenden usan la regla de movimiento de los agentes que no aprenden no se moverían en la mayoría de los ensayos o no se moverían en lo absoluto, dependiendo de donde les toque inicialmente. Otra razón es que si siguen una regla de elección para elegir qué oferta hacerle a un rival, parece natural que también sigan una regla de elección para elegir el rival con el que quieran ofertar. Este nuevo enfoque de movimiento ejecuta el método de exploración durante los ensayos iniciales, esta etapa consiste en que los agentes que aprenden en cada ensayo eligen aleatoriamente a un rival de la población de agentes que no aprenden con el que no hayan jugado para acercarse a él, la nueva posición de los agentes que aprenden estará en un rango de -0.5 y 0.5 unidades para x y -0.5 y 0.5 unidades para y alrededor del rival seleccionado. Una vez que hayan jugado con todos los agentes que no aprenden de la población empieza la ejecución del método de explotación, esta etapa consiste en que los agentes que aprenden en cada ensayo eligen aleatoriamente a uno de los agentes que no aprenden que les haya aceptado la oferta más alta para acercarse a él, la nueva posición de los agentes que aprenden está en un rango de -0.5 y 0.5 unidades para x y -0.5 y 0.5 unidades para y alrededor del rival seleccionado.

Selección de estrategia de los agentes que no aprenden

Como ya se ha mencionado, en un juego de ultimátum solo hay una ronda para que un agente ofrezca y para que el otro agente decida si se acepta la partición o si se rechaza, aquí también deben dividirse un presupuesto de 100 puntos. En cada ensayo cada agente puede llevar a cabo una negociación, si los agentes seleccionados para ejecutar un juego son ambos agentes que no aprenden entonces el agente que no aprende 1 se le asigna el rol del jugador que oferta y el agente que no aprende 2 el rol del jugador que decide.

Una vez asignados los roles, el jugador que oferta debe de proponer una partición del presupuesto total para ambos, esto lo hace de acuerdo a la regla de ofertas que se propuso en la fase de aprendizaje, es decir, el agente con preferencia baja a resultados ventajosos asigna probabilidad uniforme a su conjunto de ofertas con utilidad positiva ($a \sim \mathcal{U}$ para a con $u(a) > 0$), el agente con preferencia media a resultados justos asigna una probabilidad normal a su conjunto de ofertas con media en la mitad del conjunto de ofertas ($a \sim \mathcal{N}(\mu, \sigma^2)$ para a con $u(a) >$

0), el agente con preferencia alta a resultados ventajosos le asigna probabilidad exponencial a su conjunto de ofertas ($a \sim \text{Exp}(\lambda)$ para a con $u(a) > 0$), siendo las más bajas para el rival a las que les asigna más probabilidad, y el agente con preferencia por resultados eficientes hace la oferta más baja posible. La intención es que el primero ofrezca aleatoriamente, el segundo ofrezca cercano a lo equitativo para ambos, el tercero ofrezca en su mayoría muy favorecedor para él y que el cuarto ofrezca la menor cantidad posible sin llegar a cero. De esta manera, el jugador que decide aceptar o rechazar la partición considera, de acuerdo a sus preferencias del modelo de IA, si la oferta que le hicieron tuvo una utilidad mayor a cero, si este fue el caso entonces la oferta es aceptada y cada agente recibe la ganancia acordada, por el otro lado, si es rechazada ninguno recibe nada, al final del juego a ambos se les descuenta el costo por jugar.

Después los mismos agentes vuelven a jugar pero se invierten los roles, ahora el agente que no aprende 2 es el jugador que oferta y el agente que no aprende 1 es el jugador que decide, se repite el juego y se vuelve a descontar el costo por jugar. Este proceso es ejecutado para todos los agentes que estén a una unidad de distancia o menos, si un agente está a menos de una unidad de distancia de, por ejemplo, 5 agentes, ese agente debe jugar contra esos 5 agentes, si ese agente no está lo suficientemente cerca de nadie, entonces ese agente pasa ese ensayo sin enfrentar a ningún agente y, por lo tanto, también pasa el ensayo sin ganancias ni pérdidas.

Selección de estrategia de los agentes que aprenden

Los agentes que aprenden no pueden jugar entre ellos. Si los agentes seleccionados para llevar a cabo una negociación es un agente que aprende y un agente que no aprende, entonces al agente que aprende siempre se le asigna el rol del jugador que oferta y al agente que no aprende se le asigna el rol del jugador que decide, esto con el objetivo de que la política óptima aprendida sea lo único que influya en los resultados de los agentes que aprenden. Una vez que se asignaron los roles, el agente que aprende ejecuta su función de categorización de rival desconocido, la función ejecuta juegos de prueba con el agente que no aprende y determina la probabilidad de que el agente que no aprende se asemeje a uno de los rivales aprendidos, el rival aprendido que tenga mayor probabilidad de ser el agente que no aprende será el rival al que el agente que aprende asume que se está enfrentando, y con esta información presente le hará una oferta óptima aprendida. Este proceso es ejecutado para todos los agentes que estén a una unidad de distancia o menos.

4.7.3. Fase de prueba de distribuciones de población

Experimento simulado de tres grupos. Todos los agentes que no aprenden de la población responden de acuerdo al modelo de Aversión a la Inequidad, cada una de las siete distribuciones de poblaciones de agentes que no aprenden fue descrita en el apartado de "Muestra". El hecho de que los agentes que no aprenden estén distribuidos de manera diferente en cada grupo, quiere decir que cada agente tiene diferente probabilidad de surgir en la población cuando es generada. Sin embargo, es importante recordar que las últimas dos poblaciones son las únicas que incluyen a un agente que no aprende nuevo y desconocido para los agentes RL y FEWA, este nuevo agente tiene preferencia a resultados eficientes. Todas las poblaciones incluyen a los agentes FEWA y RL, ambos serán referenciados una vez más como agentes que aprenden. Se tomaron medidas repetidas de los agentes que aprenden como categorización del rival enfrentado, posiciones, estrategias seleccionadas y ganancias de cada agente que aprende a lo largo del proceso de prueba. También se toman medidas repetidas de los agentes que no aprenden sobre la estrategia seleccionada, posiciones, ganancias totales y ganancias por agente a lo largo del proceso de prueba.

Tarea

La tarea es la misma que en la fase de prueba de tamaños de población.

Espacio

Las condiciones del espacio son las mismas que en la fase de prueba de tamaños de población.

Movimiento de agentes que no aprenden

La regla de movimiento es la misma que en la fase de prueba de tamaños de población.

Movimiento de agentes que aprenden

Las reglas de movimiento son las mismas que en la fase de prueba de tamaños de población.

Selección de estrategia de los agentes que no aprenden

La elección de estrategia es la misma que en la fase de prueba de tamaños de población.

Selección de estrategia de los agentes que aprenden

La elección de estrategia es la misma que en la fase de prueba de tamaños de población.

Sin embargo, como se mencionó en la fase de aprendizaje, las funciones de categorización de rival desconocido de RL y FEWA son diferentes. Por lo que RL asiganará probabilidad a cada agente que no aprende de acuerdo a los rivales que ya aprendió y FEWA asiganará probabilidad a cada agente que no aprende de acuerdo con los rivales que ya aprendió y, además, podrá identificar si un agente que no aprende al que se está enfrentando no se asemeja a ninguno de los rivales conocidos, si este es el caso FEWA ejecutará toda una simulación de fase de aprendizaje con ese agente y después continuará con la tarea.

4.8. Instrumentos

Computadora y softwares de programación compatibles con Python.

5. Cronograma

Programación del ambiente de negociación que incluya, reglas del juego (presupuesto, jugadores, rondas e información), orden de procedimientos, secuencia de funciones e iteraciones. Programación del modelo. Integración del modelo al ambiente como método para que el agente que aprenda obtenga respuestas y recompensas. Programación del agente RL. Integración del agente RL al ambiente como método de toma de decisiones por medio de exploración con incertidumbre, explotación por medio de asignación de probabilidades. Programación del agente FEWA. Integración del agente FEWA al ambiente como método de toma de decisiones por medio de exploración con incertidumbre, explotación por medio de asignación de probabilidades.

Programación del ambiente de negociación de una ronda que incluya las reglas del juego, orden de procedimientos, secuencia de funciones e iteraciones. Programación del modelo. Integración del modelo al ambiente como método para obtener n agentes con diferentes ponderaciones de los que el agente que aprendió obtenga respuestas y recompensas. Programación de movimiento de agentes en el entorno y restricciones de este mismo. Integración de los agentes RL y FEWA como método de toma de decisiones por medio de explotación por medio de asignación de probabilidades.

6. Análisis

6.1. Reinforcement Learning Q-Learning

1. ¿Cómo afectan los parámetros del modelo de Aversión a la Inequidad a las estrategias de negociación óptimas aprendidas por un agente RL Q-Learning?

   a) El agente RL Q-Learning aprenderá a hacer ofertas más equitativas a medida que los parámetros de aversión a la inequidad del rival aumentan.

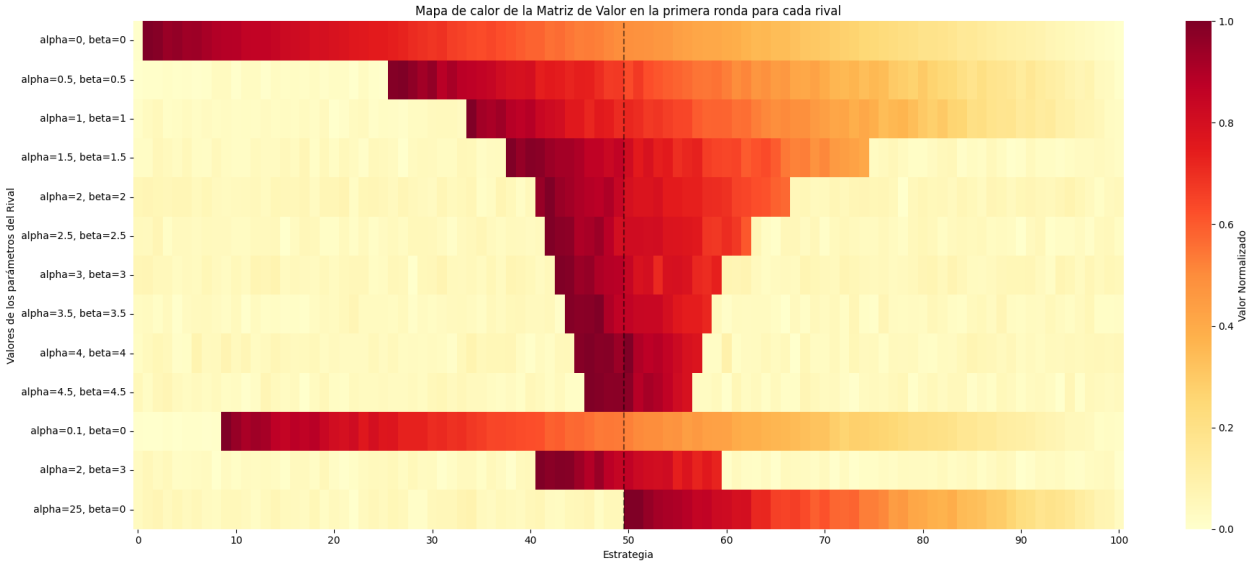

Figura 9: Valor normalizado de las ofertas de la primera ronda de RL contra rivales con diferentes preferencias de IA.

Se corrieron un total de 260 simulaciones, 20 simulaciones por rival, cada simulación consta de 2500 juegos de negociación contra el rival seleccionado. Dado que el agente aprende que ofrecer la mejor oferta desde la primera ronda es lo más conveniente, el Heatmap contiene los valores promediados de las 20 simulaciones de las acciones de la primera ronda, esto para cada uno de los 13 rivales. Se probaron 10 rivales más, a parte de los rivales de la fase de aprendizaje, para ilustrar los efectos de los parámetros del rival sobre la matriz de valores aprendidos de RL. Los valores fueron normalizados. Las ofertas que el agente RL le haga a cualquier rival de 0 a 49 son ponderadas por alpha para el rival y las que el agente RL le haga de 51 a 100 son ponderadas por beta para el rival, la línea punteada muestra esta distinción. Conforme los valores de los parámetros del rival incrementan se acota más el rango de ofertas con valor para el agente RL.

1. ¿La política óptima aprendida por el agente RL Q-learning se acerca o se aleja del Equilibrio Perfecto en Subjuegos de Nash?

   a) El Equilibrio Perfecto en Subjuegos de Nash (SPNE) será diferente para cada rival, y la política óptima aprendida por el agente será la misma que el mismo SPNE que aplique para ese rival.

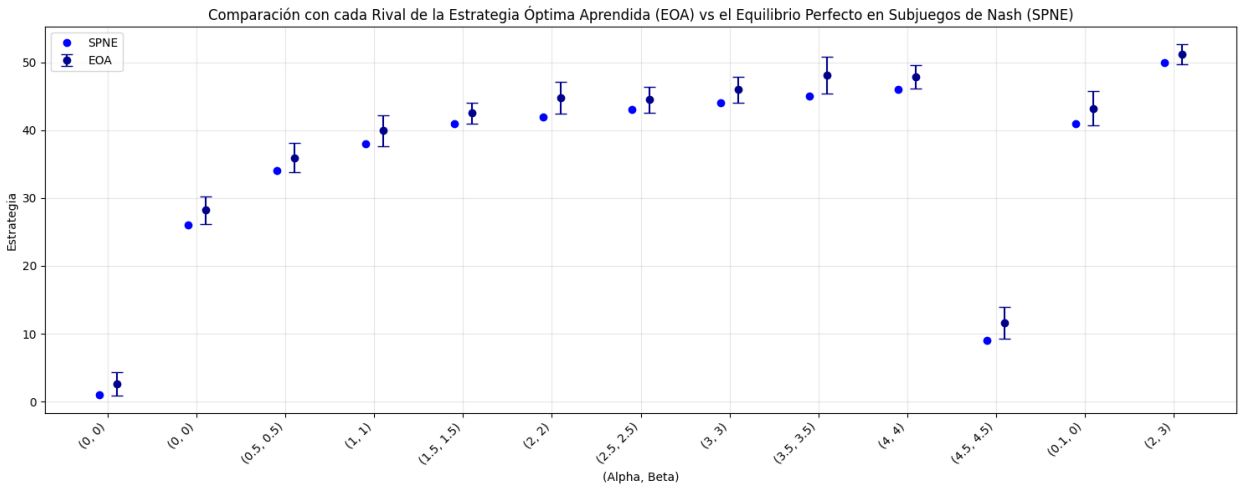

Figura 10: SPNE de cada rival y Estrategia Óptima aprendida por RL contra cada rival.

Si el agente aprende las preferencias del rival, y hacemos inducción hacia atrás, la mejor estrategia posible (SPNE) es que el agente ofrezca la oferta óptima aceptable desde la primera ronda y que el rival la acepte. Un caso particular es que si el agente aprende las preferencias del rival, el límite de rondas, que el rival también tiene conocimiento del límite de rondas y que en la última ronda le corresponde ofertar a él y no al rival, la mejor estrategia en este caso (SPNE) es que el agente ofrezca ofertas que el rival rechazará, que rechace sus contraofertas y que ofrezca la mejor oferta posible en la última ronda, lo cuál el agente RL logra aprender. Sin embargo, nos interesa comparar las estrategias óptimas aprendidas por el agente de cada rival con sus respectivos SPNE, y este escenario en particular se deshace de las preferencias del rival, por lo que no lo tomamos en cuenta.

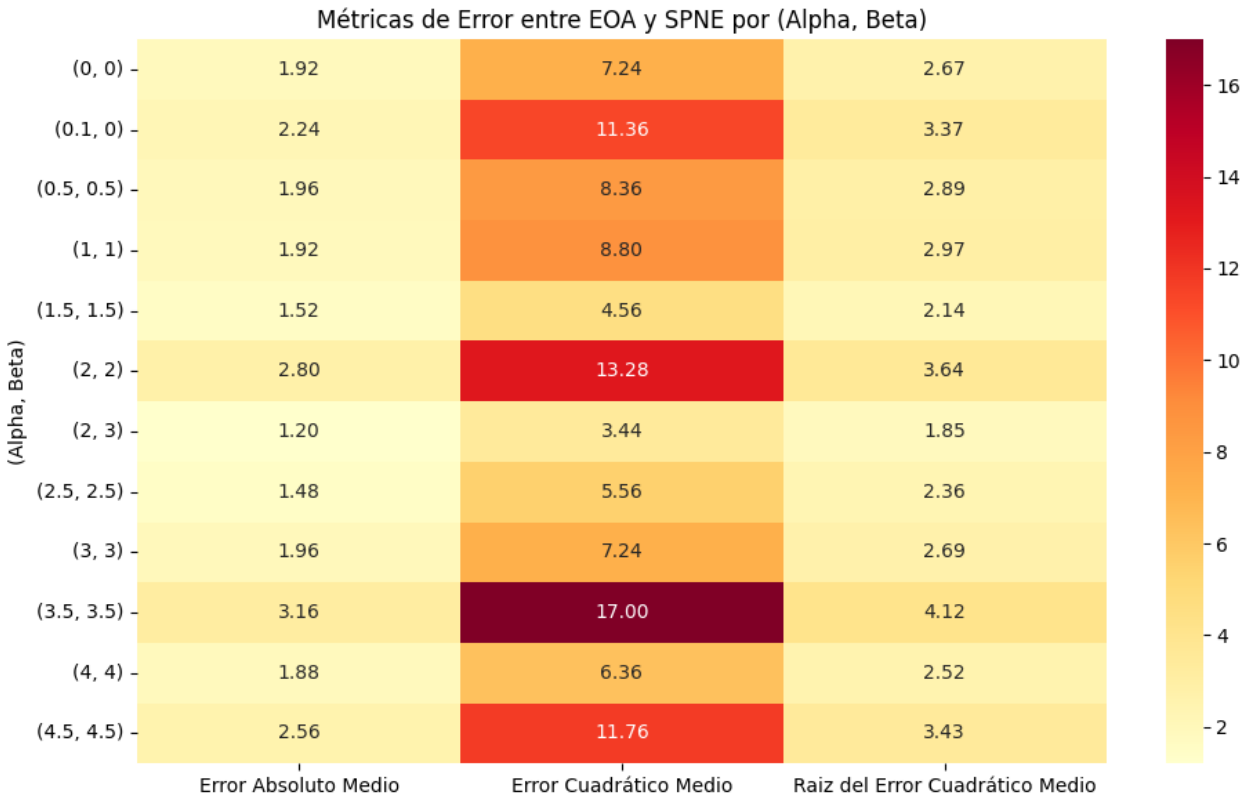

Figura 11: Error entre SPNE de cada rival y Estrategia Óptima aprendida por RL contra cada rival.

Tomando en cuenta lo anterior, se corrieron un total de 260 simulaciones, 20 simulaciones por rival, cada simulación consta de 2500 juegos de negociación contra el rival seleccionado. Los gráficos muestran el promedio de los errores medios por rival en las 20 simulaciones, considerando que el error absoluto medio entre la estrategia óptima aprendida para cada rival y su SPNE es en promedio 1.8923 el agente RL lograba aprender constantemente estrategias cercanas al SPNE.

6.2. Functional Experience Weighted Attraction

1. ¿Cómo afectan los parámetros del modelo de Aversión a la Inequidad a las estrategias de negociación óptimas aprendidas por un agente FEWA?

    a) El agente FEWA aprenderá a hacer ofertas más equitativas a medida que los parámetros de aversión a la inequidad del rival aumentan.

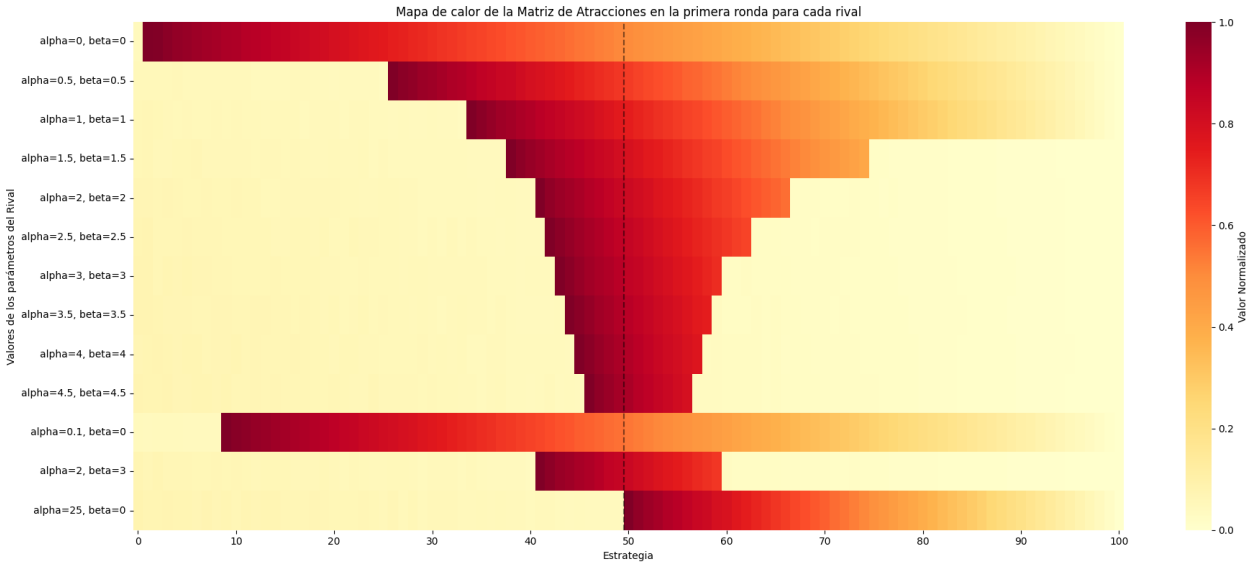

Figura 12: Valor normalizado de las ofertas de la primera ronda de FEWA contra rivales con diferentes preferencias de IA.

Se corrieron un total de 260 simulaciones, 20 simulaciones por rival, cada simulación consta de 2500 juegos de negociación contra el rival seleccionado. Dado que el agente aprende que ofrecer la mejor oferta desde la primera ronda es lo más conveniente, el Heatmap contiene los valores promediados de las 20 simulaciones de las acciones de la primera ronda, esto para cada uno de los 13 rivales. Se probaron 10 rivales más, a parte de los rivales de la fase de aprendizaje, para ilustrar los efectos de los parámetros del rival sobre la matriz de valores aprendidos de FEWA. Los valores fueron normalizados. Las ofertas que el agente FEWA le haga a cualquier rival de 0 a 49 son ponderadas por alpha para el rival y las que el agente FEWA le haga de 51 a 100 son ponderadas por beta para el rival, la línea punteada muestra esta distinción. Conforme los valores de los parámetros del rival incrementan se acota más el rango de ofertas con valor para el agente FEWA.

1. ¿La política óptima aprendida por el agente FEWA se acerca o se aleja del Equilibrio Perfecto en Subjuegos de Nash?

    a) El Equilibrio Perfecto en Subjuegos de Nash (SPNE) será diferente para cada rival, y la política óptima aprendida por el agente será la misma que el mismo SPNE que aplique para ese rival.

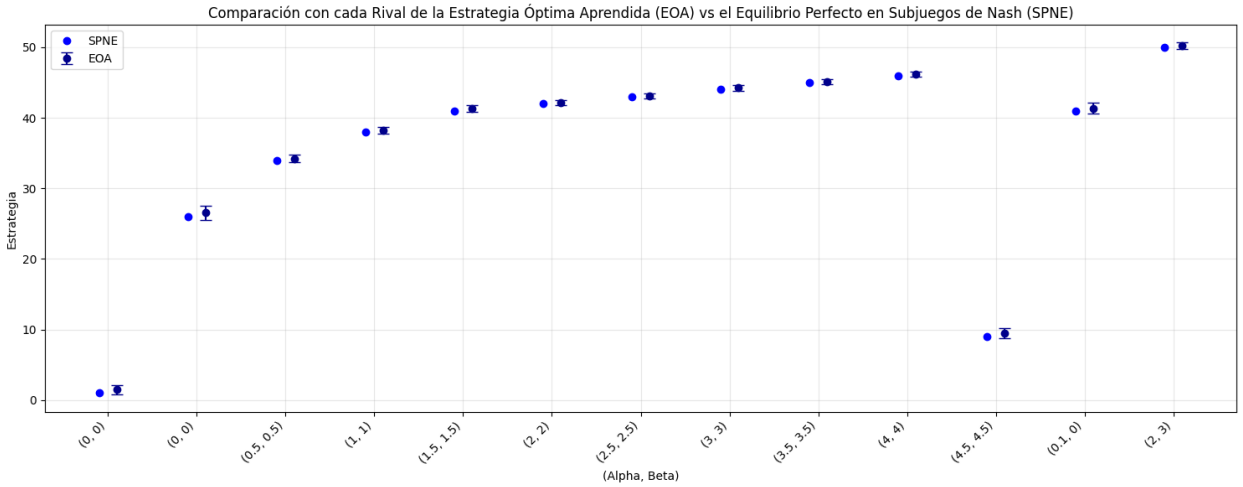

Figura 13: SPNE de cada rival y Estrategia Óptima aprendida por FEWA contra cada rival.

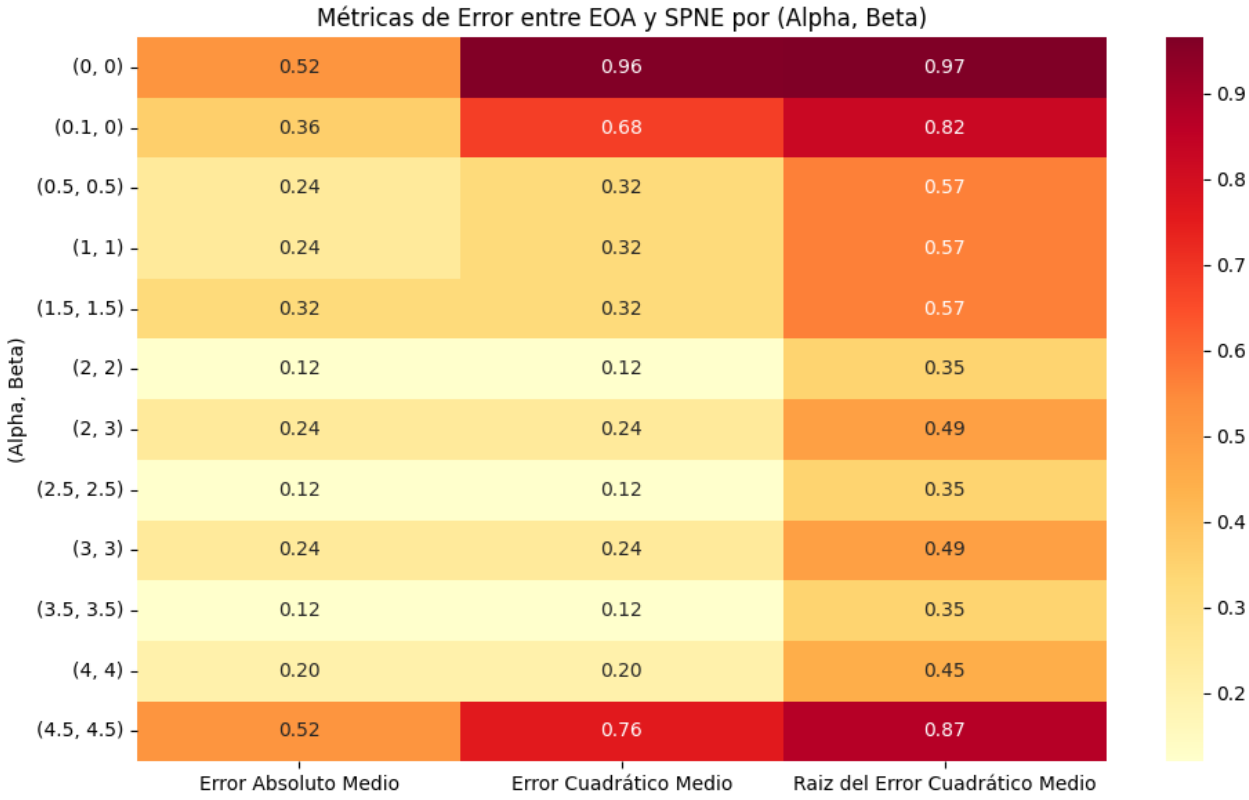

Figura 14: Error entre SPNE de cada rival y Estrategia Óptima aprendida por FEWA contra cada rival.

Se corrieron un total de 260 simulaciones, 20 simulaciones por rival, cada simulación consta de 2500 juegos de negociación contra el rival seleccionado. Los gráficos muestran el

promedio de los errores medios por rival en las 20 simulaciones, considerando que el error absoluto medio entre la estrategia óptima aprendida para cada rival y su SPNE es en promedio 0.2492 el agente FEWA lograba aprender constantemente el SPNE o estrategias cercanas.

6.3. Reinforcement Learning Q-Learning y Functional Experience Weighted Attraction

1. ¿En qué se diferenciarán las estrategias óptimas aprendidas por los modelos?

    a) Ambos modelos aprenderán la política óptima que sea igual al SPNE para cada rival.

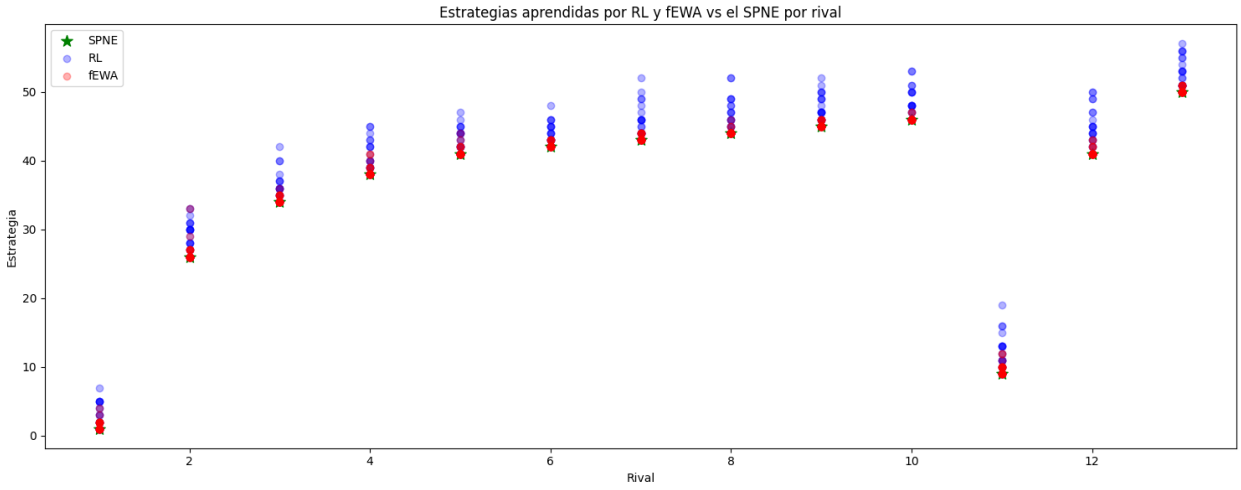

Figura 15: SPNE de cada rival y Estrategia Óptima aprendida por FEWA y RL contra cada rival.

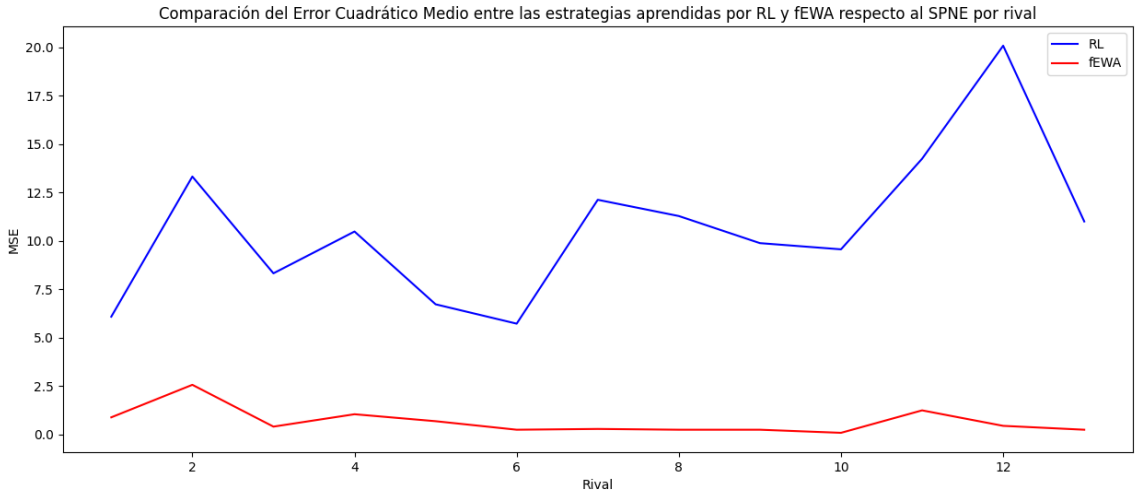

Figura 16: Error entre SPNE de cada rival y Estrategia Óptima aprendida por FEWA y RL contra cada rival.

El error cuadrático medio de las estrategias óptimas de FEWA promediado para los 13 rivales es 0.3385, mientras que el de RL es 8.0738, por lo que podemos notar que FEWA en general se acerca más al SPNE que RL.

1. ¿En qué se diferenciará el desempeño de los modelos?

    a) Ambos modelos obtendrán resultados similares debido a que usan la misma política de exploración y explotación.

Frecuencias de selección:

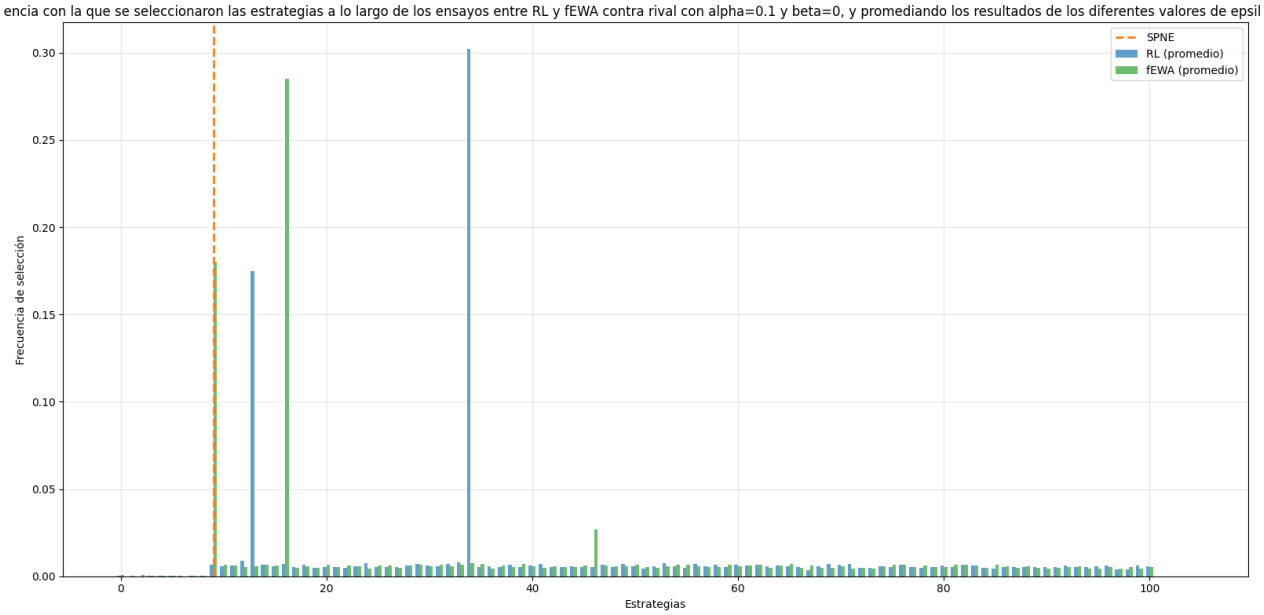

Figura 17: Frecuencia de selección de estrategias de RL y FEWA contra el rival con preferencia baja por resultados ventajosos promediando lo obtenido con las tres diferentes ϵ (0.1, 0.5 y 0.999).

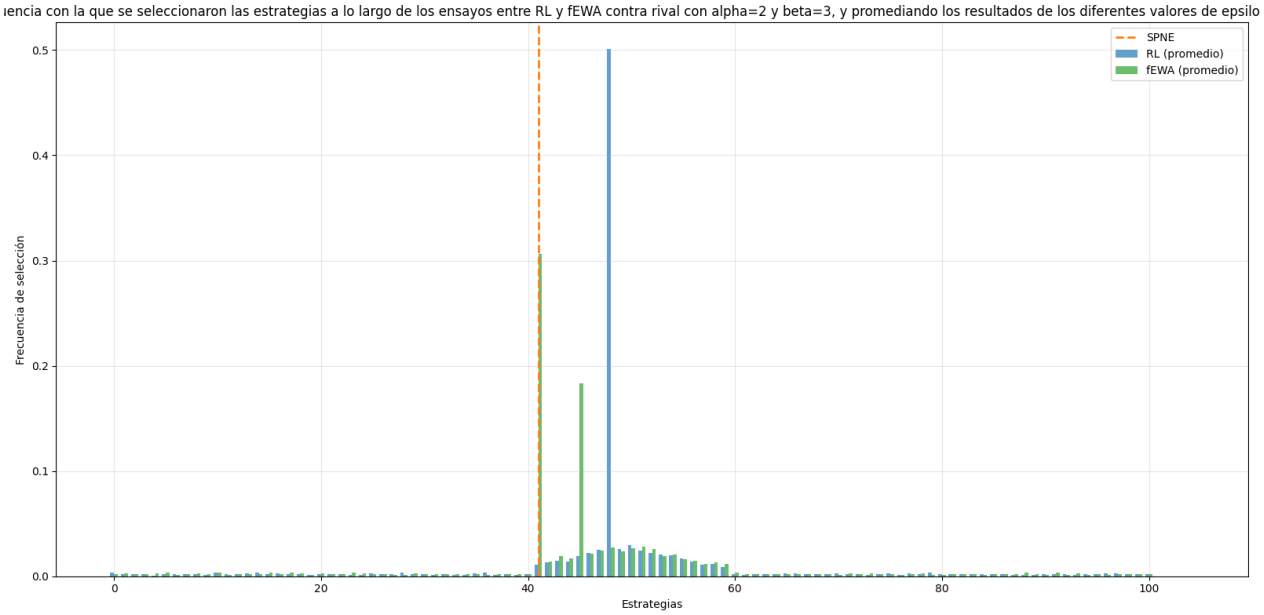

Figura 18: Frecuencia de selección de estrategias de RL y FEWA contra el rival con preferencia media por resultados justos promediando lo obtenido con las tres diferentes ϵ (0.1, 0.5 y 0.999).

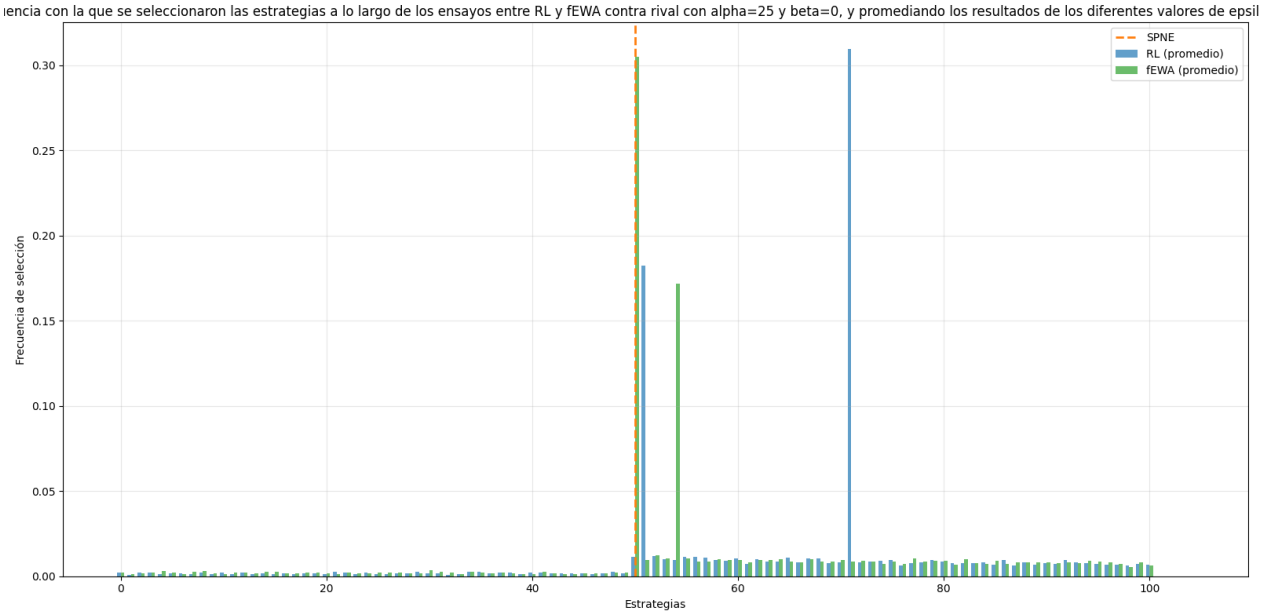

Figura 19: Frecuencia de selección de estrategias de RL y FEWA contra el rival con preferencia alta por resultados ventajosos promediando lo obtenido con las tres diferentes ϵ (0.1, 0.5 y 0.999).

Valor de Estrategias:

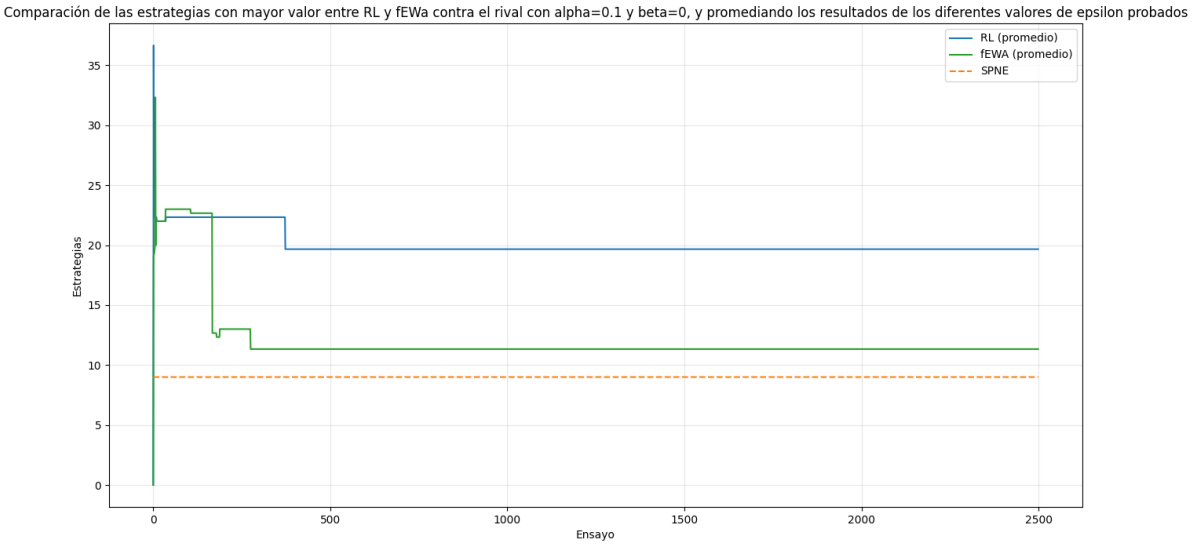

Figura 20: Valor normalizado que RL y FEWA le asignaron a cada estrategia contra el rival con preferencia baja por resultados ventajosos promediando lo obtenido con las tres diferentes ϵ (0.1, 0.5 y 0.999).

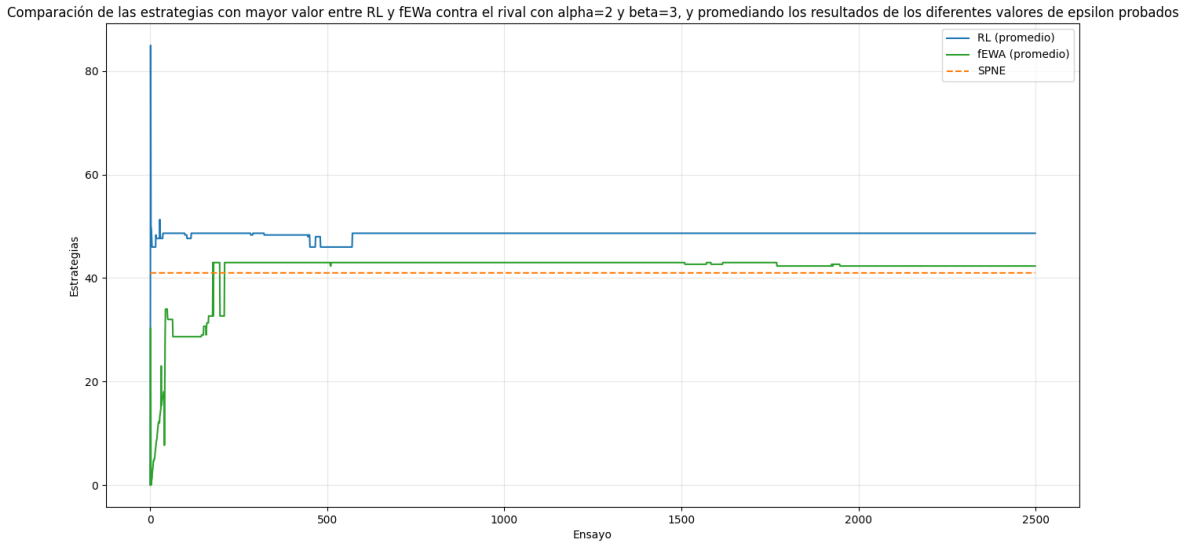

Figura 21: Valor normalizado que RL y FEWA le asignaron a cada estrategia contra el rival con preferencia media por resultados justos promediando lo obtenido con las tres diferentes ϵ (0.1, 0.5 y 0.999).

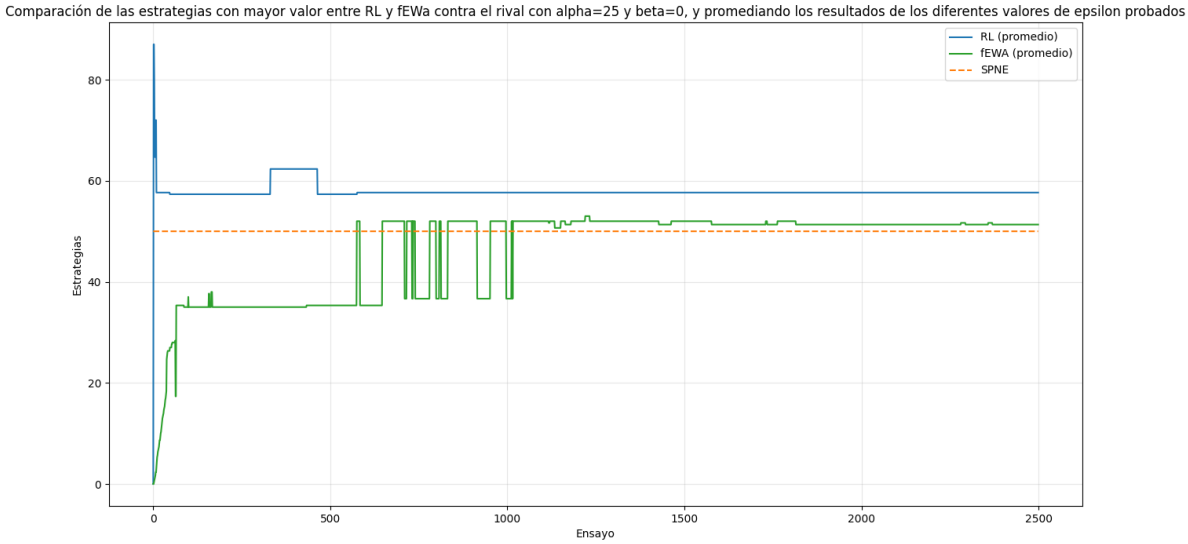

Figura 22: Valor normalizado que RL y FEWA le asignaron a cada estrategia contra el rival con preferencia alta por resultados ventajosos promediando lo obtenido con las tres diferentes ϵ (0.1, 0.5 y 0.999).

Acumulación de ganancias:

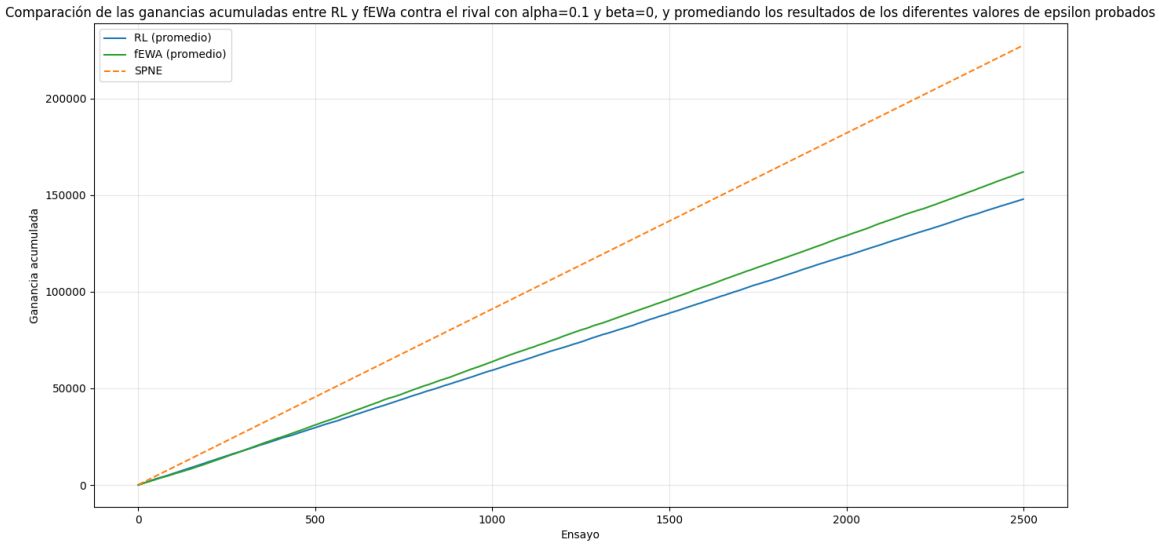

Figura 23: Ganancias acumuladas a lo largo de los ensayos de RL y FEWA contra el rival con preferencia baja por resultados ventajosos promediando lo obtenido con las tres diferentes ϵ (0.1, 0.5 y 0.999).

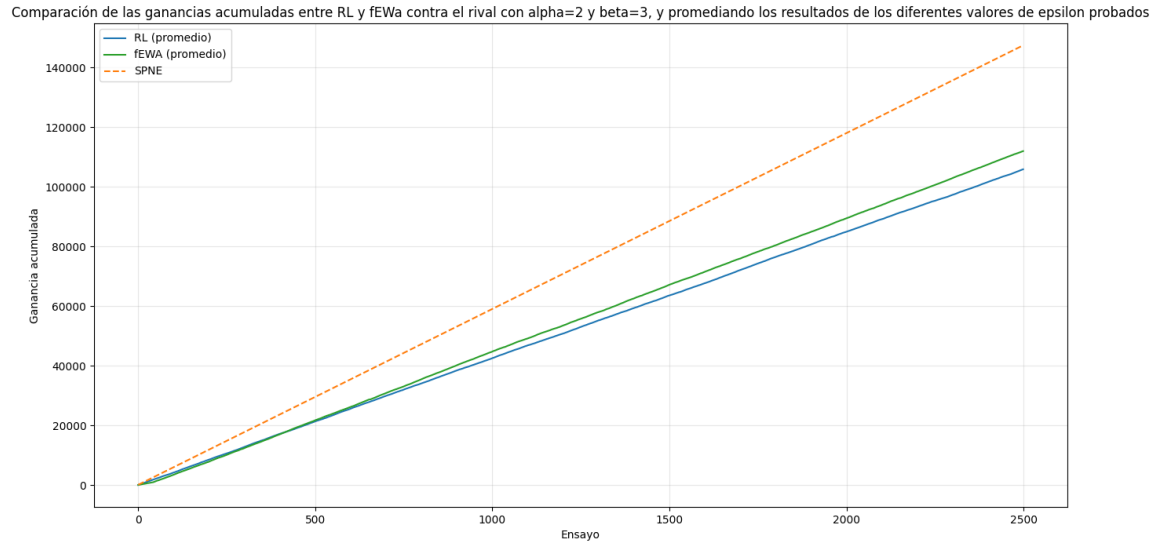

Figura 24: Ganancias acumuladas a lo largo de los ensayos de RL y FEWA contra el rival con preferencia media por resultados justos promediando lo obtenido con las tres diferentes ϵ (0.1, 0.5 y 0.999).

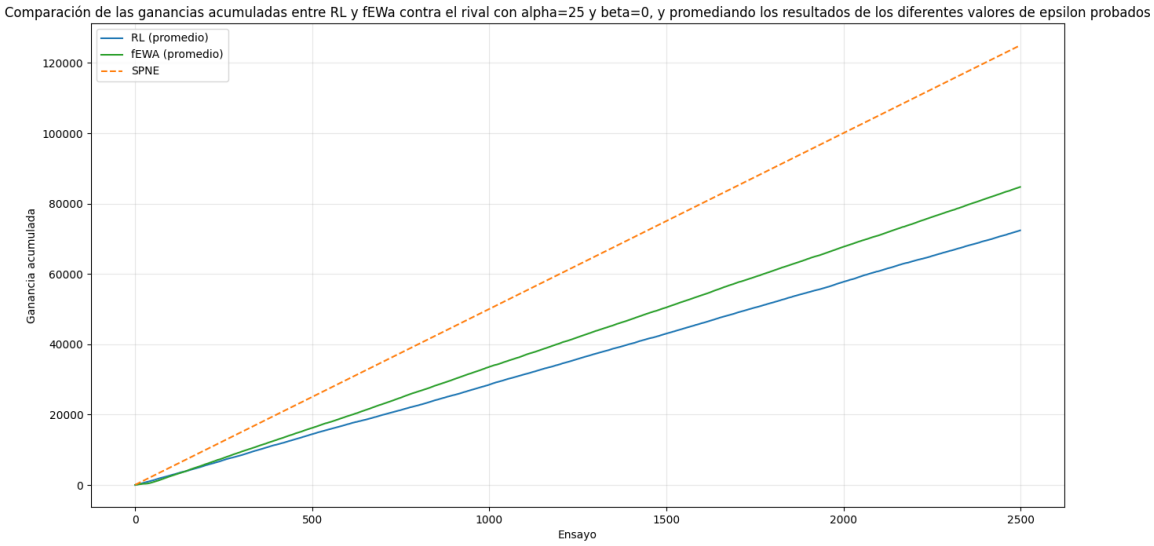

Figura 25: Ganancias acumuladas a lo largo de los ensayos de RL y FEWA contra el rival con preferencia alta por resultados ventajosos promediando lo obtenido con las tres diferentes ϵ (0.1, 0.5 y 0.999).

A diferencia de las anteriores simulaciones, aquí solo se recopilaron los resultados de los tres rivales de interés de la fase de aprendizaje, al rival con preferencia baja a resultados ventajosos, al rival con preferencia media a resultados justos y al rival con preferencia alta a resultados ventajosos. Además, se corrieron 20 simulaciones por cada rival con tres epsilon diferentes, siendo 0.1, 0.5 y 0.999 las diferentes epsilon probadas, dando un total de 180 simulaciones, cada simulación consta de 2500 juegos de negociación contra el rival seleccionado. Los gráficos de cada epsilon para cada rival se presentan en el apartado de Anexos, los gráficos mostrados aquí promedian los resultados de los diferentes valores de epsilon de cada rival.

En cuanto al desempeño, FEWA en general le otorgaba el valor más alto a la estrategia óptima, o estrategias adyacentes a la óptima, más rápido que RL, incluso cuando el epsilon es bajo y es probable que aún no se hayan probado todas las acciones disponibles, esto se debía al elemento BL de FEWA que le otorga valor hipotético incluso a estrategias que no fueron probadas. Por esto, para FEWA la estrategia óptima tenía más frecuencia de selección, incluso variando epsilon para ambos modelos, y por lo tanto FEWA en la mayoría de los casos obtuvo más ganancias que RL.

1. ¿En qué se diferenciará el aprendizaje de los modelos?

a) El valor de las matrices aprendidas por los agentes serán diferentes, la matriz del agente FEWA tendrá valores más altos en general y la diferencia proporcional en los valores de las estrategias será más notoria en la matriz del agente FEWA.

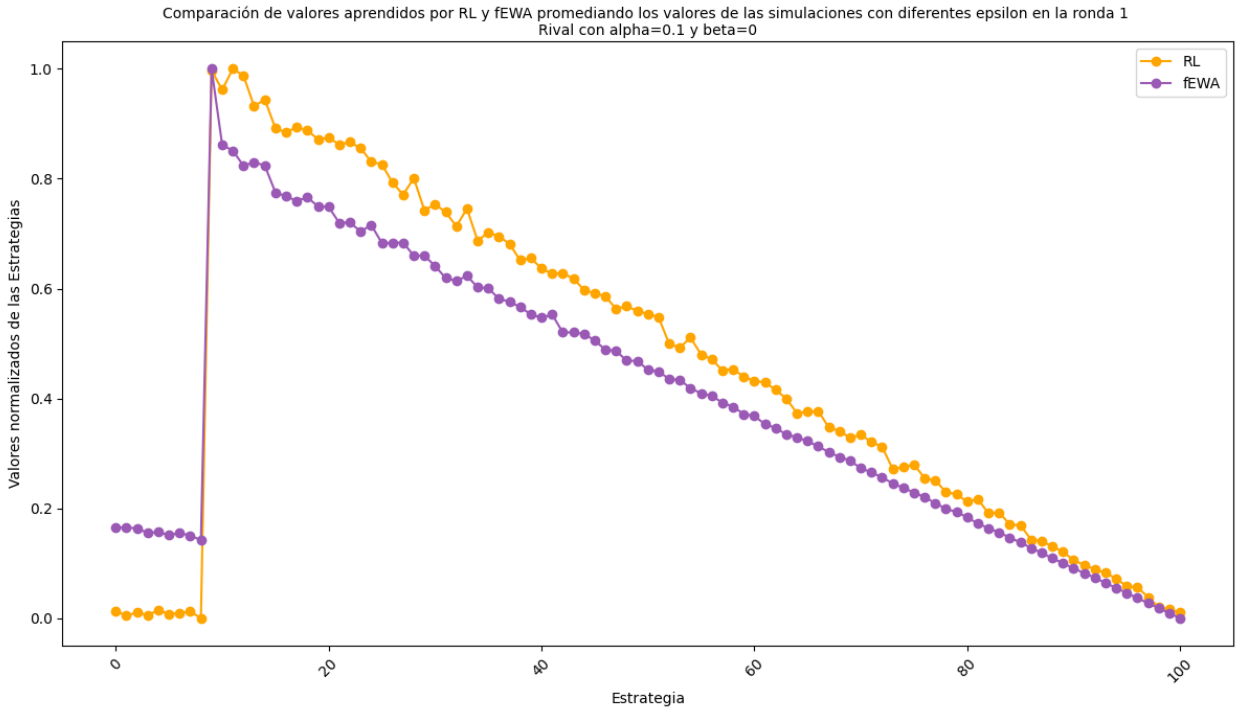

Figura 26: Valor normalizado del aprendizaje de estrategias de RL y FEWA en la ronda 1 contra el rival con preferencia baja por resultados ventajosos cuando se promedia ϵ.

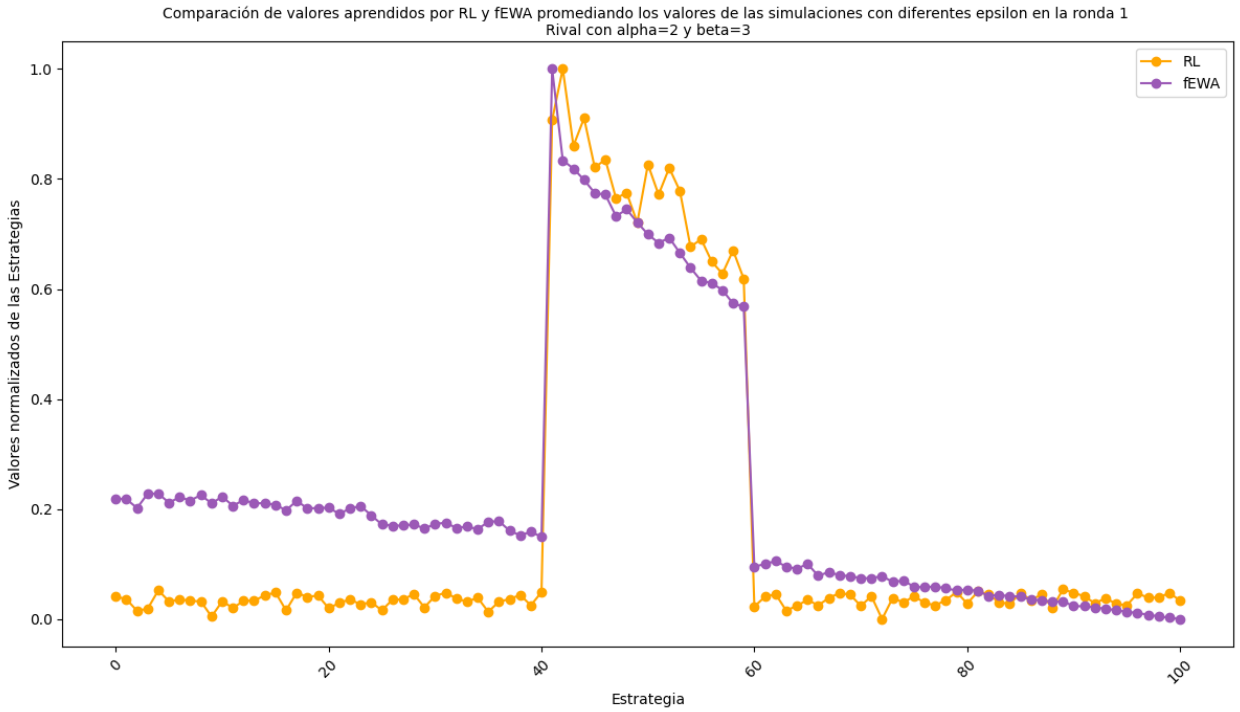

Figura 27: Valor normalizado del aprendizaje de estrategias de RL y FEWA en la ronda 1 contra el rival con preferencia media por resultados justos cuando se promedia ϵ.

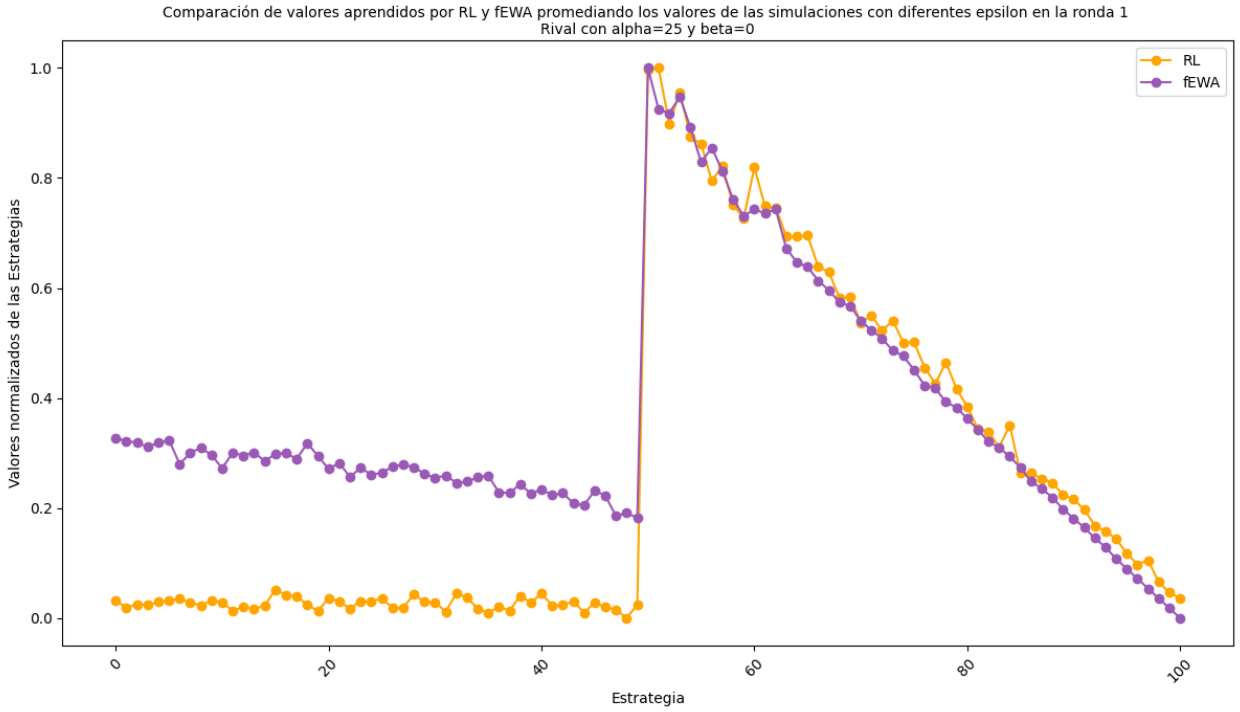

Figura 28: Valor normalizado del aprendizaje de estrategias de RL y FEWA en la ronda 1 contra el rival con preferencia alta por resultados ventajosos cuando se promedia ϵ.

Se recopilaron los resultados de los tres rivales de interés de la fase de aprendizaje, se corrieron 20 simulaciones por cada rival con tres epsilon diferentes, siendo 0.1, 0.5 y 0.999 las diferentes epsilon probadas, dando un total de 180 simulaciones, cada simulación consta de 2500 juegos de negociación contra el rival seleccionado. Los gráficos de cada epsilon para cada rival se presentan en el apartado de Anexos, los gráficos mostrados aquí promedian los resultados de los diferentes valores de epsilon de cada rival.

Cuando los valores no están normalizados FEWA asigna valores mucho más altos que RL a los conjuntos estados-acciones por ser un modelo acumulativo. Sin embargo, cuando se normalizan, la diferencia entre las estrategias aceptadas y no aceptadas por el rival son más notorias en para RL, esto debido al elemento BL de FEWA que le otorga valor incluso a estrategias que no fueron probadas. Además, la variabilidad entre las estrategias adyacentes no es tan notoria en FEWA como en RL, incluso cuando el epsilon es bajo y no ha habido una exploración tan uniforme para ambos modelos.

1. ¿Ambos agentes serán capaces de clasificar correctamente a un rival con un rango de ofertas aceptables conocido y de clasificar correctamente a un rival con un rango de ofertas aceptables desconocido?

    a) Tanto RL como FEWA serán capaces de clasificar correctamente rivales con un rango de ofertas aceptables que sea diferenciable y que hayan aprendido previamente. Sin embargo, solo el agente FEWA será capaz de identificar y aprender rivales con un rango de ofertas aceptables desconocido y diferenciable, debido a su parámetro de detección de cambio.

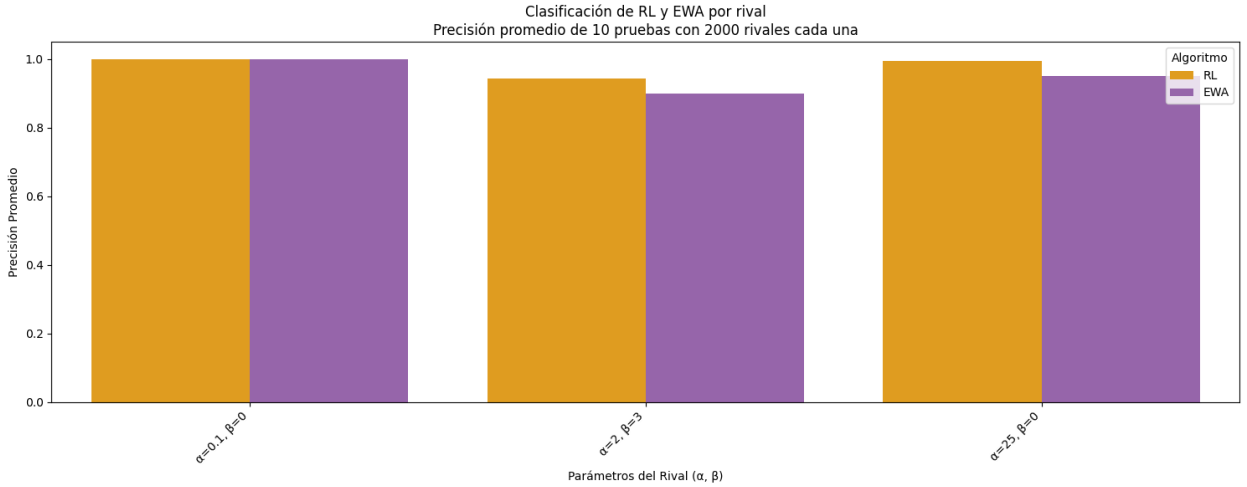

Figura 29: Precisión al clasificar 3 rivales cuando ambos agentes aprendieron 3 rivales.

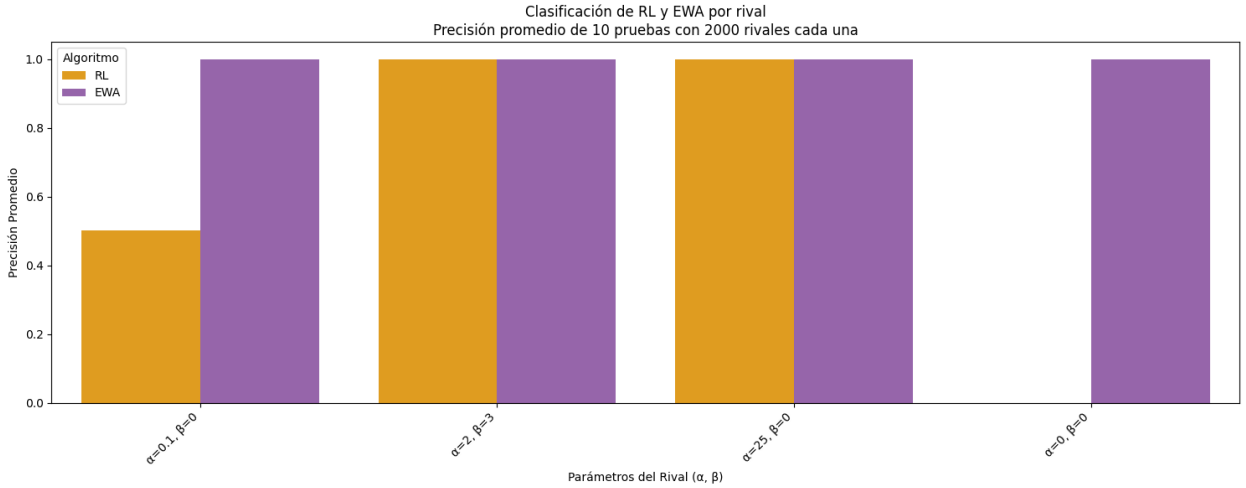

Figura 30: Precisión al clasificar 4 rivales cuando ambos agentes aprendieron 3 rivales y se agrega un rival nuevo con un rango de ofertas aceptable muy distinguible.

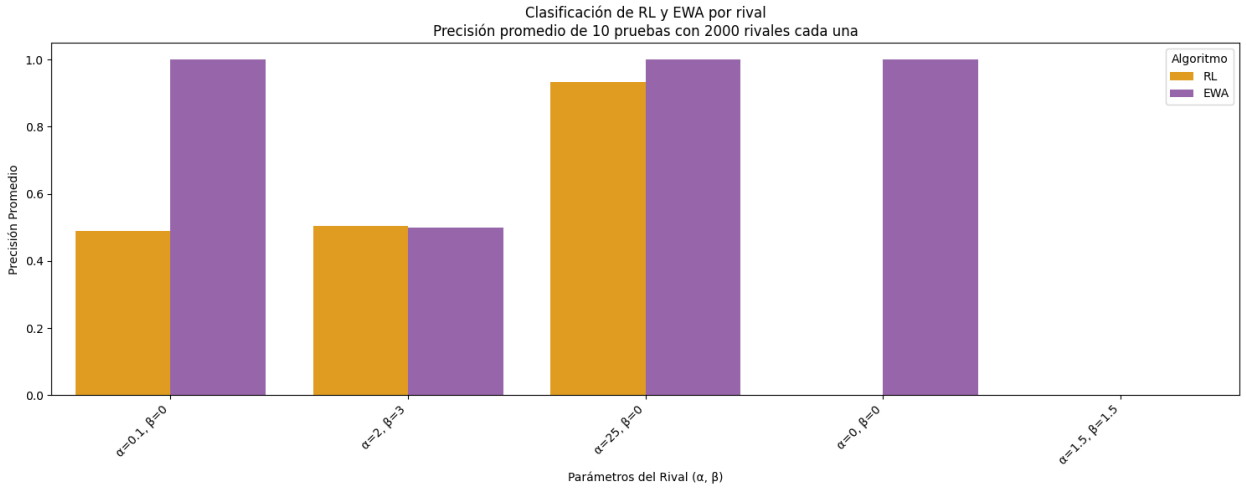

Figura 31: Precisión al clasificar 5 rivales cuando ambos agentes aprendieron 3 rivales y se agrega dos rivales nuevos, uno con un rango de ofertas aceptables muy distinguible y otro poco distinguible.

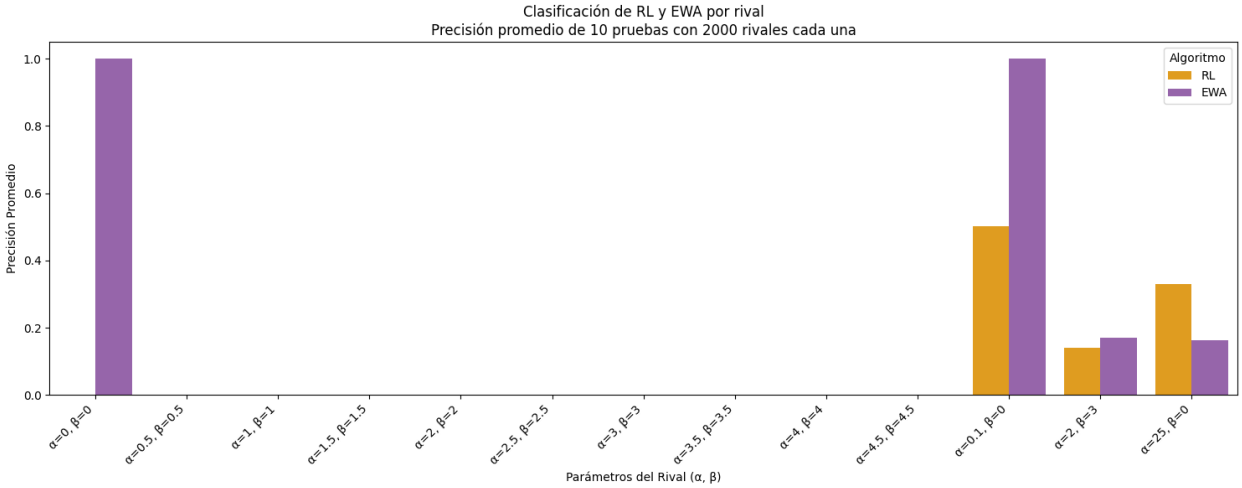

Figura 32: Precisión al clasificar 13 rivales cuando ambos agentes aprendieron 3 rivales.

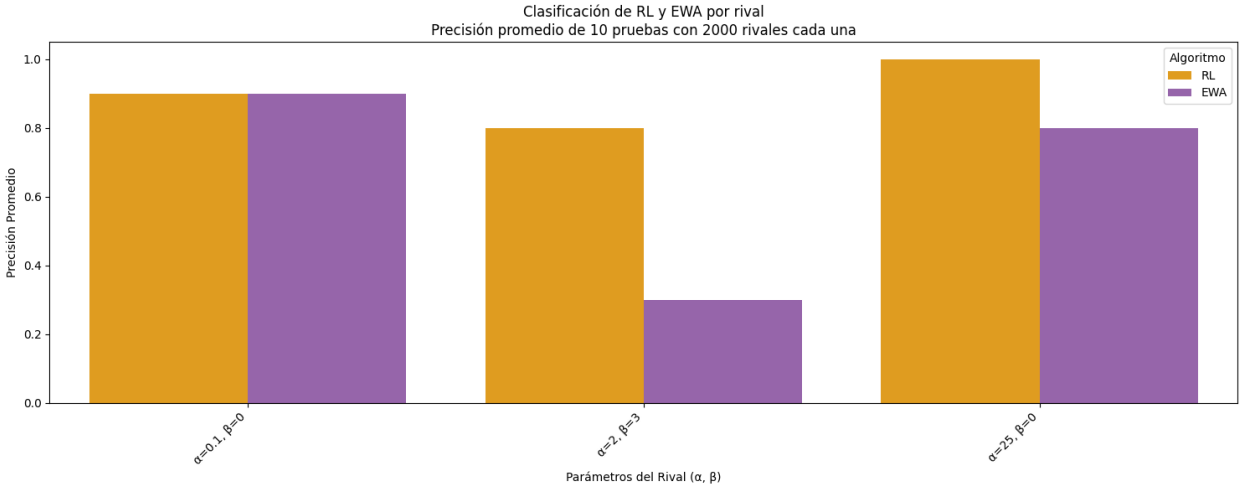

Figura 33: Precisión al clasificar 3 rivales cuando ambos agentes aprendieron 13 rivales.

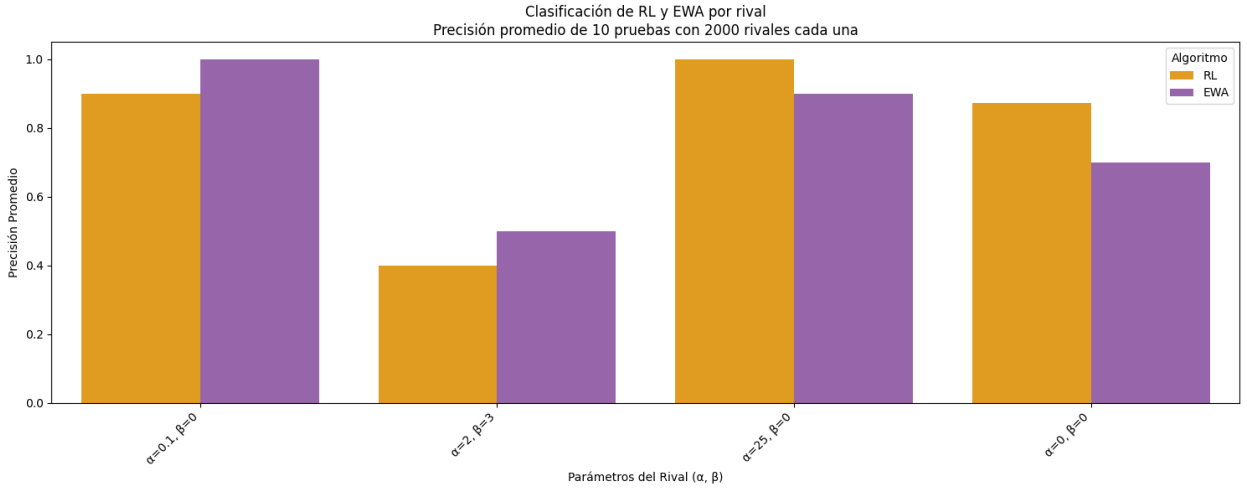

Figura 34: Precisión al clasificar 4 rivales cuando ambos agentes aprendieron 13 rivales y se agrega un rival nuevo con un rango de ofertas aceptable muy distinguible.

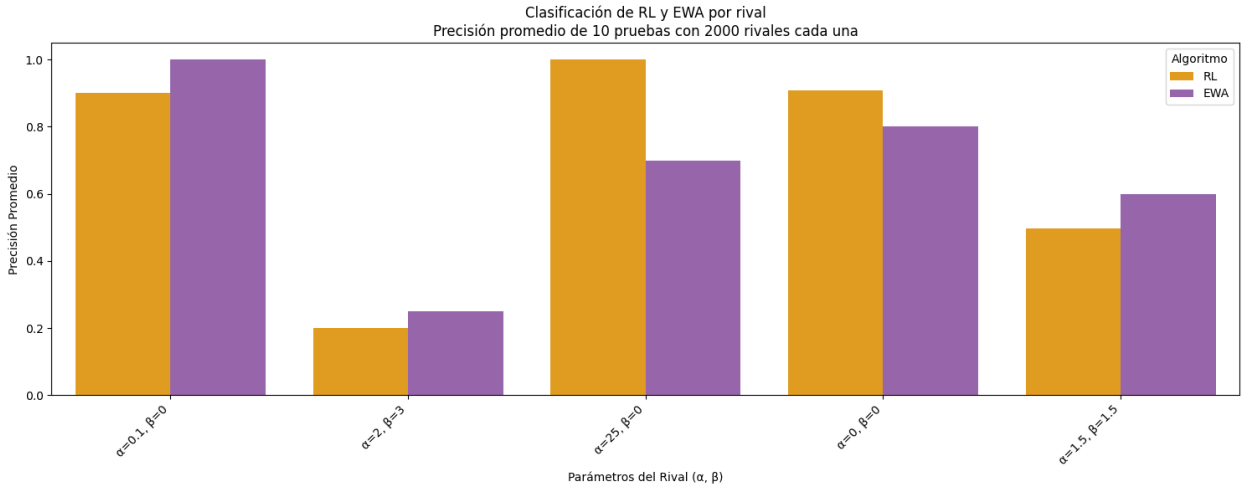

Figura 35: Precisión al clasificar 5 rivales cuando ambos agentes aprendieron 13 rivales y se agrega dos rivales nuevos, uno con un rango de ofertas aceptables muy distinguible y otro poco distinguible.

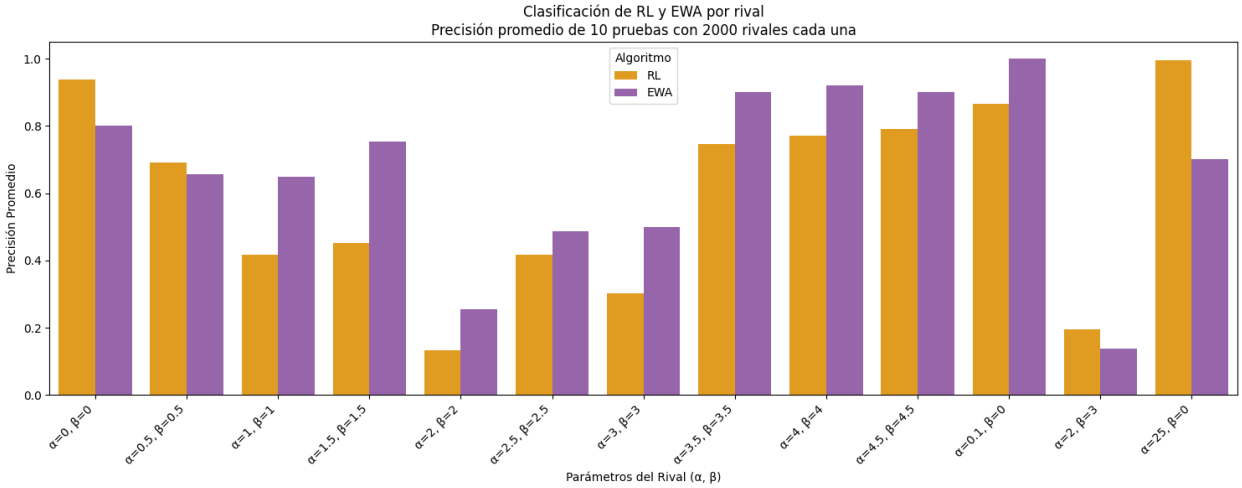

Figura 36: Precisión al clasificar 13 rivales cuando ambos agentes aprendieron 13 rivales.

Se recopilaron los resultados de 10 simulaciones, en cada una los agentes RL y FEWA debían de jugar contra un grupo de 2000 rivales, las preferencias de los rivales (valores en los parámetros del modelo IA) contra los que jugaron están en los gráficos, todos los rivales siempre tenían probabilidad uniforme de aparecer. Los rivales que RL y FEWA aprendieron con anticipación también se encuentran en los gráficos, estos únicamente variaban entre los tres rivales de la fase de aprendizaje (preferencia baja por resultados ventajosos, preferencia media por resultados justos, preferencia alta por resultados ventajosos) o todos los tipos de rivales de prueba (13 rivales). Los resultados de precisión, recall y f1-score son el promedio de las 10 simulación, dando un total de 32 simulaciones de aprendizaje y 80 simulaciones para probar la categorización.

En general, cuando los agentes aprenden únicamente 3 rivales y deben clasificar 3 rivales la precisión de ambos es alta clasificando a todos los rivales. Cuando los agentes aprenden únicamente 3 rivales y deben clasificar 4 rivales, uno nuevo pero muy reconocible, la precisión de FEWA es alta para todos los rivales, la de RL es nula con el nuevo rival y disminuye su precisión con el rival con preferencia baja a resultados ventajosos porque clasifica al nuevo rival como uno de ese tipo. Cuando los agentes aprenden únicamente 3 rivales y deben clasificar 5 rivales, dos nuevos, uno muy reconocible y otro poco reconocible, la precisión de ambos agentes con el rival poco reconocible es nula porque es confundido con el rival con preferencia media a resultados justos, además, la precisión de este rival para ambos agentes también baja por esta confusión.

Cuando los agentes aprenden los 13 rivales y deben clasificar 3 rivales, su precisión dismi-

nuye, sobre todo con el rival con preferencia media a resultados justos, ya que es el rival con el rango de ofertas más similar a otros. Cuando los agentes aprenden los 13 rivales y deben clasificar 4 o 5 rivales, la precisión de ambos va disminuyendo conforme crecen los valores en los parámetros de los rivales, ya que son confundidos con el rival con preferencia media a resultados justos.

Sin embargo, la diferencia de clasificación es muy notoria cuando se enfrentan a los 13 rivales y aprendieron a 13 rivales en comparación a cuando solo aprendieron 3 rivales, cuando a aprenden 13 rivales, el rival con preferencia media a resultados justos es confundido con rivales con valores adyacentes en los parámetros, sin embargo, la precisión de ambos agentes a rivales con valores en los parámetros muy bajos o muy altos es buena, particularmente la de FEWA, esto se debe a que las estrategias óptimas que aprende este agente son más cercana a los SPNE que las que aprende el agente RL.

Por otra parte, cuando solo aprenden 3 rivales y deben clasificar a 13, RL es incapaz de distinguir a todos los rivales desconocidos, y FEWA es incapaz de distinguir a todos los rivales desconocidos excepto a uno, al que tiene un rango de ofertas aceptables muy distinguible, es decir, al rival con preferencia por resultados eficientes, y esto se debe a la oferta adicional que hace desde el inicio FEWA contemplando sus valores a priori por introspección pre-juego. Sin embargo, los resultados de esta prueba se deben a que, para ambos agentes, todos los rivales con valores en los parámetros medios-bajos son confundidos con el rival con preferencia media a resultados justos, y todos los rivales con valores en los parámetros altos con confundidos con el rival con preferencia alta a resultados ventajosos. Si somos más laxos con el umbral de FEWA, para que con menos errores identifique a un rival nuevo, desafortunadamente esto provocaría muchas falsas alarmas.

Este mismo patrón de confusión y reconocimiento de rivales dependiendo de los agentes aprendidos descrito para los gráficos de Presición también puede ser observado en los gráficos Recall y F1-score en la sección de Anexos.

6.4. Agent Based

1. ¿En qué se diferenciará el desempeño de los modelos de aprendizaje en el sistema AB si en la población solo hay rivales aprendidos previamente?

    a) Entre más similares sean las estrategias óptimas aprendidas por los agentes RL y

FEWA serán más similares las ganancias finales de ambos agentes y viceversa.

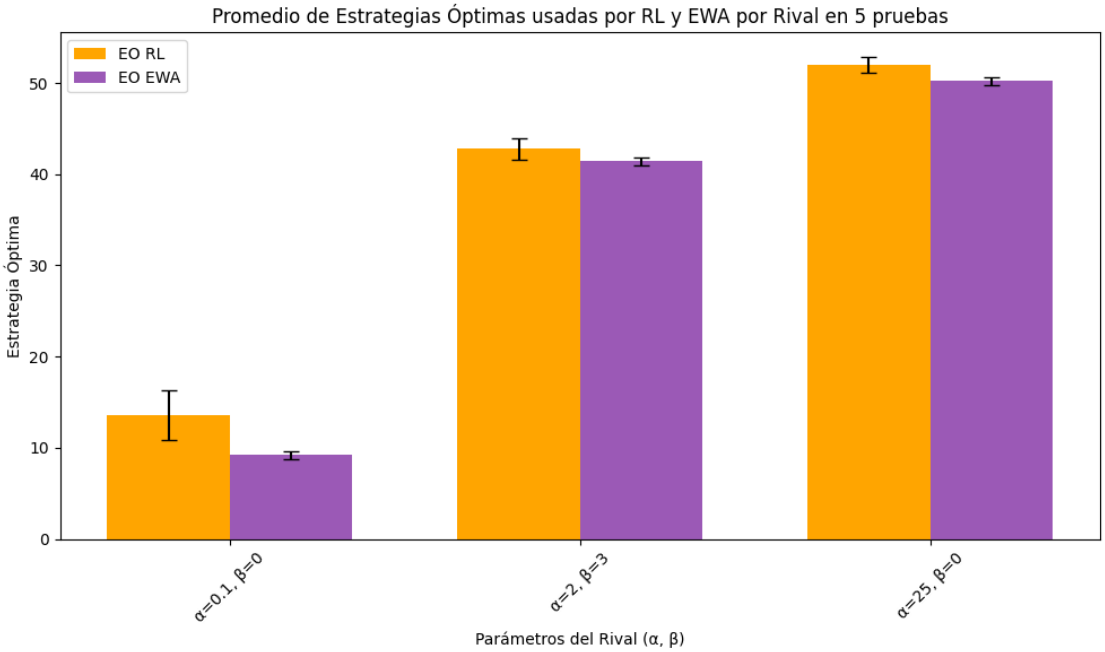

Figura 37: Proporción de estrategias usadas por los agentes RL y FEWA cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA donde todos tienen una proporción uniforme.

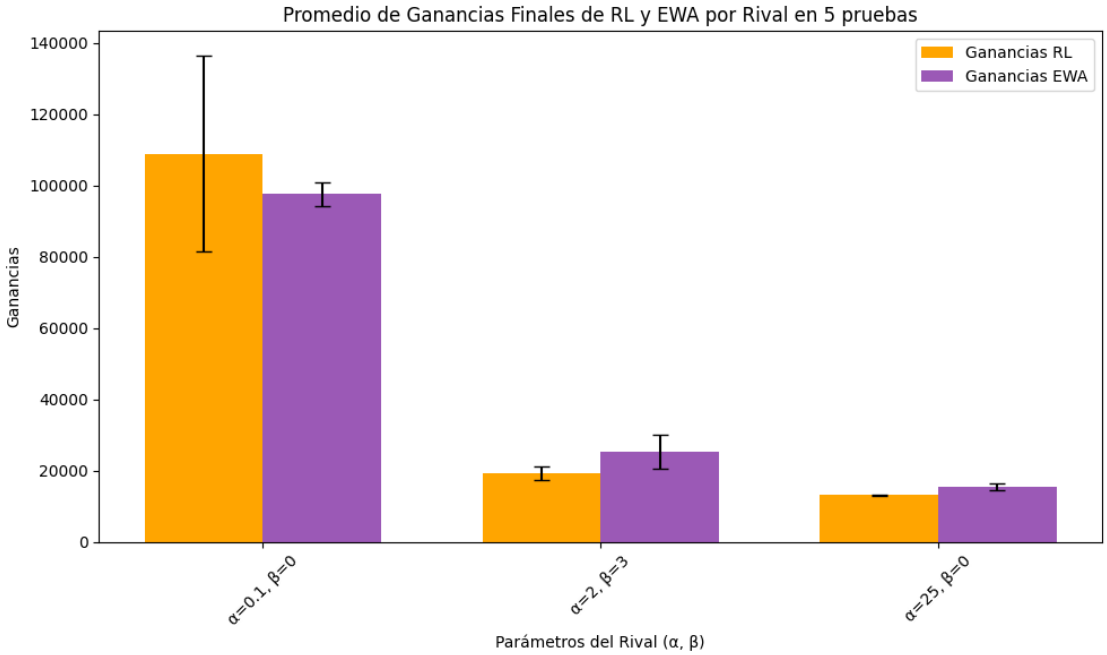

Figura 38: Ganancias que obtivieron RL y FEWA por cada grupo cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA donde todos tienen una proporción uniforme.

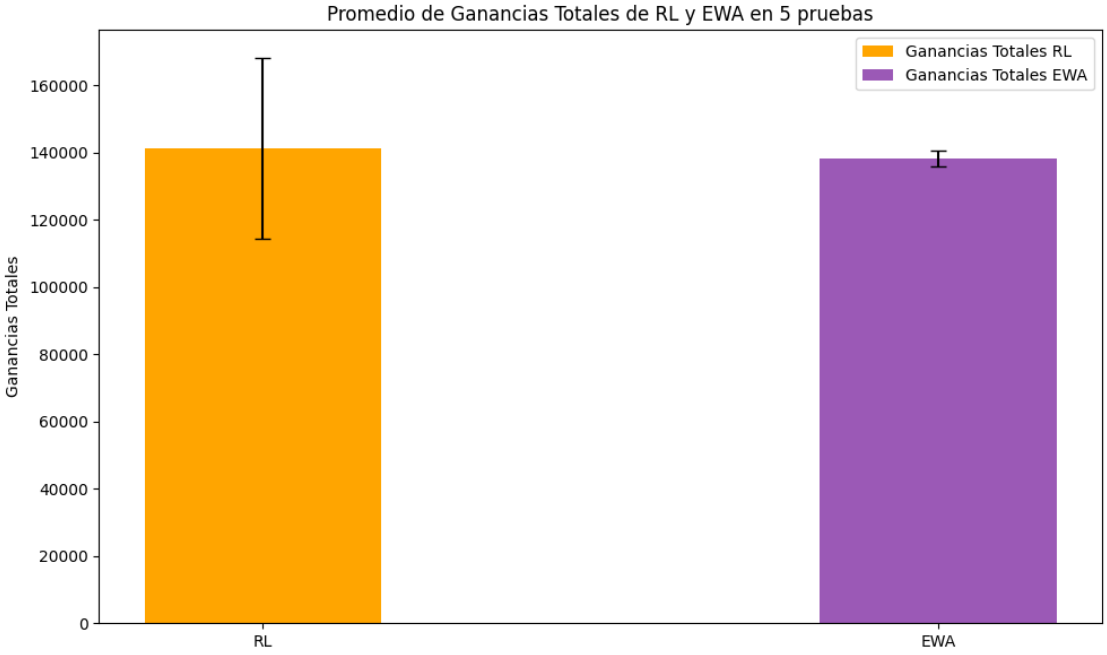

Figura 39: Ganancias totales que obtivieron RL y FEWA cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA donde todos tienen una proporción uniforme.

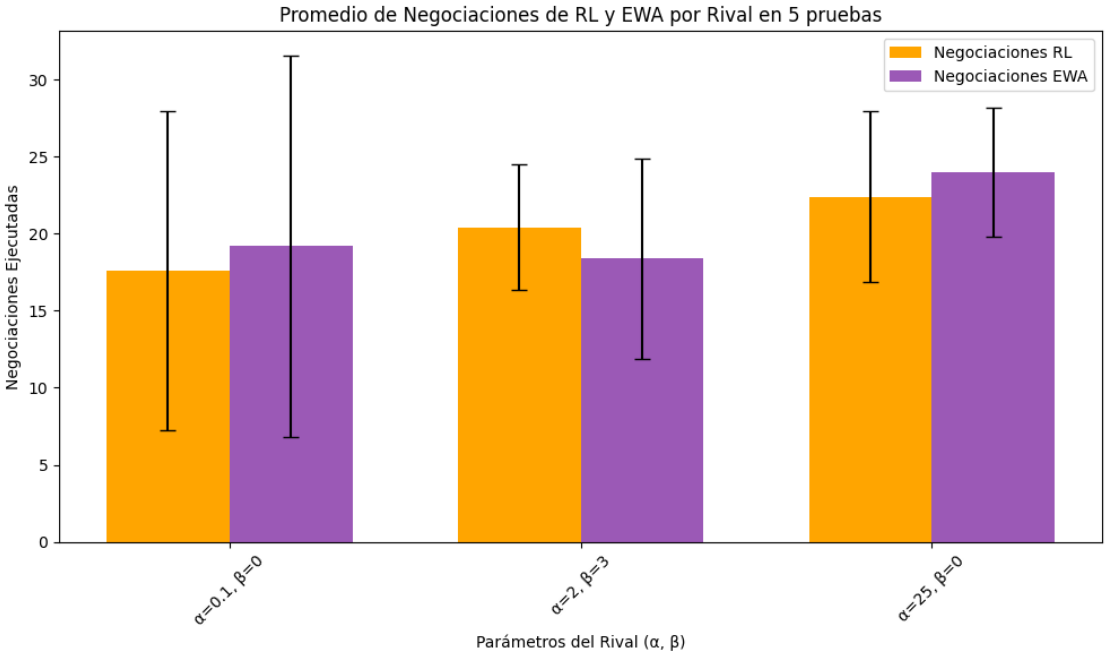

Figura 40: Negociaciones que llevaron a cabo RL y FEWA por cada grupo cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA donde todos tienen una proporción uniforme.

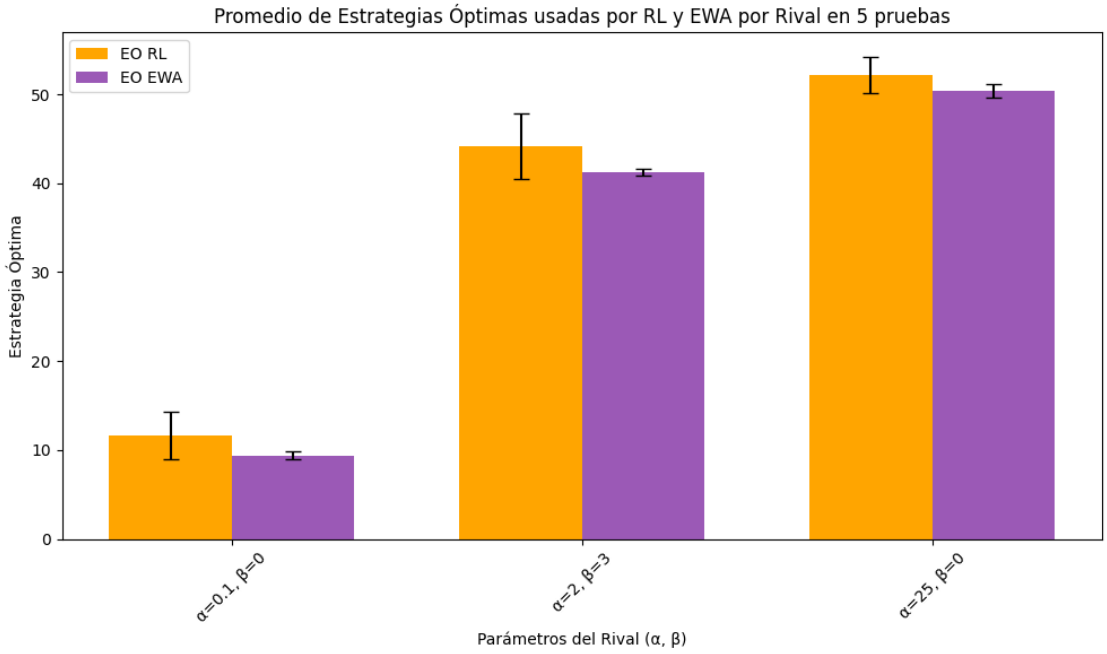

Figura 41: Estrategias usadas por los agentes RL y FEWA cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia alta a resultados ventajosos tienen mayoría.

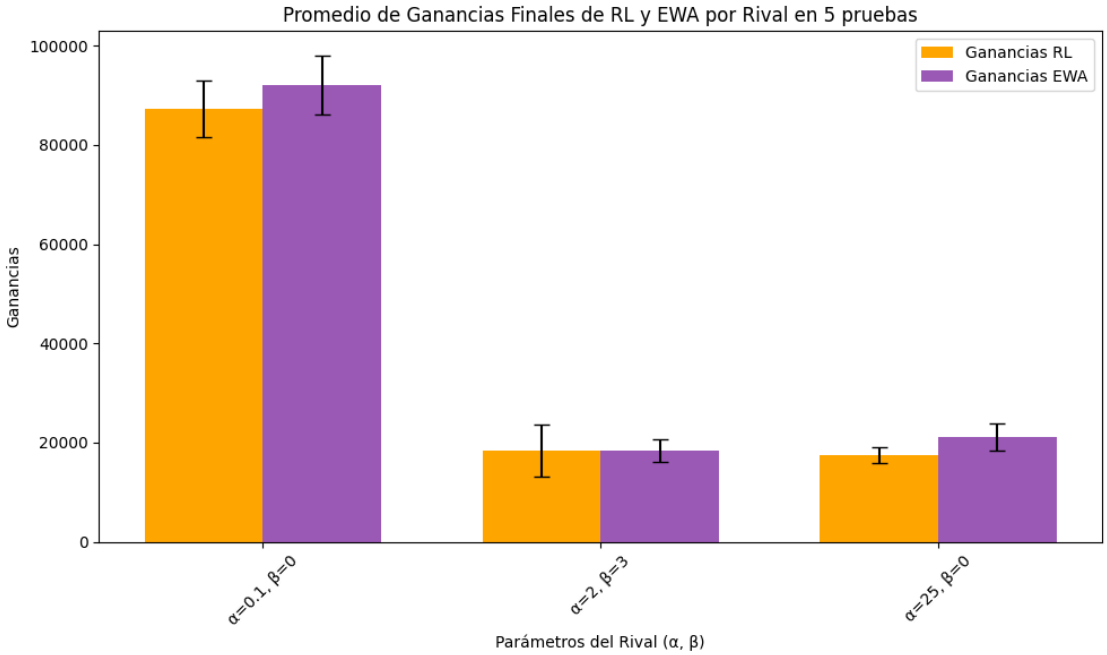

Figura 42: Ganancias que obtivieron RL y FEWA por cada grupo cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia alta a resultados ventajosos tienen mayoría.

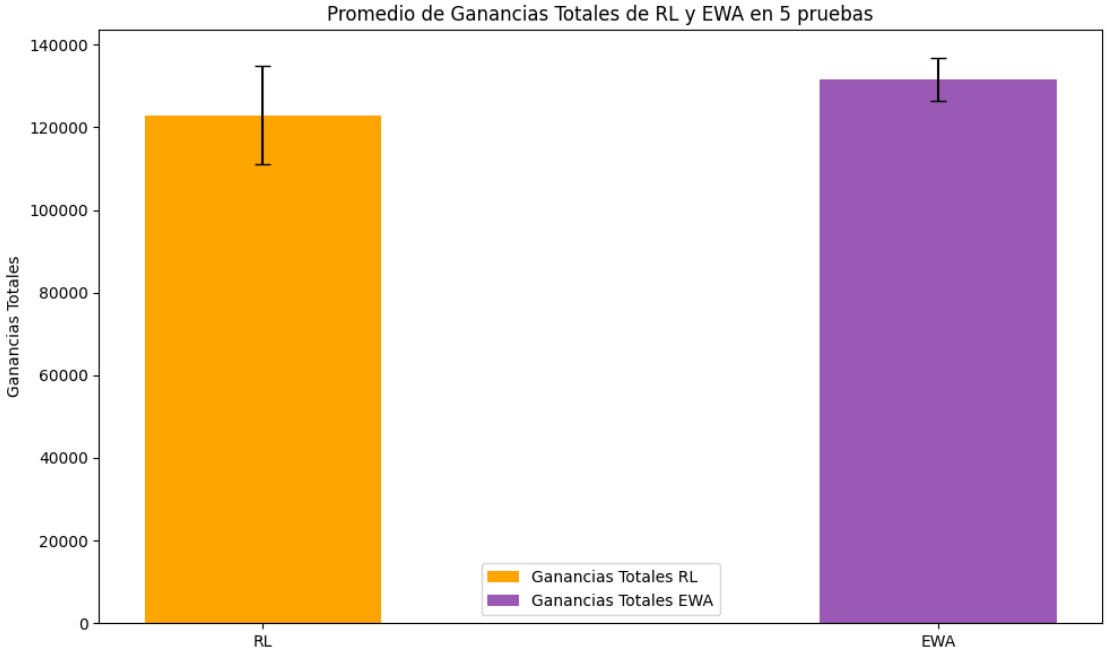

Figura 43: Ganancias totales que obtivieron RL y FEWA cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia alta a resultados ventajosos tienen mayoría.

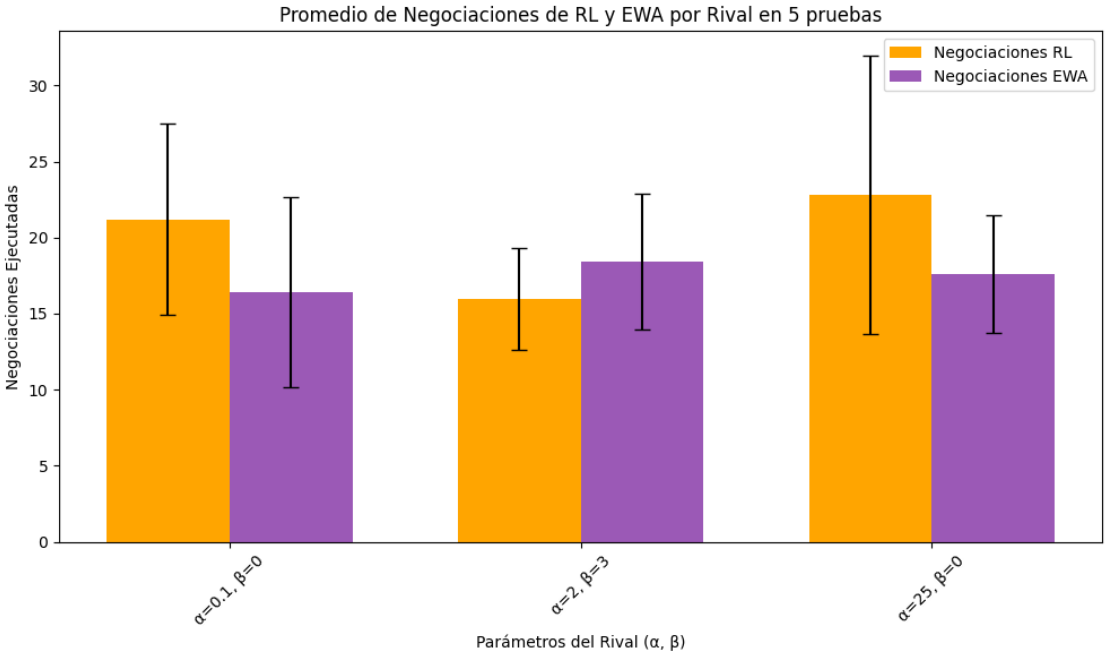

Figura 44: Negociaciones que llevaron a cabo RL y FEWA por cada grupo cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia alta a resultados ventajosos tienen mayoría.

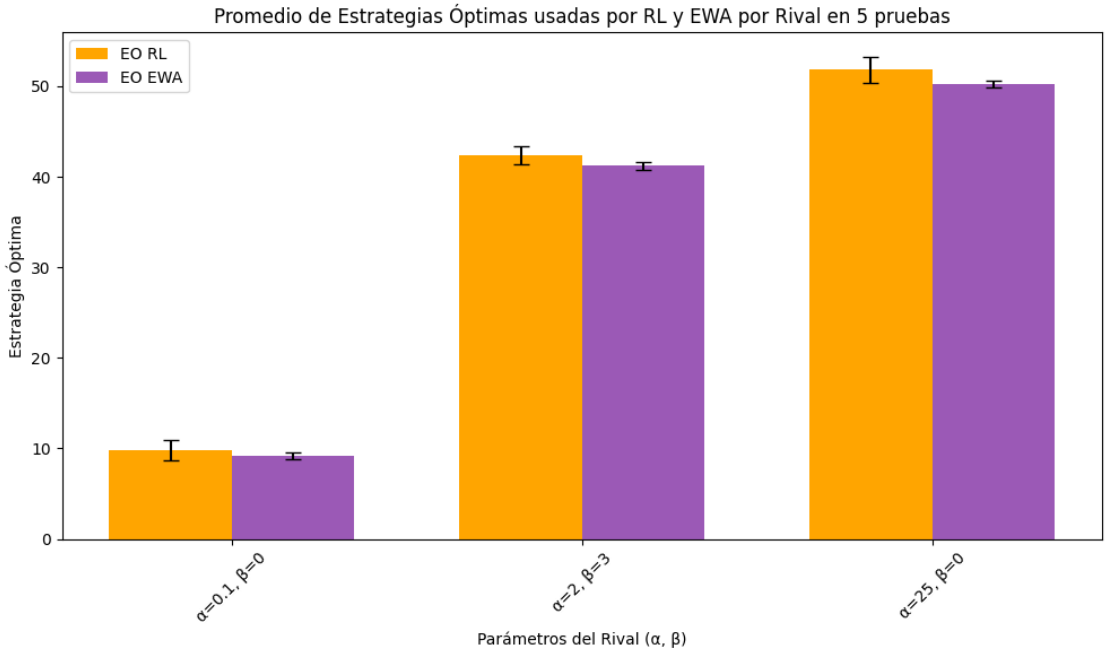

Figura 45: Estrategias usadas por los agentes RL y FEWA cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia baja a resultados ventajosos tienen mayoría.

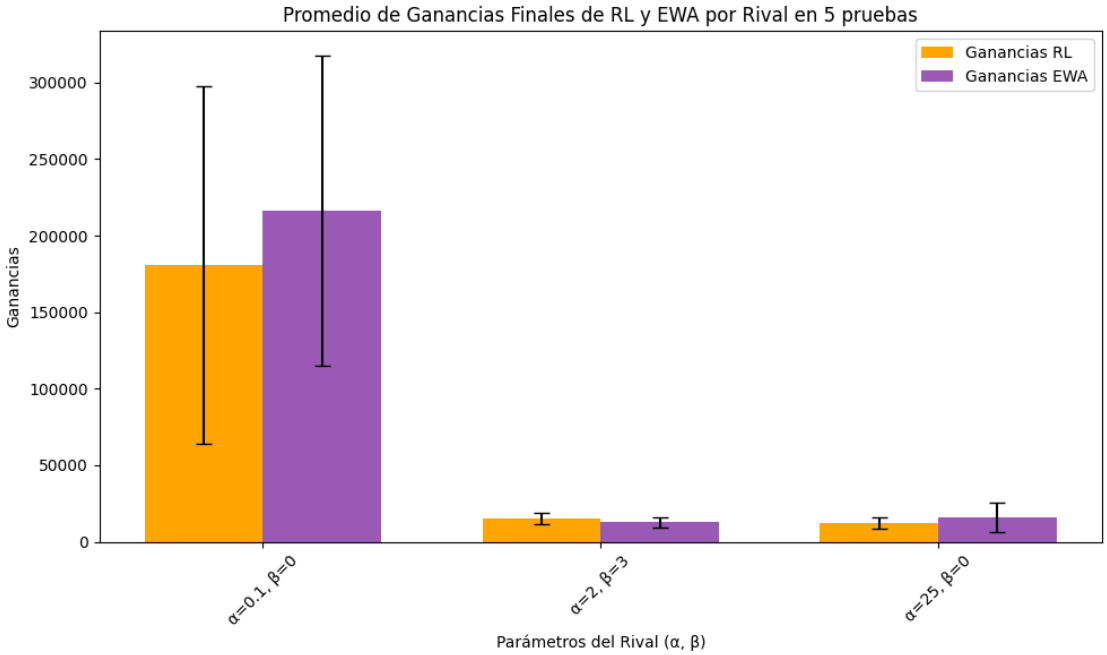

Figura 46: Ganancias que obtivieron RL y FEWA por cada grupo cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia baja a resultados ventajosos tienen mayoría.

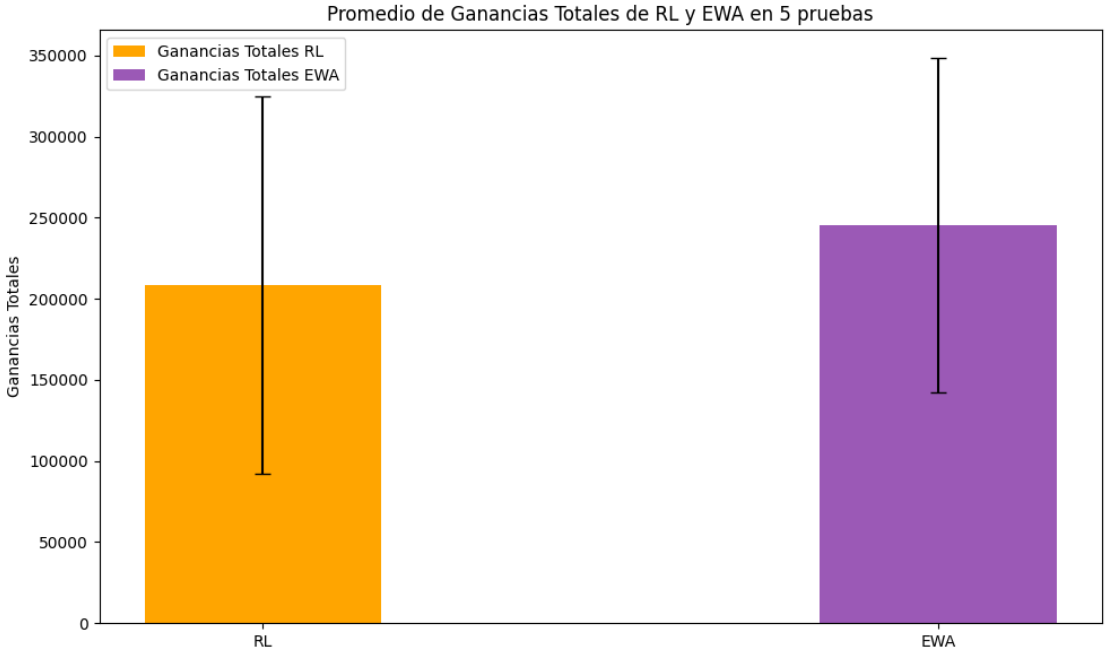

Figura 47: Ganancias totales que obtivieron RL y FEWA por cada grupo cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia baja a resultados ventajosos tienen mayoría.

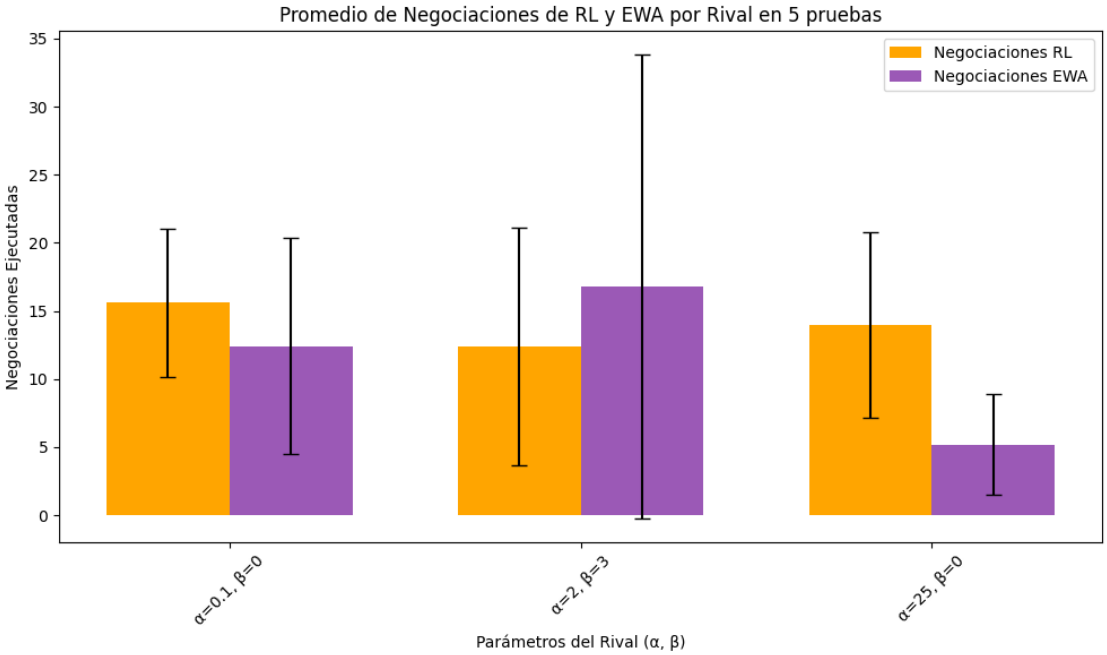

Figura 48: Negociaciones que llevaron a cabo RL y FEWA por cada grupo cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia baja a resultados ventajosos tienen mayoría.

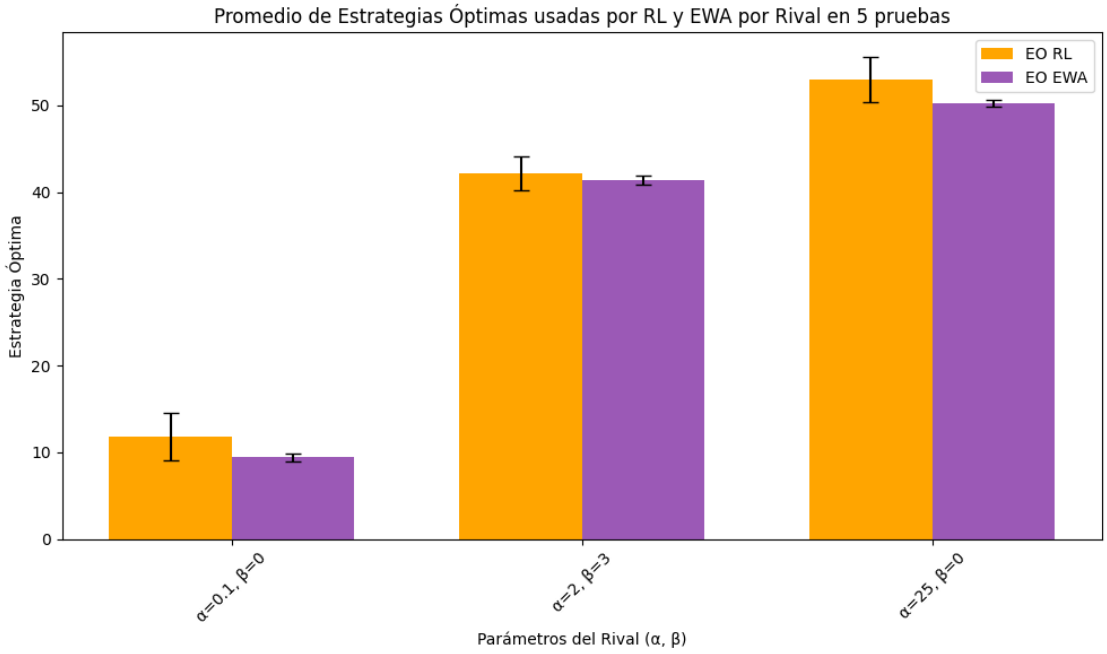

Figura 49: Estrategias usadas por los agentes RL y FEWA cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia media a resultados justos tienen mayoría.

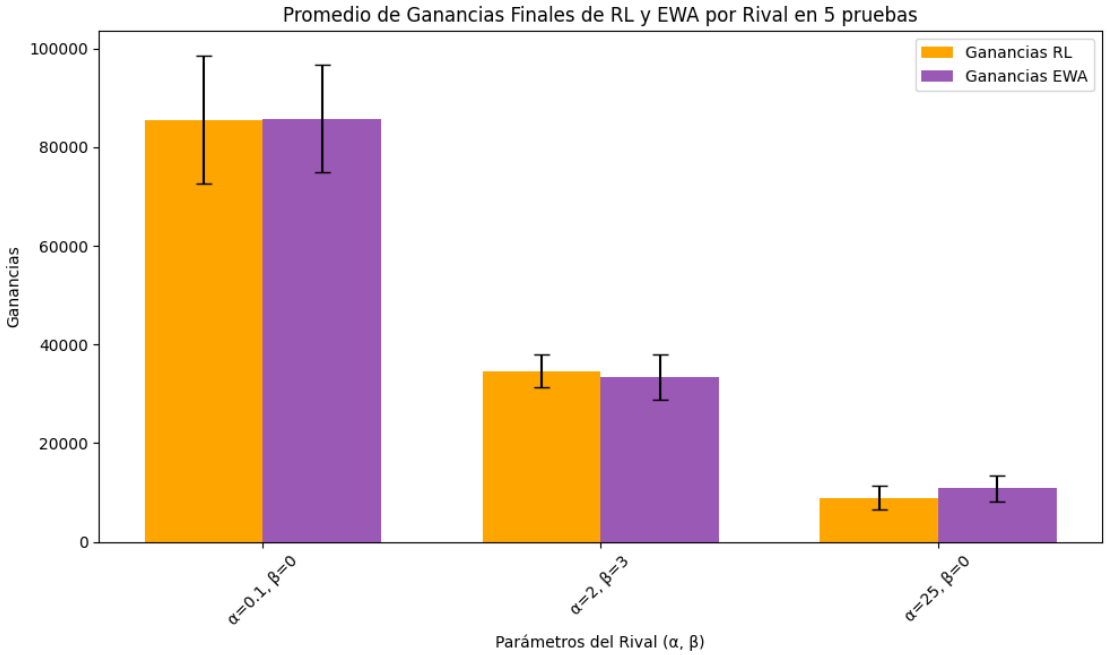

Figura 50: Ganancias que obtivieron RL y FEWA por cada grupo cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia media a resultados justos tienen mayoría.

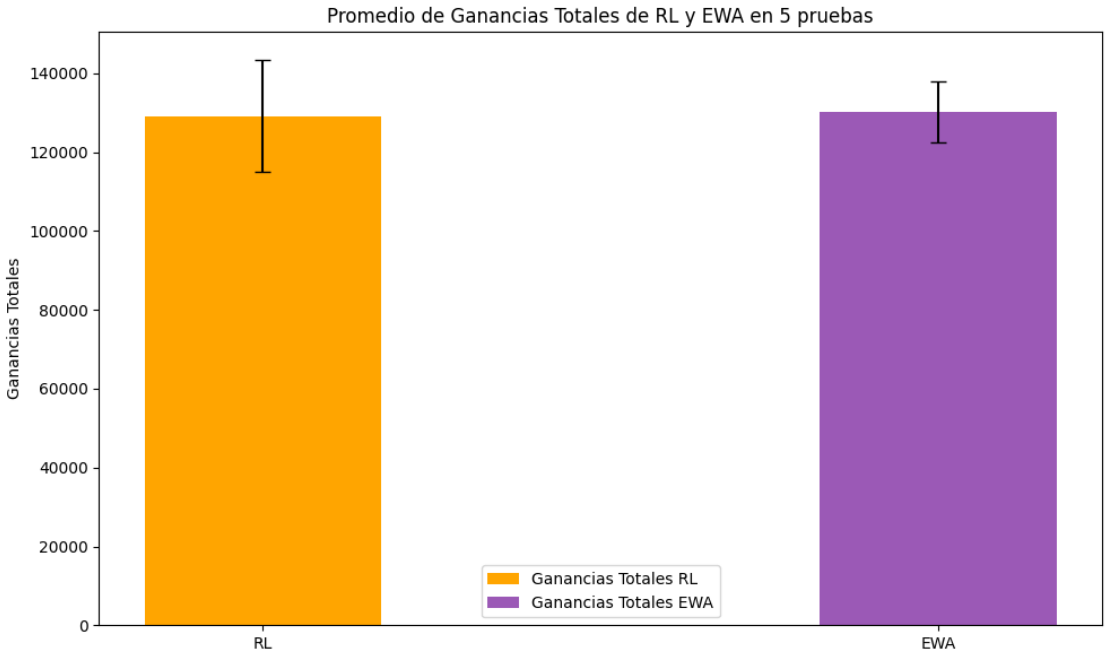

Figura 51: Ganancias totales que obtivieron RL y FEWA cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia media a resultados justos tienen mayoría.

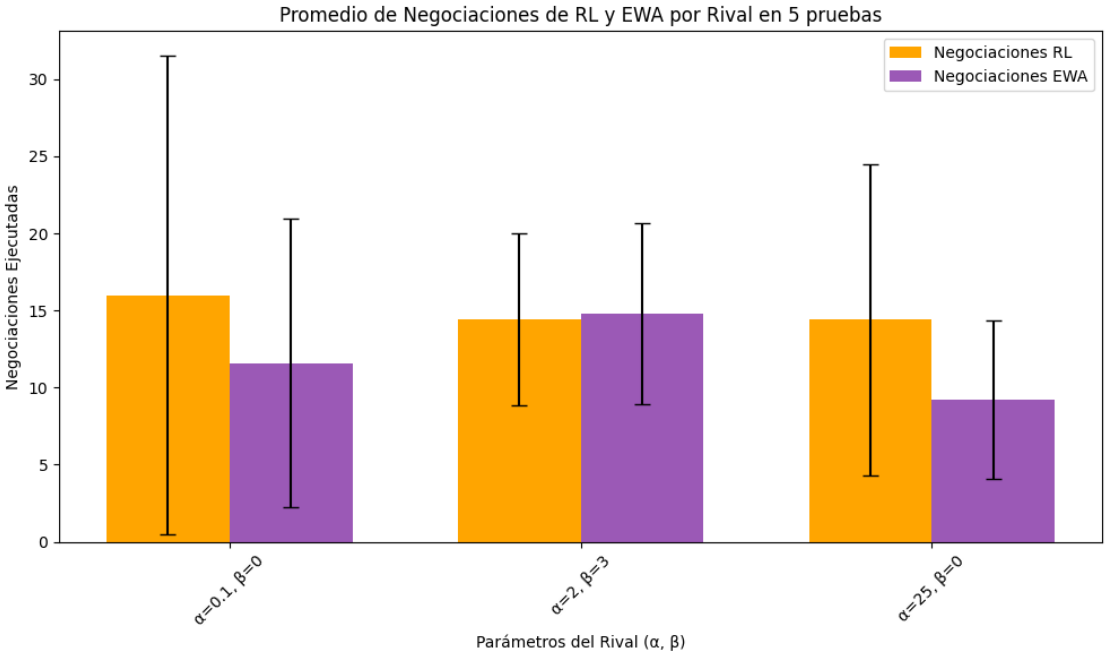

Figura 52: Negociaciones que llevaron a cabo RL y FEWA por cada grupo cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia media a resultados justos tienen mayoría.

1. ¿En qué se diferenciará el desempeño de los modelos si en la población se incluye a un rival que no fue aprendido previamente por los modelos, que ofrece más ganancias y tiene un rango de ofertas aceptables diferenciable?

    a) FEWA será el único capaz de reconocer a los rivales que no fueron aprendidos previamente y, debido a que este nuevo rival ofrece más ganancias, FEWA obtendrá más ganancias que RL al terminar la simulación.

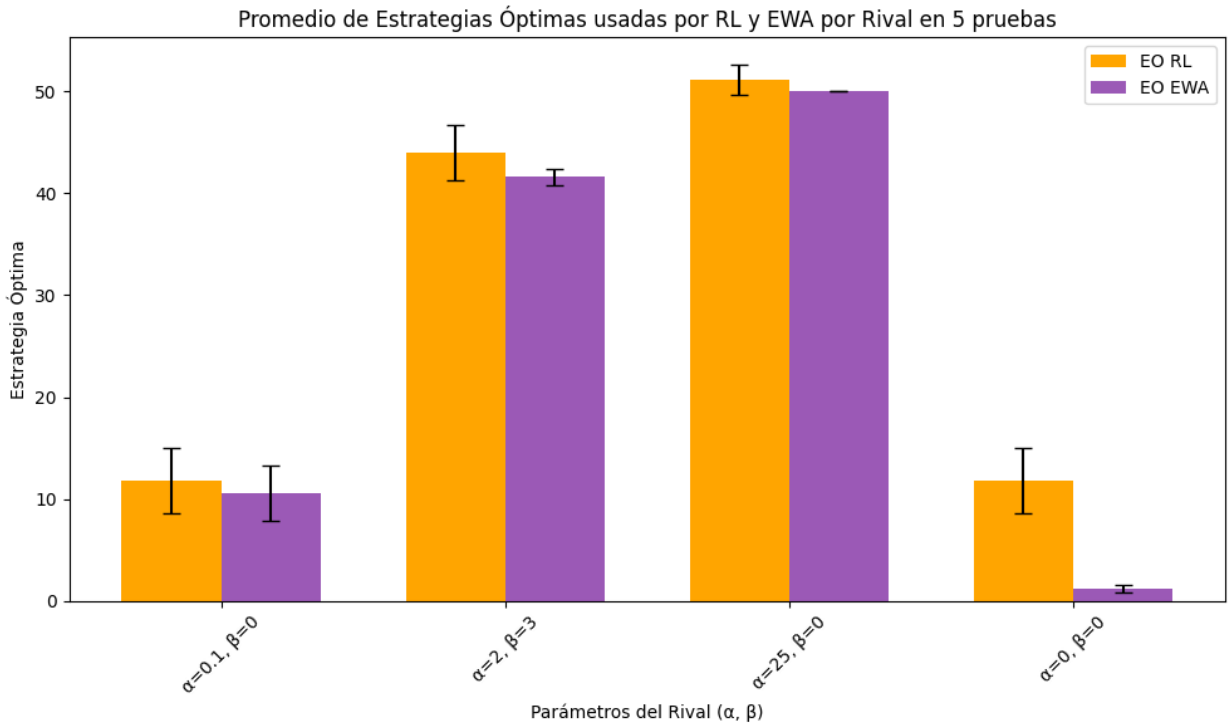

Figura 53: Estrategias usadas por los agentes RL y FEWA cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde todos tienen una proporción uniforme.

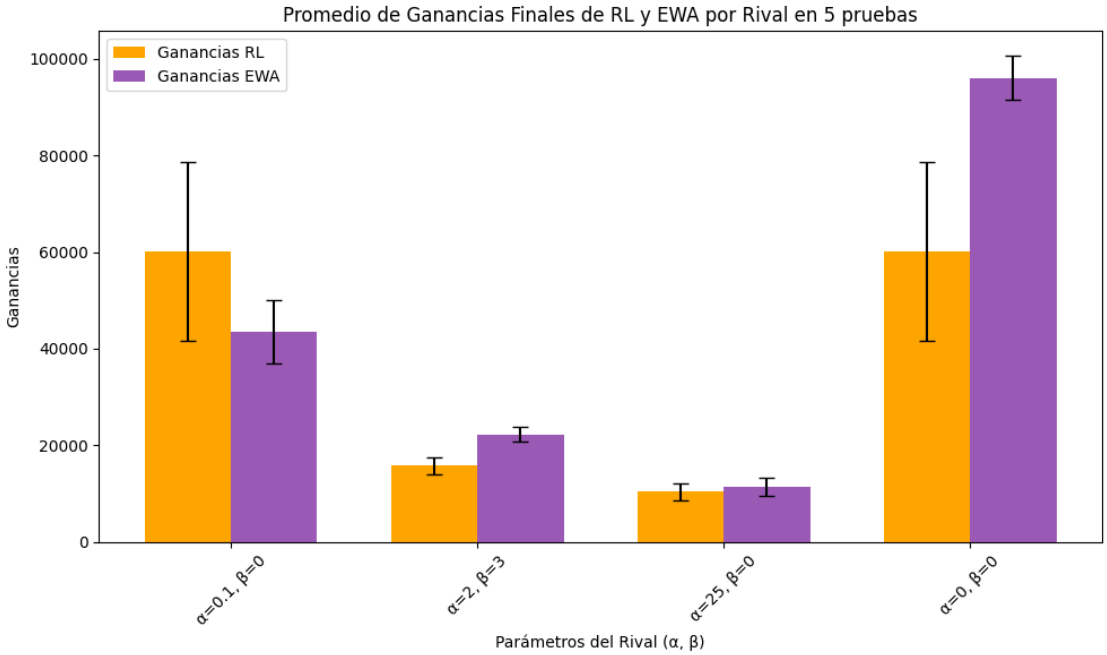

Figura 54: Ganancias que obtivieron RL y FEWA por cada grupo cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde todos tienen una proporción uniforme.

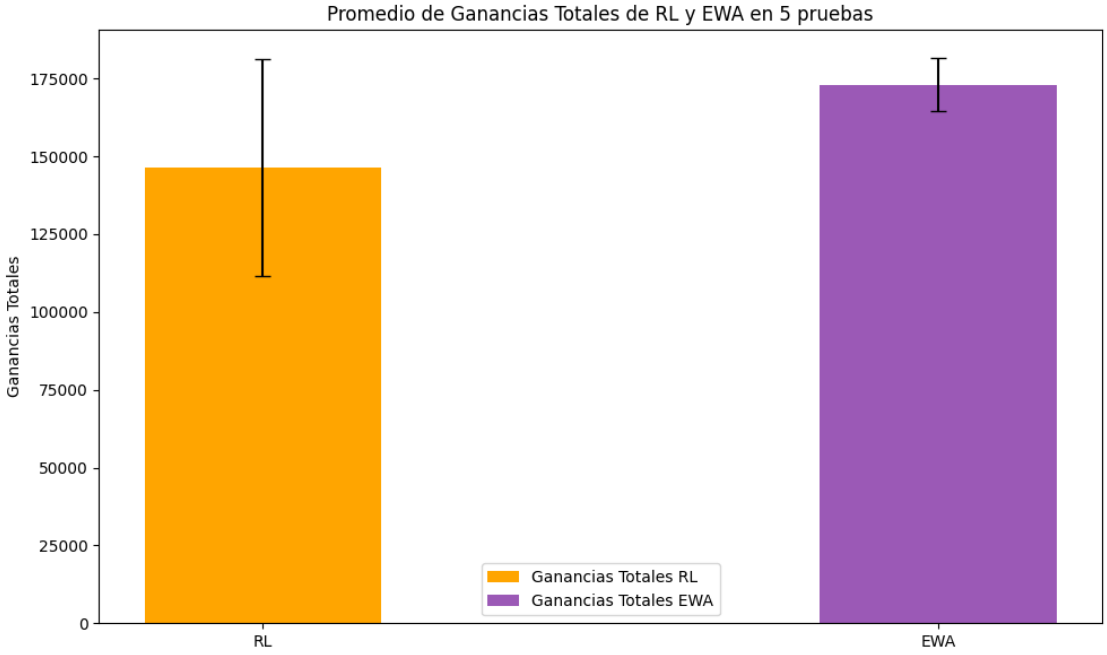

Figura 55: Ganancias totales que obtivieron RL y FEWA cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde todos tienen una proporción uniforme.

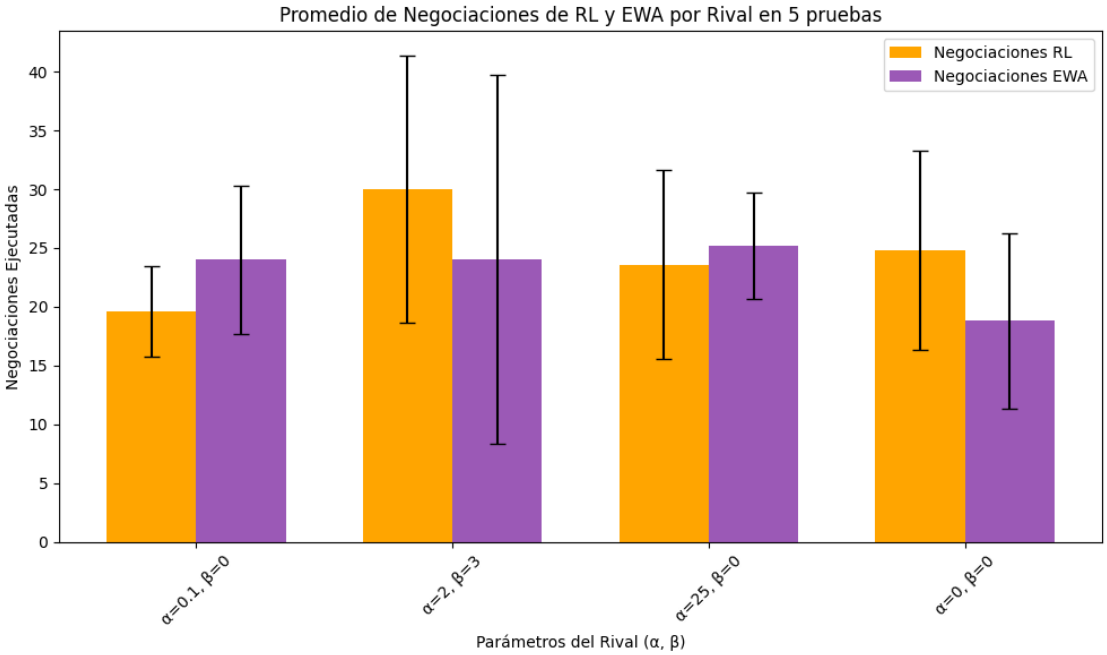

Figura 56: Negociaciones que llevaron a cabo RL y FEWA por cada grupo cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde todos tienen una proporción uniforme.

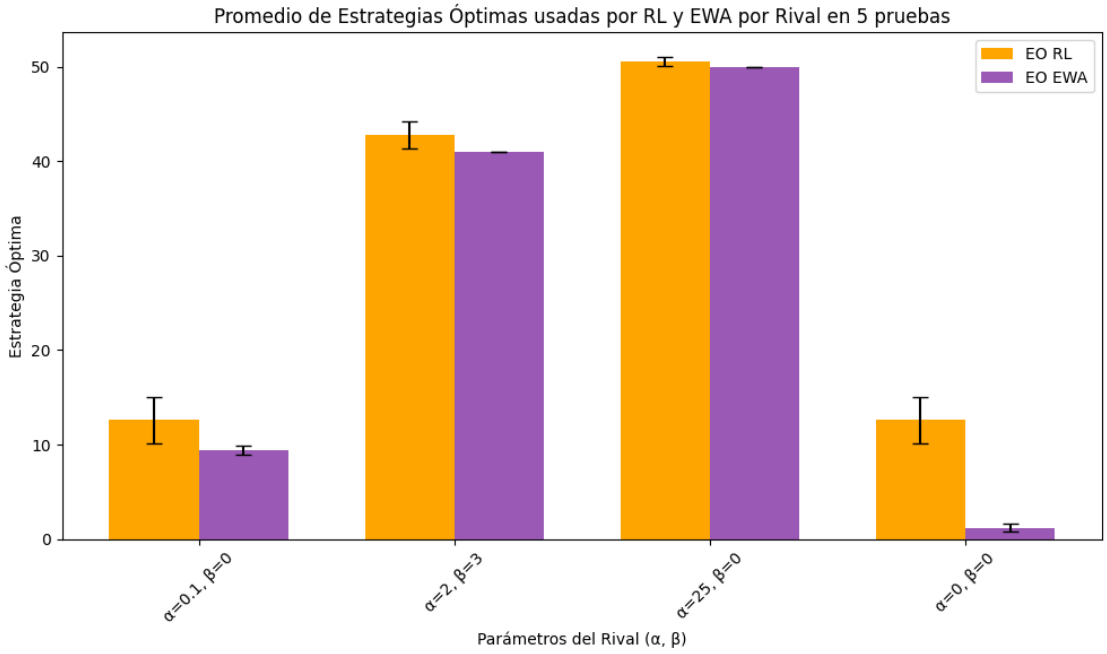

Figura 57: Estrategias usadas por los agentes RL y FEWA cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene minoría.

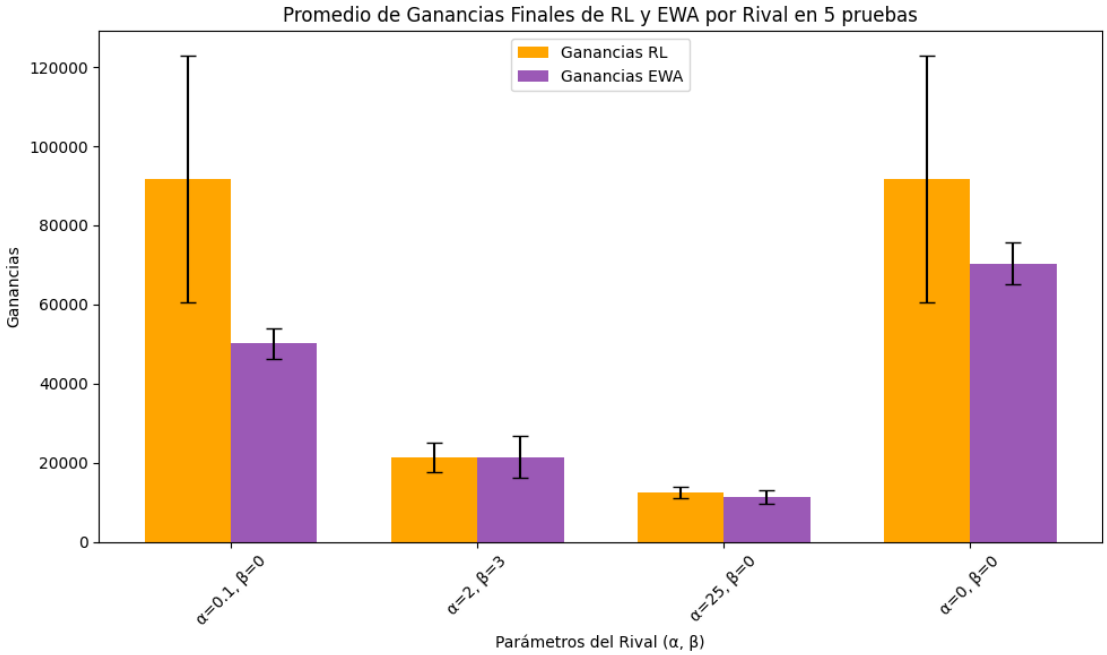

Figura 58: Ganancias que obtivieron RL y FEWA por cada grupo cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene minoría.

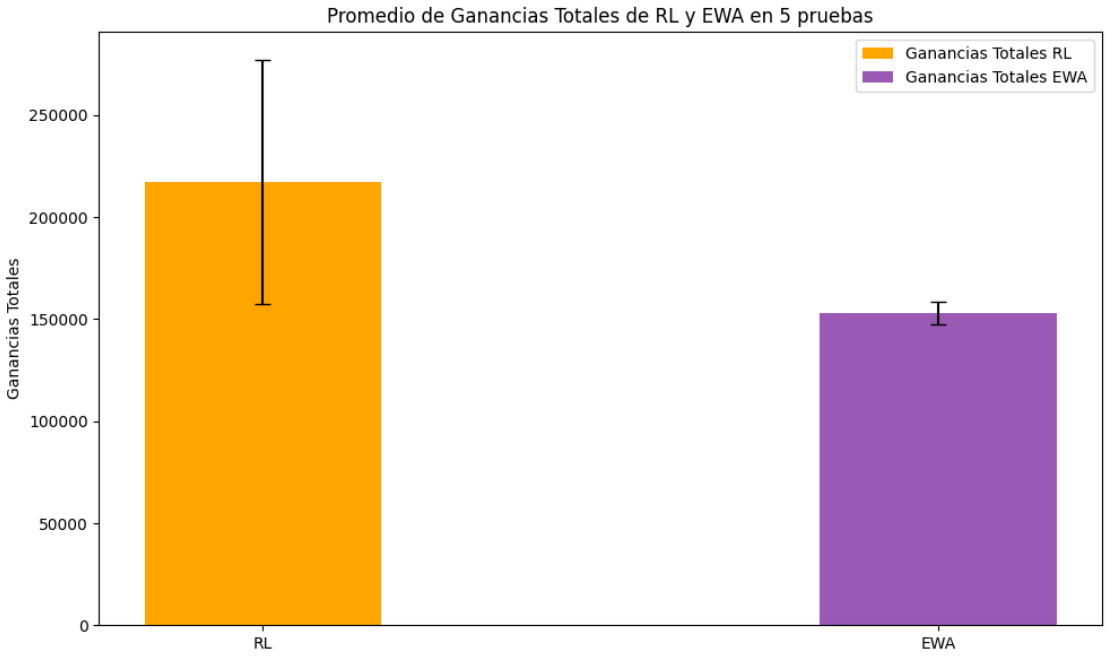

Figura 59: Ganancias totales que obtivieron RL y FEWA cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene minoría.

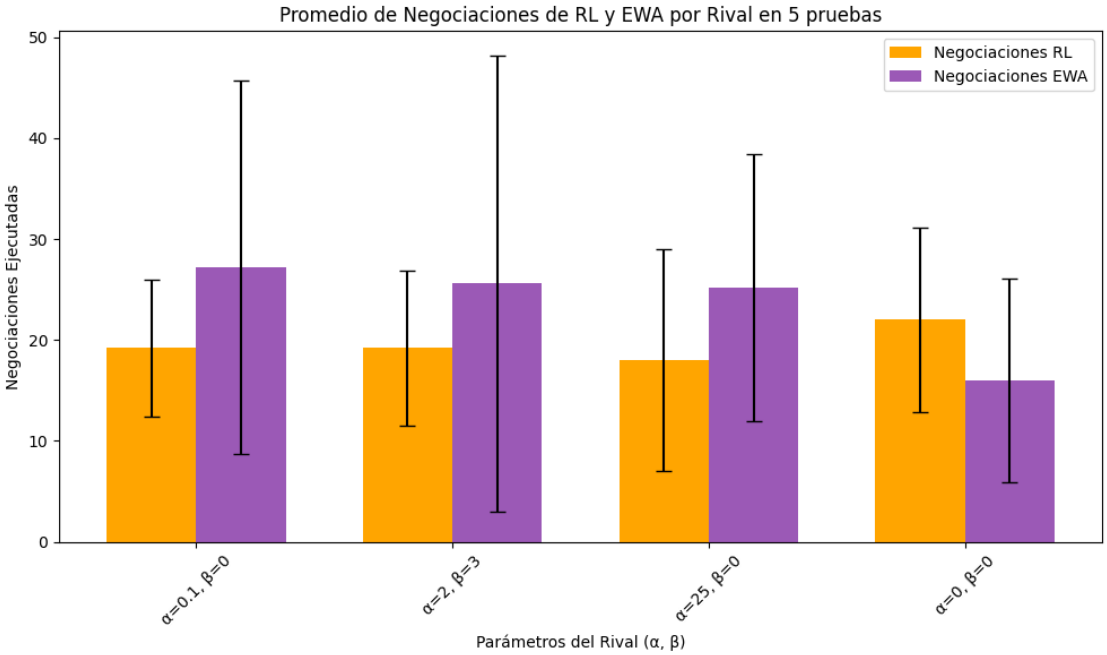

Figura 60: Negociaciones que llevaron a cabo RL y FEWA por cada grupo cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene minoría.

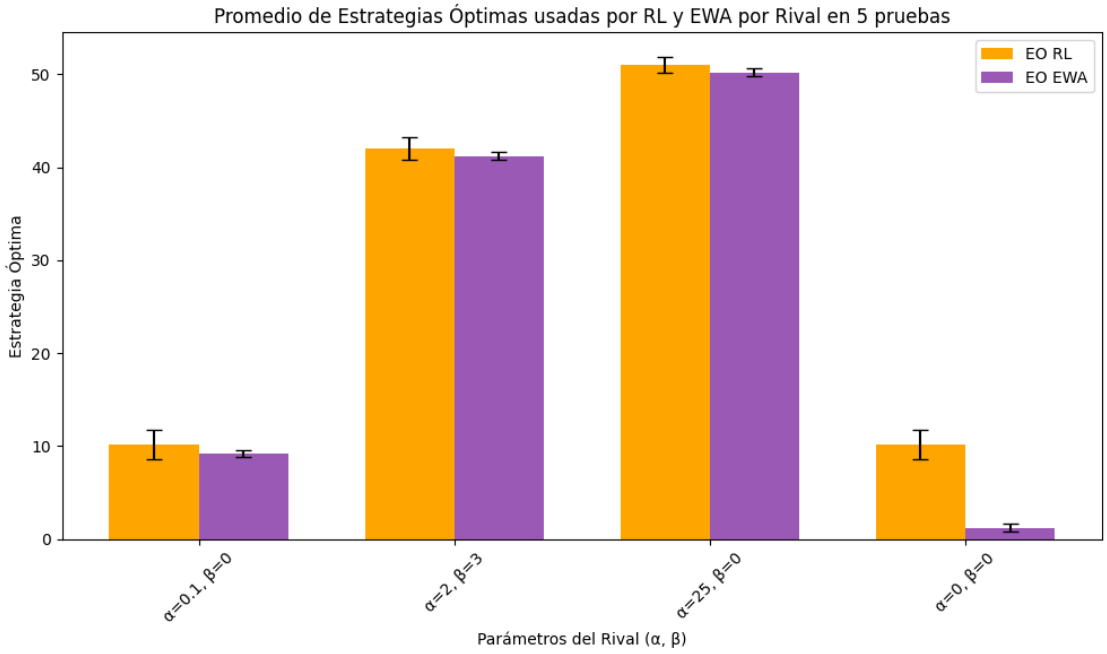

Figura 61: Estrategias usadas por los agentes RL y FEWA cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene mayoría.

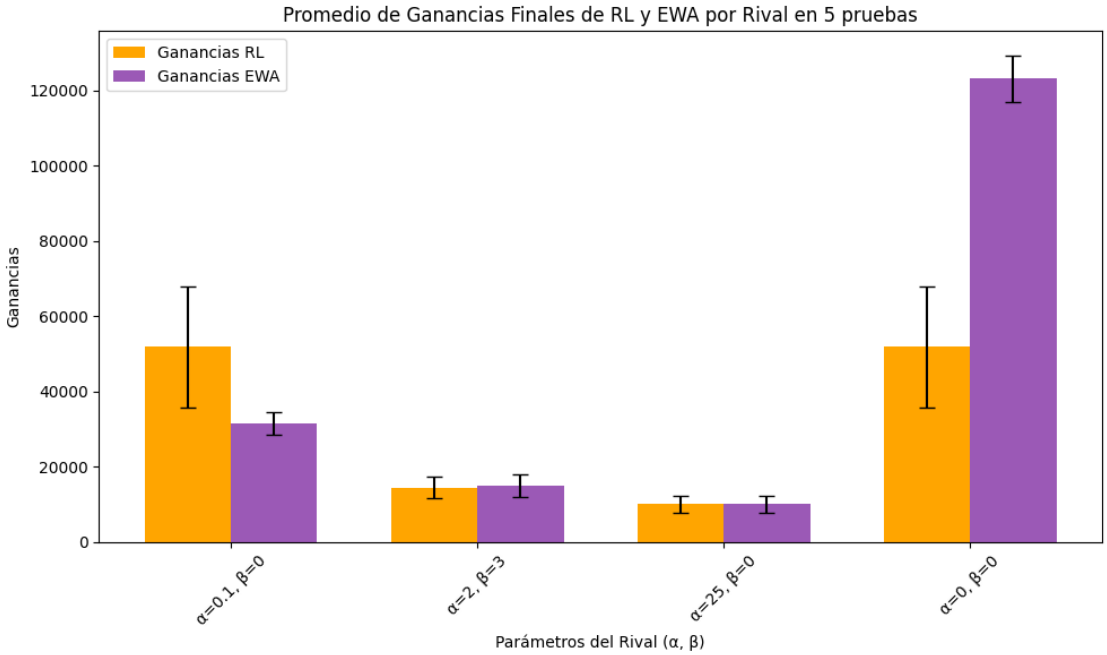

Figura 62: Ganancias que obtivieron RL y FEWA por cada grupo cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene mayoría.

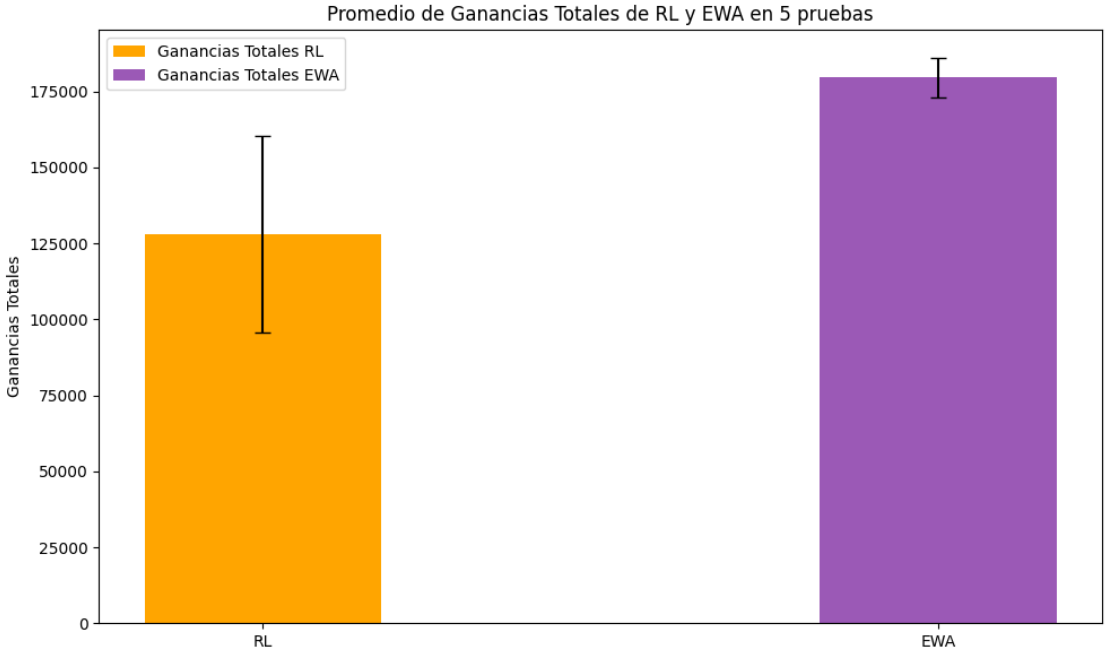

Figura 63: Ganancias totales que obtivieron RL y FEWA cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene mayoría.

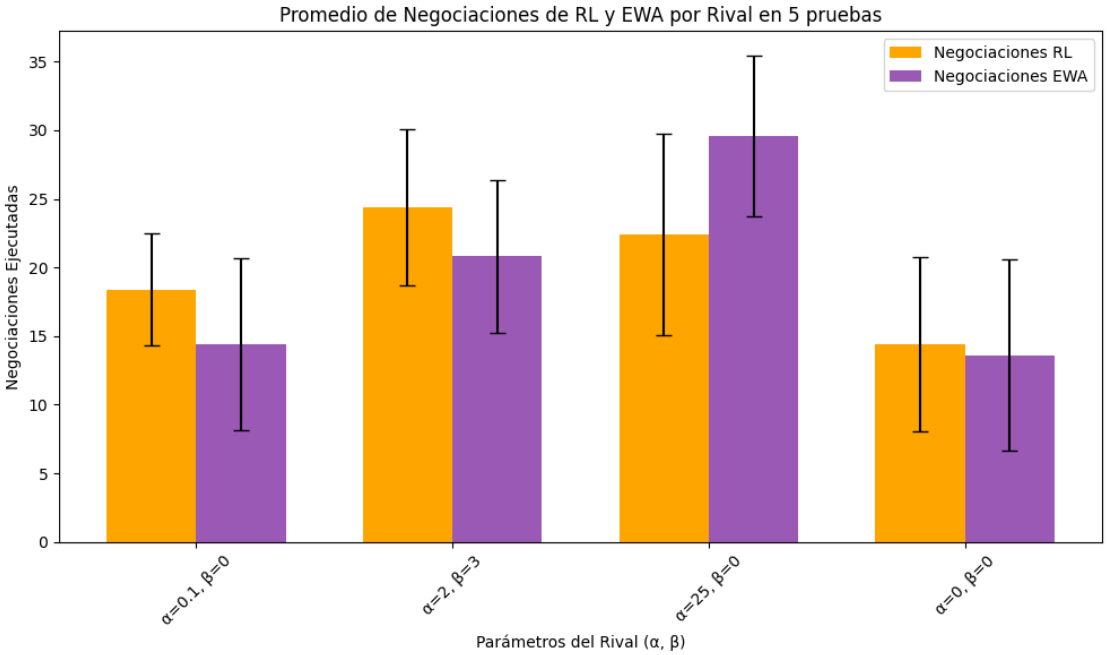

Figura 64: Negociaciones que llevaron a cabo RL y FEWA por cada grupo cuando cuando la población es de 150 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene mayoría.

Se corrieron 5 simulaciones por cada una de las 7 distribuciones, donde cada simulación constaba de 500 iteraciones del sistema AB, y en cada iteración cada agente de la población podía ejecutar uno o más juegos de ultimatum, se siguió este procedimiento para cada tamaño de población, dando un total de 6 simulaciones de aprendizaje y 105 del sistema AB.

Por otra parte, comenzando con los rivales aprendidos previamente, para la distribución de rivales conocidos para RL y FEWA donde los rivales con preferencia alta a resultados ventajosos tienen mayoría y en la distribución donde los rivales con preferencia baja a resultados ventajosos tienen mayoría, FEWA sobrepasó ligeramente a RL en ganancias, mientras que en la distribución de rivales donde los rivales con preferencia media a resultados justos tienen mayoría, RL sobrepasó, también ligeramente, a FEWA en ganancias.

En cuanto a la distribución de rivales conocidos para RL y FEWA y un rival desconocido donde todos tienen una proporción uniforme, FEWA es capaz de reconocer al nuevo rival, que es el rival con preferencia a resultados eficientes, y por ello, FEWA lo explota siendo el grupo de rivales con el que negoció más y con el que obtuvo más ganancias. Por su parte, RL clasificó al grupo de rivales con preferencia a resultados eficientes como rivales con preferencia baja a la ventaja, por ello obtuvo las mismas ganancias de ambos grupos de agentes. Por

esto, vemos que en la distribución mencionada FEWA sobrepasó a RL por haber reconocido y explotado al nuevo rival.

En el caso de la distribución de rivales conocidos para RL y FEWA donde el rival desconocido tiene minoría, FEWA se centra en explotar al grupo de rivales con preferencia a resultados eficientes, ya que le ofrece más ganancias, sin embargo, esto termina por provocar que FEWA obtenga menos ganancias totales que RL, debido a que es el grupo más pequeño y RL pudo explotar agrupamientos de rivales con preferencia baja a la ventaja. No obstante, en el caso contrario, cuando la distribución de rivales conocidos para RL y FEWA donde el rival desconocido tiene mayoría, FEWA termina por obtener más ganancias totales que RL, siendo a los rivales con preferencia a resultados eficientes el grupo que más explotó.

En los Anexos, se pueden observar prácticamente los mismos resultados para las distribuciones de la población de 50 y de 300 agentes, con la diferencia de que con la población de 150, los agentes que aprenden explotan alrededor del 0.8 de las simulaciones, en la población de 300 alrededor del 0.6 y en la población de 50 alrededor del 0.95. Asimismo, en cuanto a diferencias de desempeño por tamaño de población, para la población de 300, en comparación a la de 150, en la distribución de rivales conocidos para RL y FEWA donde los rivales con preferencia baja a resultados ventajosos tienen mayoría, RL sobrepasa a FEWA en ganancias totales, y para la misma población en la distribución de rivales conocidos para RL y FEWA donde todos tienen una proporción uniforme, en lugar de que los agentes que aprenden obtengan ganancias totales similares como en 50 y en 150, FEWA una vez más sobrepasa ligeramente a RL.

En general, los resultados de las ganancias pueden cambiar dependiendo de los agrupamientos que surjan en la población. Sin embargo, lo que es común, independientemente del tamaño de la población, es que el promedio de estrategias que usó en FEWA en comparación a RL es ligeramente más conveniente para el agente. Además, cuando hay rivales con preferencia a resultados eficientes en la población, las ganancias del agente RL tienen más variabilidad, debido a que el éxito en ganancias de RL depende de que se formen agrupamientos de rivales con preferencia baja a la ventaja, y la presencia de los rivales eficientes hace más volátil esta posibilidad. Asimismo, cuando la distribución está conformada sólo de rivales conocidos, en general, FEWA y RL suelen tener ganancias similares, usualmente FEWA sobrepasa levemente a RL en ganancias totales y, en pocos casos, uno sobrepasa de forma notoria al otro. También, en los casos de la distribución de rivales conocidos para RL y FEWA donde el rival desconocido tiene minoría, RL obtiene más ganancias totales que

FEWA y en la distribución de rivales conocidos para RL y FEWA donde el rival desconocido tiene mayoría, FEWA obtiene más ganancias totales que RL.

1. ¿En qué medida las variaciones en la distribución de rivales resultan en segregación?

    a) Entre mayor sea la probabilidad de aparición de agentes con preferencia alta a resultados ventajosos y de agentes con preferencia por resultados eficientes en la distribución de población habrá más desacuerdos y, por lo tanto, más movimineto, lo que se traduce en menos formación de grupos y menos segregación.

1. ¿En qué medida las variaciones en el tamaño de la población resultan en segregación?

    a) Sin importar el tamaño de la muestra, si la probabilidad de aparición de agentes con preferencia alta a resultados ventajosos y de agentes con preferencia por resultados eficientes es alta en la distribución de población, seguirá habiendo desacuerdos, y por lo tanto, menos segregación.

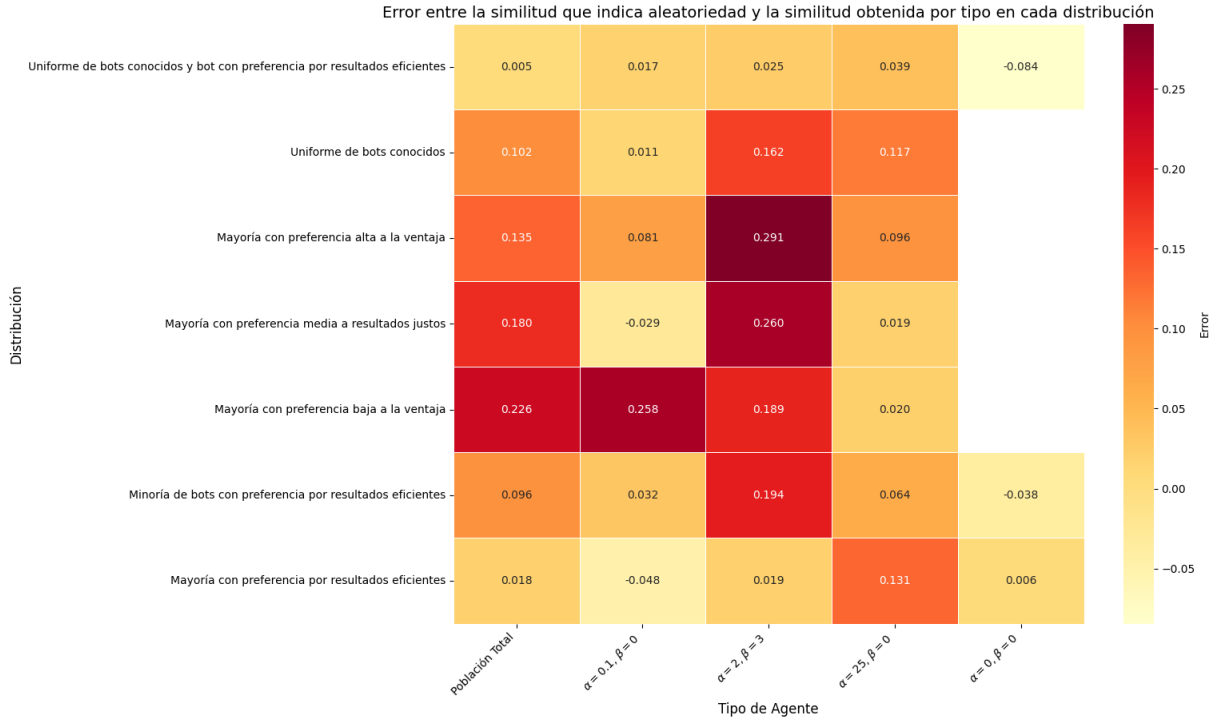

Figura 65: Error entre la similitud esperada por cada grupo (punto donde no hay segregación ya que la similitud de los vecinos de un grupo es equivalente a su proporción en la población y, por lo tanto sus posiciones son aleatorias) y la similitud entre los agentes de un grupo obtenida en las simulaciones. Esto por cada grupo, para cada población y para cada distribución cuando la muestra es de tamaño de 150 agentes.

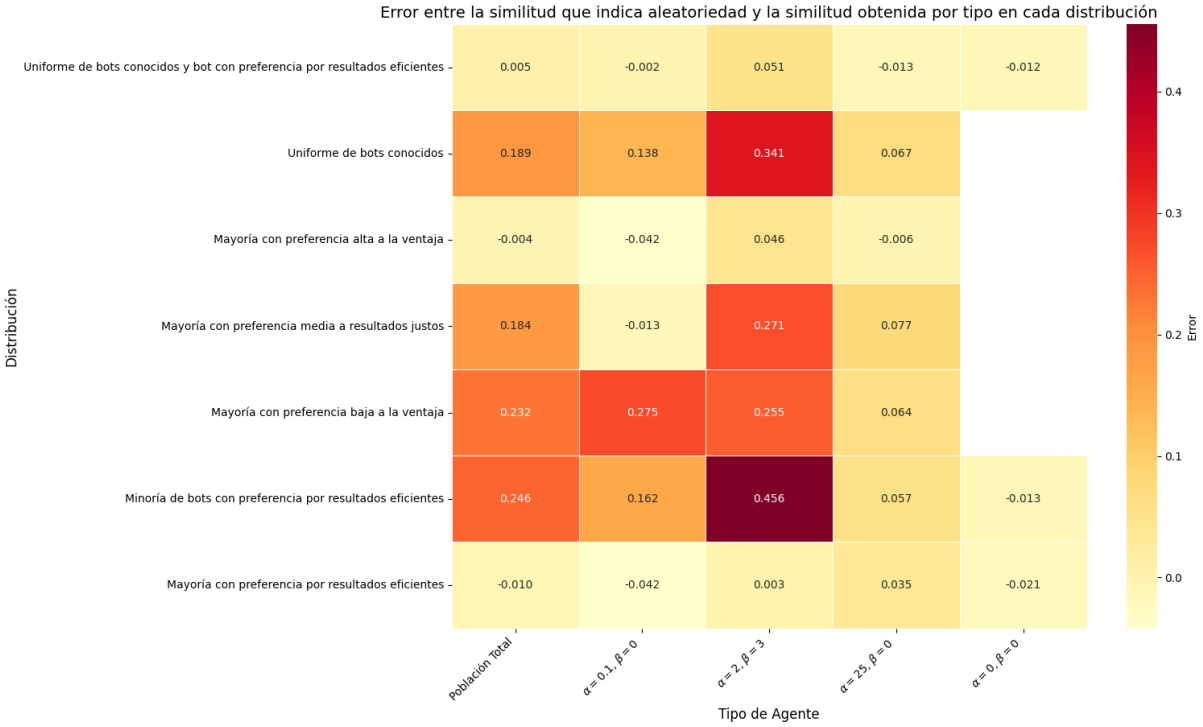

Figura 66: Error entre la similitud esperada por cada grupo (punto donde no hay segregación ya que la similitud de los vecinos de un grupo es equivalente a su proporción en la población y, por lo tanto sus posiciones son aleatorias) y la similitud entre los agentes de un grupo obtenida en las simulaciones. Esto por cada grupo, para cada población y para cada distribución cuando la muestra es de tamaño de 300 agentes.

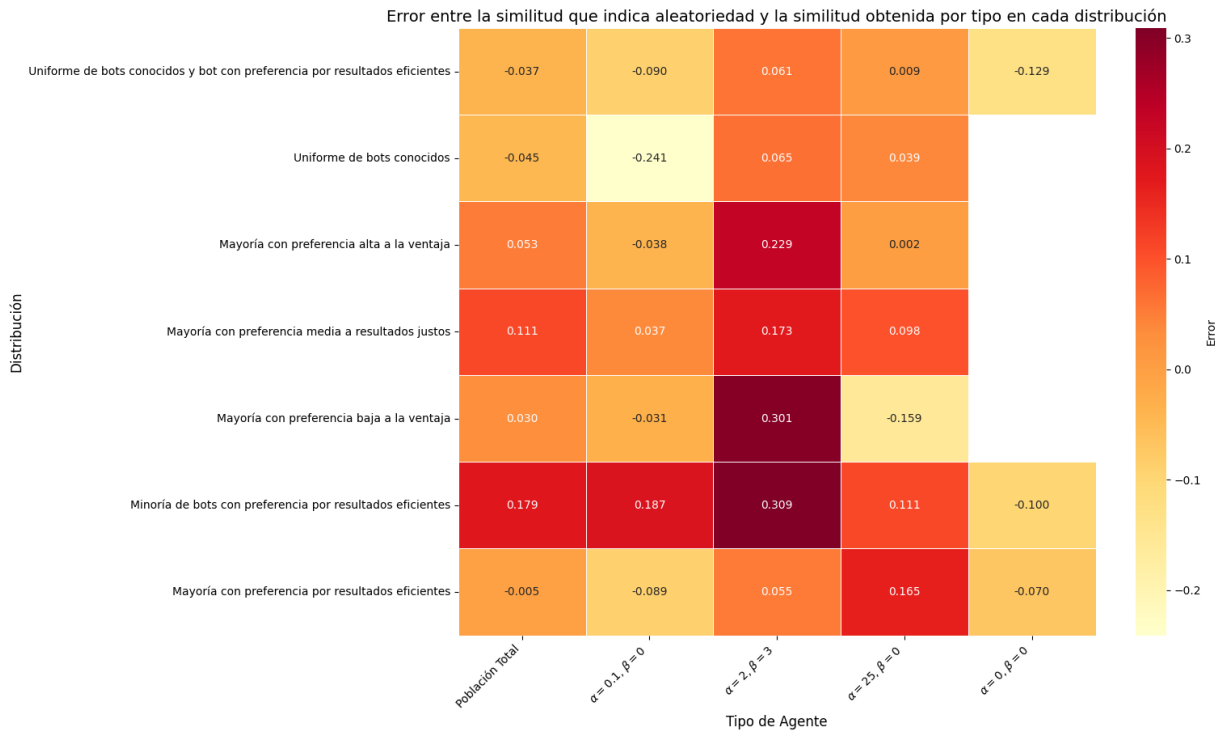

Figura 67: Error entre la similitud esperada por cada grupo (punto donde no hay segregación ya que la similitud de los vecinos de un grupo es equivalente a su proporción en la población y, por lo tanto sus posiciones son aleatorias) y la similitud entre los agentes de un grupo obtenida en las simulaciones. Esto por cada grupo, para cada población y para cada distribución cuando la muestra es de tamaño de 50 agentes.

La segregación se interpreta como el grado en que un tipo de rival está rodeado de otros rivales de su mismo tipo. De esta manera, cuando no hay segregación, la similitud esperada de los vecinos es igual a la proporción de su tipo en la población. En otras palabras; con una similitud por debajo de la esperada, los agentes se están rodeándo activamente de agentes de distinto tipo, o están evitando activamente a los de su mismo tipo; con una similitud igual a la esperada, los agentes están distribuidos aleatoriamente, por lo que no hay segregación, no hay formación de clústeres homogéneos y las interacciones son diversas; con una similitud por encima de la esperada, se sugiere que los agentes tienen cierto grado de tendencia a agruparse con otros de su mismo tipo; con una similitud alta, es decir, la proporción de vecinos del mismo tipo es cercana a 1, indica que los agentes mayoritariamente están rodeados por otros de su mismo tipo. Asimismo, la homogeneidad esperada en de toda la población es igual a $S = \sum {\eta_i}^2$, donde η es la proporción del tipo de rivales i en la población, y donde un valor igual a S indica aleatoriedad, por debajo sugiere más más heterogeneidad y por encima más homogeneidad.

Como se mencionó en los Métodos, probamos 3 tamaños de muestra, cada una constó de 7 distribuciones diferentes, se corrieron 5 simulaciones por cada distribución, donde cada simulación constaba de 500 iteraciones del sistema AB y, es importante recordar que por cada iteración cada agente de la población podía ejecutar uno o más juegos de ultimatum, lo anterior da un total de de 6 simulaciones de aprendizaje y 105 del sistema AB.

En general, como se muestra en los anexos, hay mucha variabilidad en los resultados de las diferentes simulaciones, por esto, pueden llegar a haber simulaciones con resultados ligeramente diferentes. No obstante, un caso que suele ser común es que cuando no hay mayoría de rivales con preferencia alta a la ventaja, ni de rivales con preferencia por resultados eficientes, el grupo con preferencia baja a la ventaja suele tener una tendencia leve o moderada de segregación, es decir, a agruparse con agentes del mismo tipo, y en los mismos escenarios, el grupo de rivales con preferencia media a resultados justos tiene una tendencia moderada o alta de segregación. Por lo mismo, cuando hay mayoría de rivales con preferencia alta a resultados ventajosos y con preferencia a resultados eficientes, suele haber una presencia mínima o nula de segregación. Además, parece que entre más grande es el tamaño de la población la tendencia a la segregación es más notoria. Por consiguiente los resultados también parecen indicar que entre más pequeña sea la población hay menos tendencia a segregación, de hecho, en estos casos es cuando el grupo con preferencia baja a la ventaja y el grupo con preferencia a resultados eficientes suele tener una similitud con sus vecinos por debajo de la esperada, lo que se puede deber a la sobre explotación por parte del grupo de rivales con preferencia alta a la ventaja.

1. ¿Algún estado del modelo provoca el surgimiento de un patrón reconocible o equilibrio?

    a) En la simulación, los agentes siempre deben ejecutar su regla individual, por lo que nunca se llegará a ningún punto estático. Sin embargo, se espera llegar a patrones constantes y reconocibles en los movimientos y en las ganancias, las cuales dependerán, una vez más, de la probabilidad de aparición de agentes con preferencia alta a resultados ventajosos y de agentes con preferencia por resultados eficientes, si su probabilidad es alta, habrá más constancia en las pérdidas de la población y por lo tanto en los movimientos, si su probabilidad es baja, habrá más ganancias y menos movimientos, esto sin importar el tamaño de la muestra.

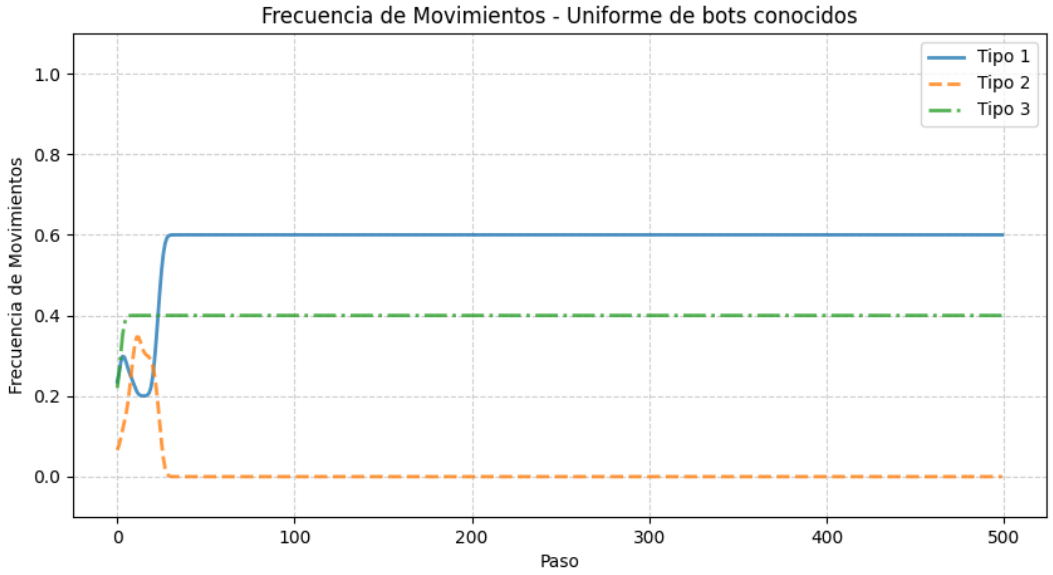

Figura 68: Movimientos de cada grupo cuando la población es de 150 agentes y la distribución es uniforme de bots conocidos.

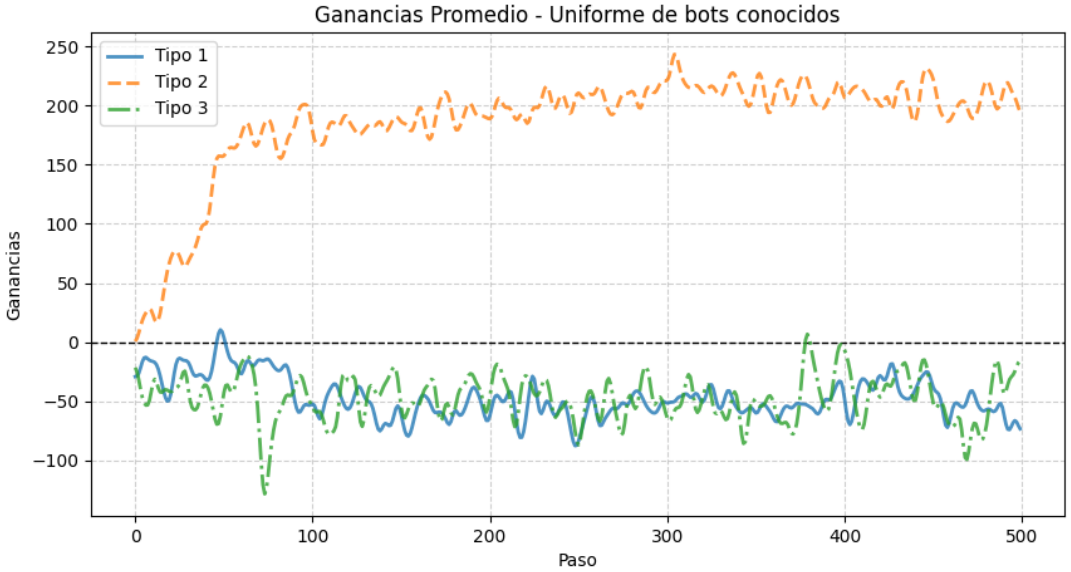

Figura 69: Ganancias de cada grupo cuando la población es de 150 agentes y la distribución es uniforme de rivales conocidos.

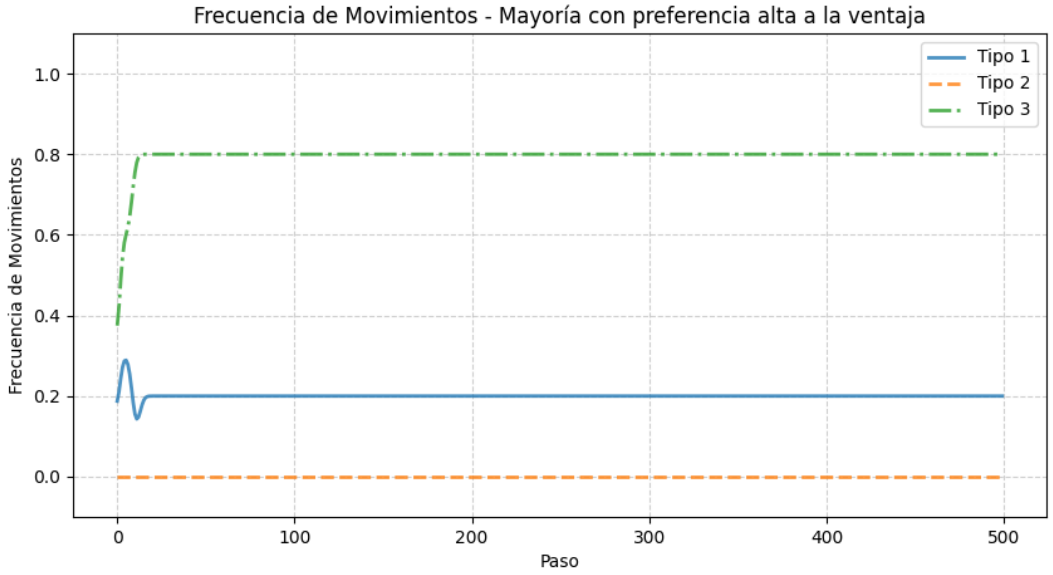

Figura 70: Movimientos de cada grupo cuando la población es de 150 agentes y la distribución es de rivales conocidos pero el rival con preferencia alta a la ventaja tiene mayoría.

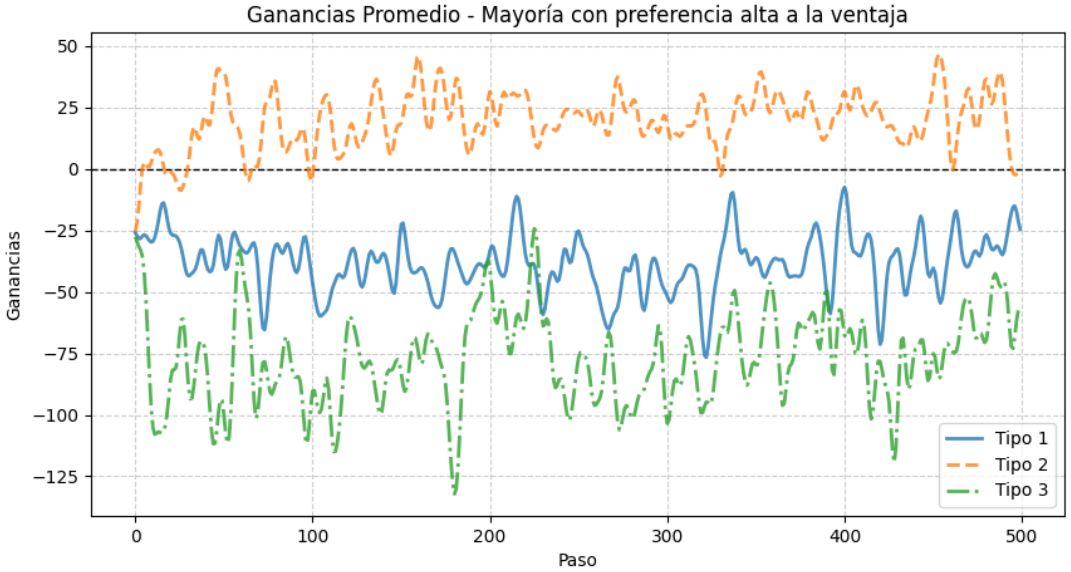

Figura 71: Ganancias de cada grupo cuando la población es de 150 agentes y la distribución es de rivales conocidos pero el rival con preferencia alta a la ventaja tiene mayoría.

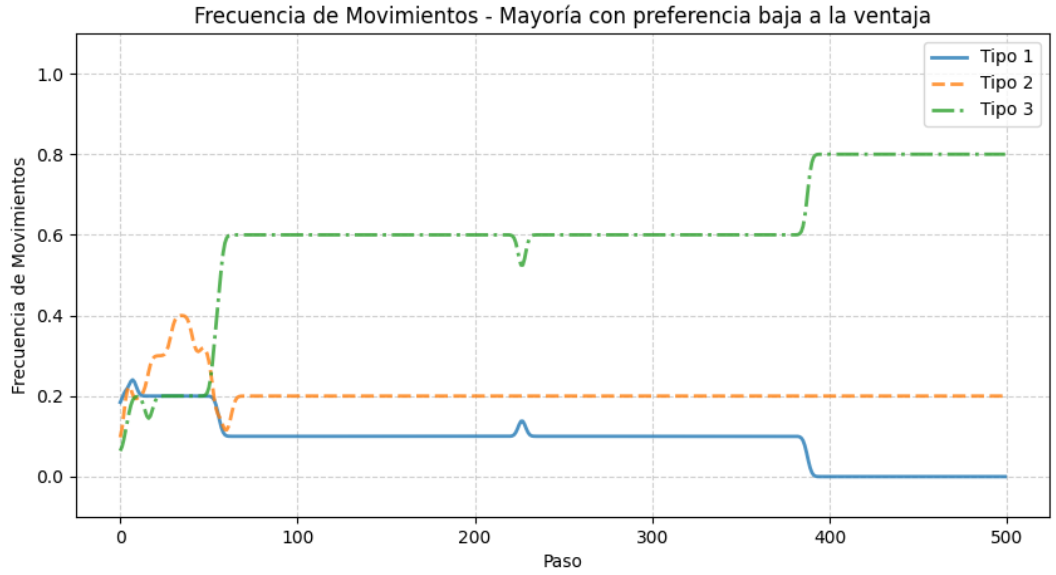

Figura 72: Movimientos de cada grupo cuando la población es de 150 agentes y la distribución es de rivales conocidos pero el rival con preferencia baja a la ventaja tiene mayoría.

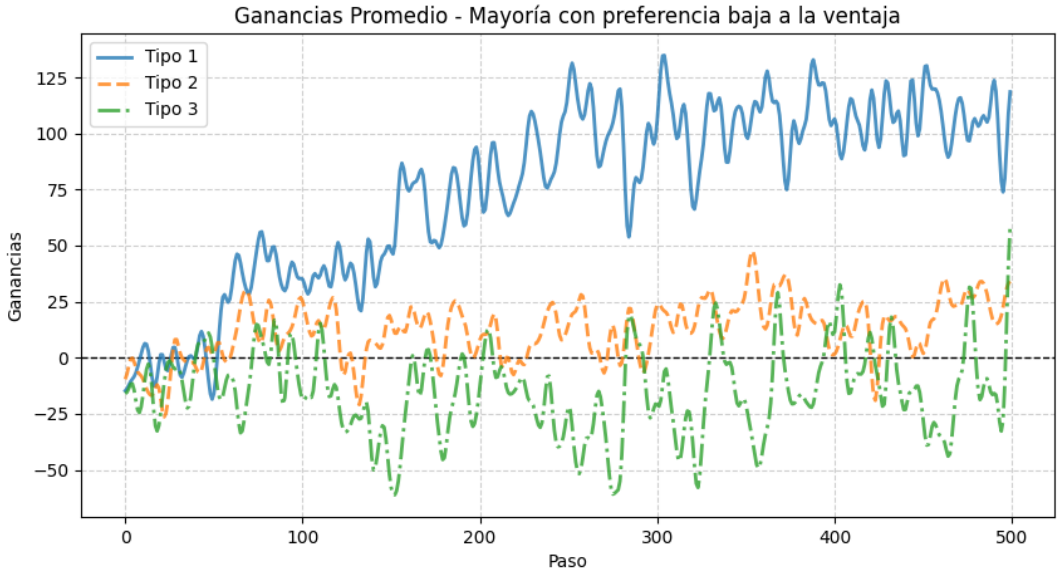

Figura 73: Ganancias de cada grupo cuando la población es de 150 agentes y la distribución es de rivales conocidos pero el rival con preferencia baja a la ventaja tiene mayoría.

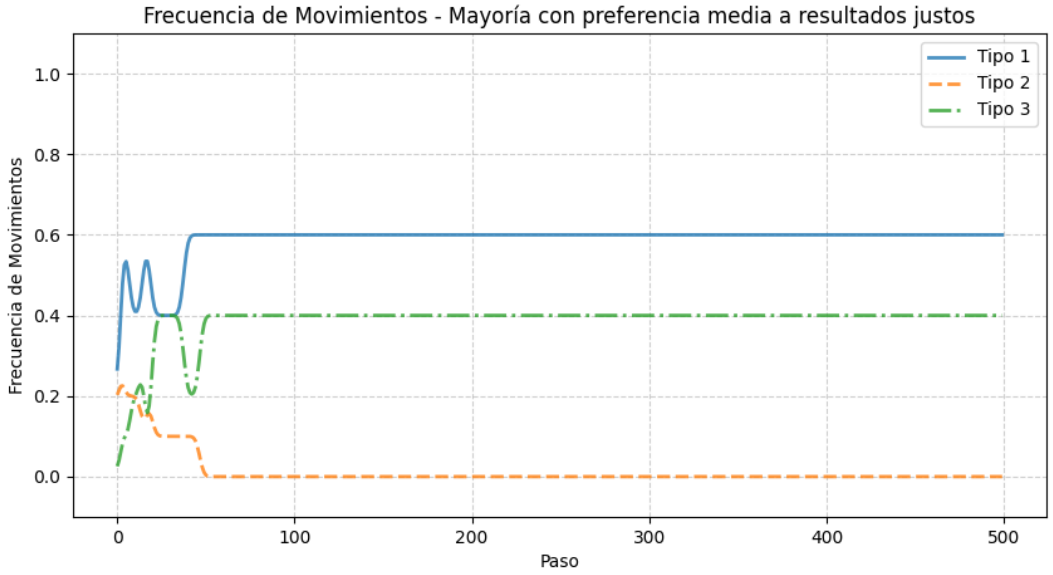

Figura 74: Movimientos de cada grupo cuando la población es de 150 agentes y la distribución es de rivales conocidos pero el rival con preferencia media a resultados justos tiene mayoría.

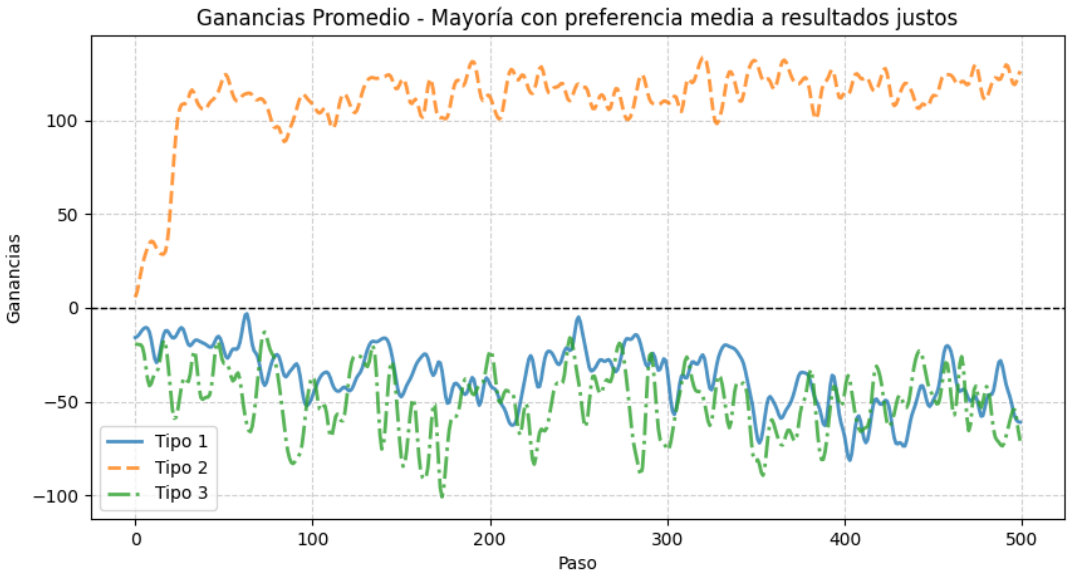

Figura 75: Ganancias de cada grupo cuando la población es de 150 agentes y la distribución es de rivales conocidos pero el rival con preferencia media a resultados justos tiene mayoría.

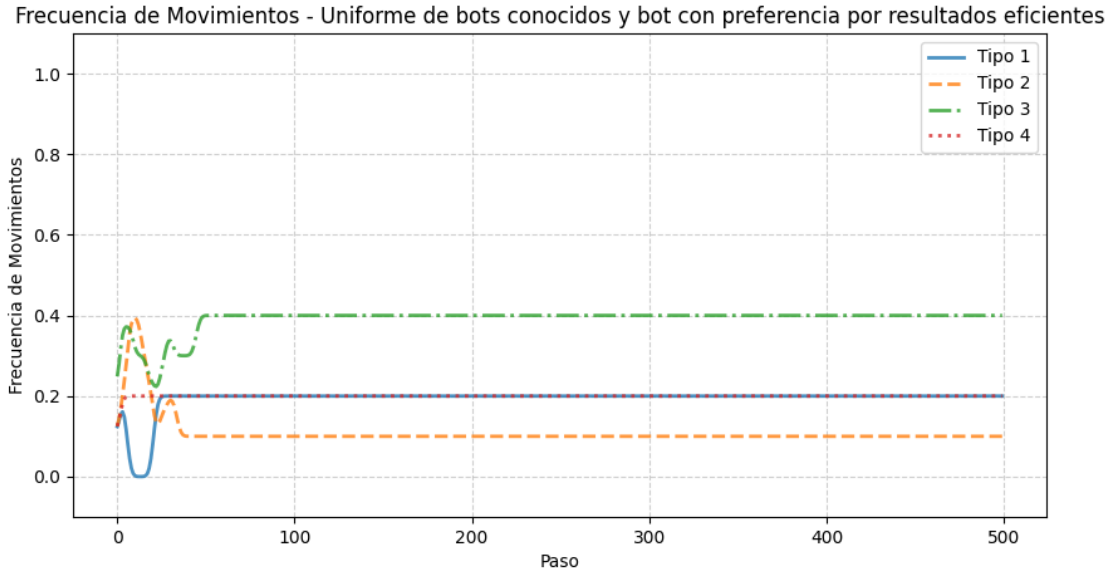

Figura 76: Movimientos de cada grupo cuando la población es de 150 agentes y la distribución es uniforme de rivales conocidos y un rival desconocido con preferencia por resultados eficientes.

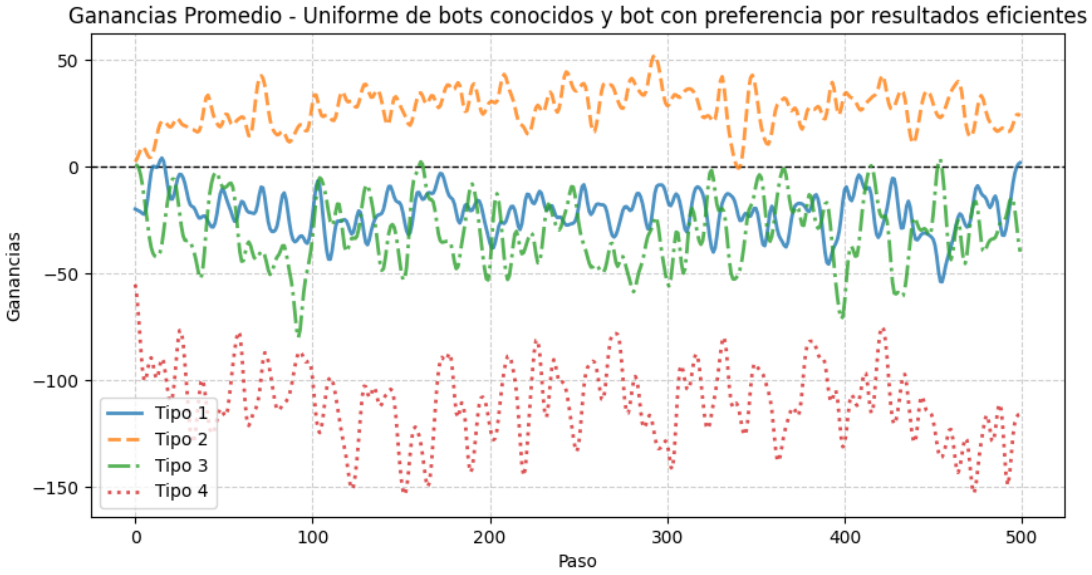

Figura 77: Ganancias de cada grupo cuando la población es de 150 agentes y la distribución es uniforme de rivales conocidos y un rival desconocido con preferencia por resultados eficientes.

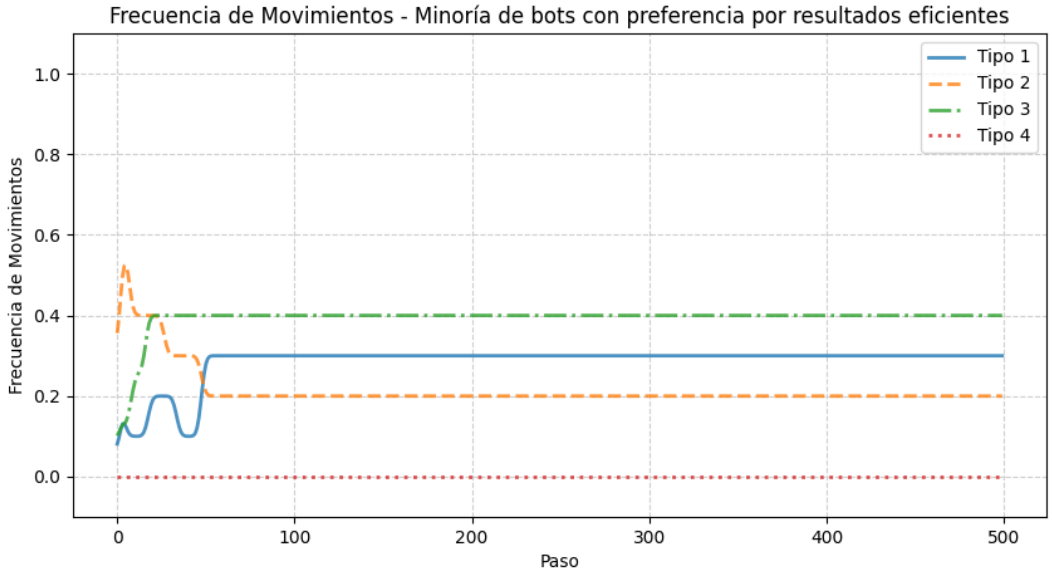

Figura 78: Movimientos de cada grupo cuando la población es de 150 agentes y la distribución de rivales conocidos y una minoría de rivales desconocido con preferencia por resultados eficientes.

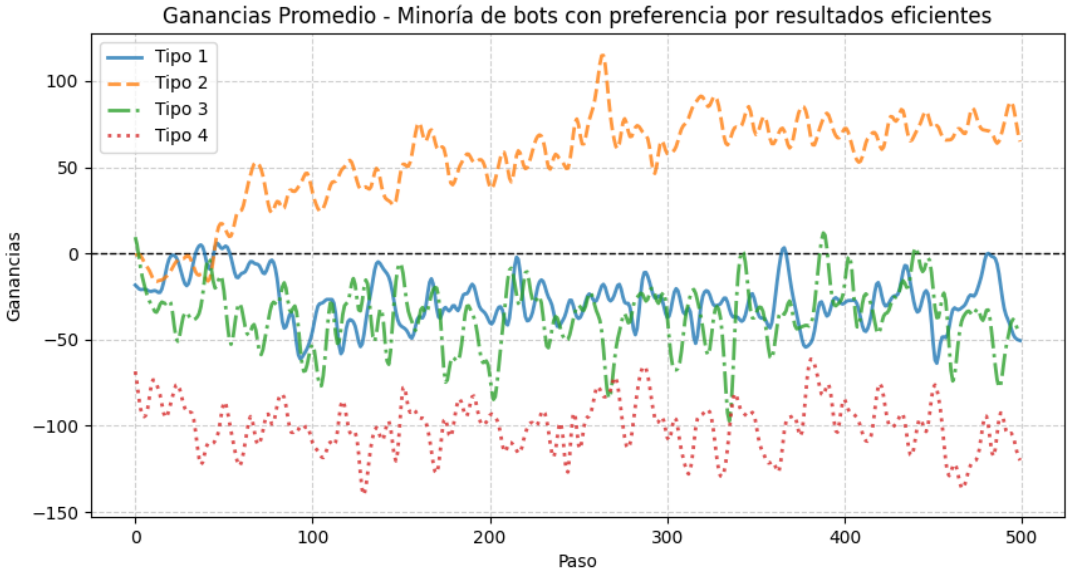

Figura 79: Ganancias de cada grupo cuando la población es de 150 agentes y la distribución de rivales conocidos y una minoría de rivales desconocido con preferencia por resultados eficientes.

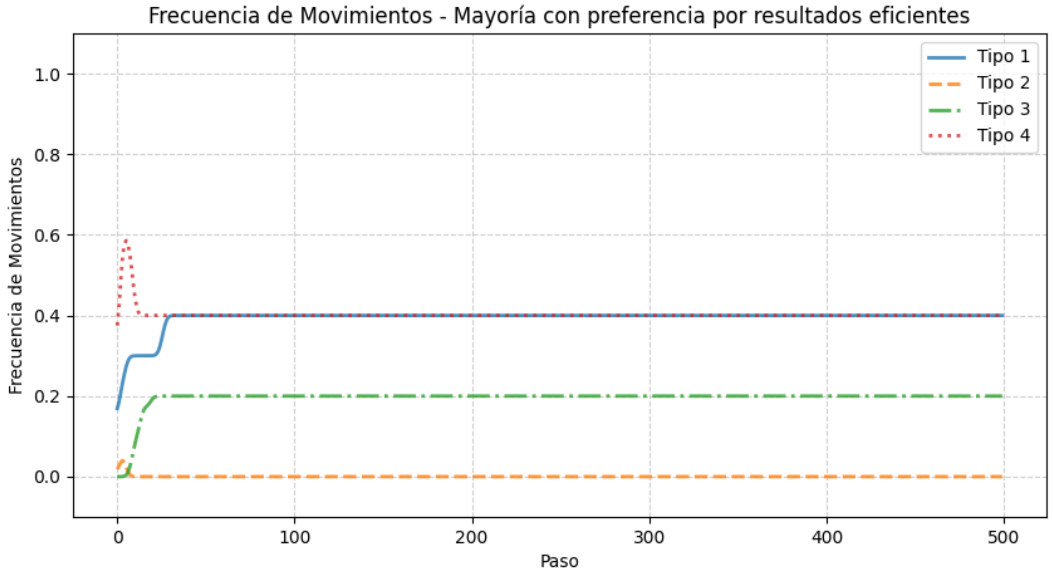

Figura 80: Movimientos de cada grupo cuando la población es de 150 agentes y la distribución de rivales conocidos y una mayoría de rivales desconocido con preferencia por resultados eficientes.

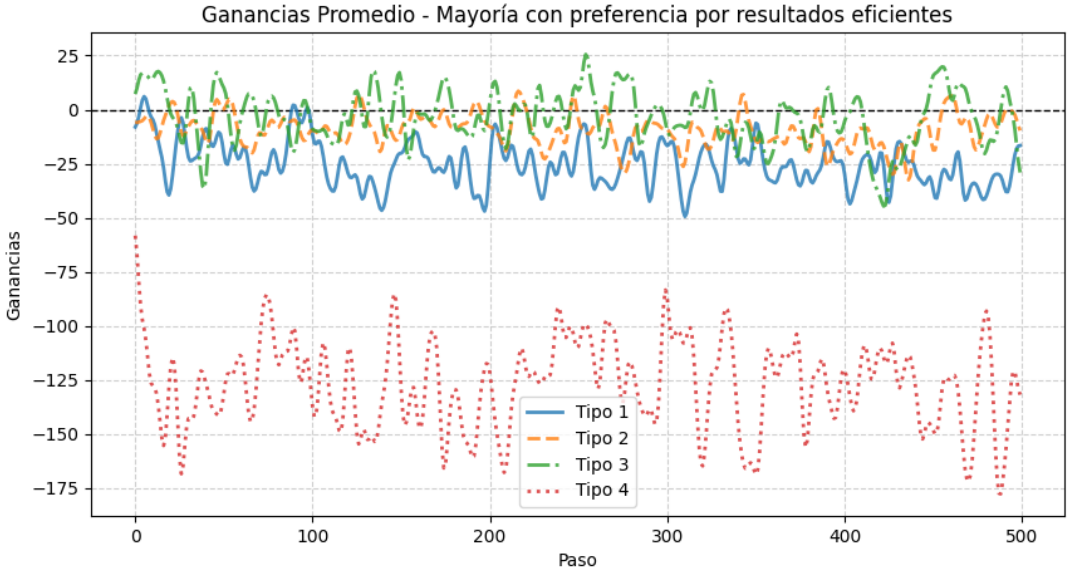

Figura 81: Ganancias de cada grupo cuando la población es de 150 agentes y la distribución de rivales conocidos y una mayoría de rivales desconocido con preferencia por resultados eficientes.

Una vez más, cada uno de los 3 tamaños de muestra constó de 7 distribuciones diferentes, se corrieron 5 simulaciones por cada distribución, donde cada simulación constaba de 500

iteraciones del sistema AB, y por cada iteración cada agente de la población podía ejecutar uno o más juegos de ultimatum, lo anterior da un total de de 6 simulaciones de aprendizaje y 105 del sistema AB.

Como se observa en los gráficos, cuando el tamaño de la población es de 150 agentes, a grandes rasgos, en todas las distribuciones parece que pronto se llega a un punto en donde la frecuencia de movimientos de un grupo se mantiene constante. Sin embargo, la excepción a este resultado es la distribución de rivales conocidos para RL y FEWA donde los rivales con preferencia baja a resultados ventajosos tienen mayoría, la disminución de movimientos de rivales con preferencia baja a la ventaja, su aumento de ganancias, y el aumento de movimientos de rivales con preferencia alta a la ventaja parece indicar que pasada la mitad de la simulación los rivales con preferencia baja a la ventaja lograron formar uno o más agrupamientos.

Además, en cuanto a las ganancias, incluso cuando tienen más variabilidad que los movimientos, parece que en casi todos los casos las ganancias de un grupo se mantienen constantemente por encima de la línea de pérdidas o por debajo de la línea de ganancias. Una vez más, la excepción a estos resultados es la distribución de rivales conocidos para RL y FEWA donde el rival desconocido tiene mayoría. Esto puede indicar que cuando hay mayoría de agentes que suelen hacer ofertas injustas, como los agentes con preferencia alta a la ventaja o agentes con preferencia por resultados eficientes, se provocan más desbalances para llegar a un equilibrio.

Como se puede observar en los Anexos, el mismo patrón de la población de 150 se repite para una población de 300 agentes, con excepción de la distribución de rivales conocidos para RL y FEWA donde todos tienen una proporción uniforme, ya que se demora un poco más en llegar a un equilibrio en la frecuencia de movimientos comparado a la población de 150 agentes. Asimismo, se observan resultados similares a los de la población de 150 en una población de 50 agentes, con la diferencia de que en la distribución de rivales conocidos para RL y FEWA donde los rivales con preferencia alta a resultados ventajosos tienen mayoría, en general, suele haber más pérdidas para el grupo con preferencia media a resultados justos.

7. Conclusiones

La investigación realizada sitúa a los agentes de aprendizaje Reinforcement Learning Q-Learning (RL) y Functional Experience Weighted Attraction (FEWA) en contextos de negociación, en interacciones con rivales con diferentes grados de justicia, así como en un sistema Basado en Agentes (AB), y ofrece una serie de conclusiones clave sobre el desempeño de los agentes que aprenden y dinámicas de los grupos de rivales.

Para comenzar, tanto el agente RL como el agente FEWA aprenden apropiadamente las preferencias del rival contra el que deben negociar, siendo FEWA el agente con las estrategias óptimas más cercanas al SPNE de cada rival. El aprendizaje de ambos agentes es diferenciable, ya que FEWA le asigna valor hipotético a estrategias no seleccionadas y, por esto, RL muestra una mayor diferenciación entre estrategias aceptadas y no aceptadas. Además, cuando se probaron diferentes epsilon, que implican períodos más cortos o más largos de exploración, FEWA demostró un aprendizaje más rápido, igualmente por su elemento de aprendizaje por creencias.

Por otra parte, ambos agentes son capaces de clasificar a rivales previamente aprendidos con alta precisión, sin mencionar que FEWA es capaz de identificar a un rival que no fue aprendido previamente. Sin embargo, el método de clasificación usado en este estudio solo es eficiente cuando se trata de rivales distinguibles entre ellos, ya que, en el caso contrario, los rivales cercanos en valores de parámetros serán confundidos.

El desempeño de RL y FEWA en el sistema AB depende de la distribución de rivales en la población, sin embargo, cuando solo hay rivales conocidos, los resultados en ganancia son muy parecidos entre los agentes que aprenden. No obstante, FEWA tiende a obtener mayores ganancias cuando puede identificar a rivales con preferencia por resultados eficientes, excepto cuando estos son minoría.

En cuanto a la formación de agrupamientos del sistema AB, los rivales con preferencia alta a resultados ventajosos y los rivales con preferencia por resultados eficientes interfieren con la segregación, por el lado contrario, los rivales con preferencia baja a la ventaja y los rivales con preferencia media a resultados justos tienden más a segregarse. El tamaño de la población también influye en la segregación, con poblaciones más grandes mostrando una mayor tendencia a la segregación.

Respecto al equilibrio del sistema AB, usualmente se llega a puntos constantes en movimientos y ganancias por grupos, siendo los rivales con preferencia media a resultados justos los que usualmente están obteniendo ganancias en lugar de pérdidas, y siendo los rivales con preferencia alta a resultados ventajosos y los rivales con preferencia por resultados eficientes los que promueven pérdidas en la población.

Finalmente, FEWA demuestra ser superior en términos de precisión en el aprendizaje de estrategias óptimas, en velocidad de aprendizaje, en capacidad para identificar rivales desconocidos y, por esto, también en desempeño, ya que su capacidad para asignar valores hipotéticos a estrategias no exploradas y su parámetro de detección de cambio le permiten adaptarse más rápidamente a nuevas situaciones. FEWA demuestra ser un modelo prometedor para la creación de agentes de inteligencia artificial que se desenvuelven en interacciones sociales estratégicas, no solo por sus resultados, sino también por su valor teórico siendo un híbrido entre aprendizaje por refuerzo y aprendizaje por creencias que implementa sustento cognitivo en sus parámetros.

8. Referencias

Referencias

Ahmed, A., & Karlapalem, K. (2014). Inequity aversion and the evolution of cooperation. *Physical Review E*, *89*(2), 022802. https://doi.org/10.1103/PhysRevE.89.022802

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390-1396. https://doi.org/10.1126/science.7466396

Baarslag, T., Hindriks, K., & Jonker, C. (2014). Effective acceptance conditions in real-time automated negotiation. *Decision Support Systems*, *60*, 68-77.

Brooks, R., Hassabis, D., Bray, D., & Shashua, A. (2012). Turing centenary: Is the brain a good model for machine intelligence? *Nature*, *482*(7386), 462-463. https://doi.org/10.1038/482462a

Brown, G. (1951). Iterative solutions of games by fictitious play. En T. Koopmans (Ed.), *Activity Analysis of Production and Allocation* (pp. [insert page range if available]). Wiley.

Brzostowski, J., & Kowalczyk, R. (2006). Adaptive negotiation with on-line prediction of opponent behaviour in agent-based negotiations. *Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'06)*.

Camerer, C. F. (2011). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2002). Sophisticated experience-weighted attraction learning and strategic teaching in repeated games. *Journal of Economic Theory*, *104*(1), 137-188.

Camerer, C., & Ho, T.-H. (1999). Experienced-Weighted Attraction Learning in Normal Form Games. *Econometrica*, *67*(4), 827-874. http://www.jstor.org/stable/2999459

Cao, M., & Kiang, M. Y. (2012). BDI agent architecture for multi-strategy selection in automated negotiation. *Journal of Universal Computer Science*, *18*(10), 1379-1404.

Carpenter, J., & Robbett, A. (2022). *Game Theory and Behavior*. MIT Press.

Charness, G., & Rabin, M. (2002). Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, *117*(3), 817-869. http://www.jstor.org/stable/4132490

Chen, S., & Weiss, G. (2015). An approach to complex agent-based negotiations via effectively modeling unknown opponents. *Expert Systems with Applications*, *42*(5), 2287-2304. https://doi.org/10.1016/j.eswa.2014.10.048

Cournot, A. (1960). *Researches into the Mathematical Principles of the Theory of Wealth* (N. Bacon, Trad.). Kelley.

Durlauf, S. N., & Blume, L. E. (2009). *Game Theory*. Palgrave Macmillan. https://doi.org/10.1057/9780230280847

Fehr, E., & Schmidt, K. (1999). A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics*, *114*, 817-851. https://doi.org/10.1162/003355399556151

Gasparrini, M., & Sánchez-Fibla, M. (2019). Loss aversion fosters coordination among independent reinforcement learners. https://doi.org/10.48550/arXiv.1912.12633

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*(4), 367-388. https://doi.org/10.1016/0167-2681(82)90011-7

Herrnstein, R. J. (1970). On the law of effect. *Journal of the Experimental Analysis of Behavior*, *13*(2), 243-266. https://doi.org/10.1901/jeab.1970.13-243

Ho, T. H., Camerer, C. F., & Chong, J. K. (2007). Self-tuning experience weighted attraction learning in games. *Journal of Economic Theory*, *133*(1), 177-198. https://doi.org/10.1016/j.jet.2005.12.008

Huang, C.-C., Liang, W.-Y., Lai, Y.-H., & Lin, Y.-C. (2010). The agent-based negotiation process for B2C e-commerce. *Expert Systems with Applications*, *37*(1), 348-359. https://doi.org/10.1016/j.eswa.2009.05.065

Hughes, E., Leibo, J., Phillips, M., Tuyls, K., Duéñez-Guzmán, E., Castañeda, A., Dunning, I., Zhu, T., McKee, K., Koster, R., Roff, H., & Graepel, T. (2018). Inequity aversion improves cooperation in intertemporal social dilemmas. *Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1803.08884

Janssen, M. (2014). An Agent-Based Model Based on Field Experiments. En A. Smajgl & O. Barreteau (Eds.), *Empirical Agent-Based Modelling - Challenges and Solutions* (pp. 175-190). Springer. https://doi.org/10.1007/978-1-4614-6134-0_10

Janssen, M. A., DeCaro, D., & Lee, A. (2022). An Agent-Based Model of the Interaction Between Inequality, Trust, and Communication in Common Pool Experiments. *Journal of Artificial Societies and Social Simulation*, *25*(4), 3. https://doi.org/10.18564/jasss.4922

Jong, S., Tuyls, K., & Verbeeck, K. (2008). Artificial agents learning human fairness. *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems*, *2*, 863-870.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263-291. https://doi.org/10.2307/1914185

Kölle, F., Sliwka, D., & Zhou, N. (2016). Heterogeneity, inequity aversion, and group performance. *Social Choice and Welfare*, *46*, 263-286. https://doi.org/10.1007/s00355-015-0912-5

Kuperman, M., & Risau-Gusman, S. (2008). The effect of the topology on the spatial ultimatum game. *European Physical Journal B*, *62*, 233-238. https://doi.org/10.1140/epjb/e2008-00133-x

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

Luce, R. D. (1959). Individual Choice Behavior. *Wiley*.

McKee, K., Gemp, I., McWilliams, B., Duéñez-Guzmán, E., Hughes, E., & Leibo, J. (2020). Social Diversity and Social Preferences in Mixed-Motive Reinforcement Learning. *Adaptive Agents and Multi-Agent Systems*. https://doi.org/10.48550/arXiv.2002.02325

Mnih, V. (2013). Playing Atari with Deep Reinforcement Learning. *arXiv preprint arXiv:1312.5602*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529-533. https://doi.org/10.1038/nature14236

Nash, J. (1953). Two person cooperative games. *Econometrica*, *21*, 128-140.

Nash, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the United States of America*, *36*(1), 48-49. https://doi.org/10.1073/pnas.36.1.48

Nash, J. F. J. (1950). The Bargaining Problem. *Econometrica*, *18*(2), 155-162. https://doi.org/10.2307/1907266

Nguyen, D., Venkatesh, S., Nguyen, P., & Tran, T. (2020). Theory of Mind with Guilt Aversion Facilitates Cooperative Reinforcement Learning. *Asian Conference on Machine Learning*. https://doi.org/10.48550/arXiv.2009.07445

Rabin, M. (1993). Incorporating Fairness into Game Theory and Economics. *American Economic Review*, *83*(5), 1281-1302.

Realpe-Gómez, J., Andrighetto, G., Nardin, L. G., & Montoya, J. A. (2018). Balancing selfishness and norm conformity can explain human behavior in large-scale prisoner's dilemma games and can poise human groups near criticality. *Physical Review E*, *97*(4), 042321.

Rubinstein, A. (1982). Perfect Equilibrium in a Bargaining Model. *Econometrica*, *50*(1), 97-109. https://doi.org/10.2307/1912531

Rummery, G., & Niranjan, M. (1994). *On-line Q-learning using connectionist systems* (inf. téc. N.º CUED/F-INFENG/TR 166). Cambridge University Engineering Department.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210-229.

Saxe, A., Nelli, S., & Summerfield, C. (2021). If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, *22*(1), 55-67. https://doi.org/10.1038/s41583-020-00310-3

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, *1*(2), 143-186.

Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence*, *299*, 103535. https://doi.org/10.1016/j.artint.2021.103535

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99-118. https://doi.org/10.2307/1884852

Smaldino, P. E. (2023). *Modeling Social Behavior: Mathematical and Agent-Based Models of Social Dynamics and Cultural Evolution*. Princeton University Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd). MIT Press.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, *2*(4).

Thorndike, E. L. (1931). *Human learning*. Appleton-Century-Crofts.

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *2*(1), 230-265. https://doi.org/10.1112/plms/s2-42.1.230

von Neumann, J. (1966). *Theory of Self-Reproducing Automata*. University of Illinois Press.

von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3), 279-292. https://doi.org/10.1007/BF00992698

Xianyu, B. (2010). Social Preference, Incomplete Information, and the Evolution of Ultimatum Game in the Small World Networks: An Agent-Based Approach. *Journal of Artificial Societies and Social Simulation*, *13*(2), 7. https://doi.org/10.18564/jasss.1534

Zeng, D., & Sycara, K. (1998). Bayesian learning in negotiation. *International Journal of Human-Computer Studies*, *48*(2), 125-141.

Zhang, J., Ren, F., & Zhang, M. (2015). Bayesian-based preference prediction in bilateral multi-issue negotiation between intelligent agents. *Knowledge-Based Systems*, *84*, 108-120.

Zhang, S., FeldmanHall, O., H'etu, S., & Otto, A. (2024). Advantageous and disadvantageous inequality aversion can be taught through vicarious learning of others' preferences. https://doi.org/10.48550/arXiv.2405.06500

9. Anexos

9.1. Gráficos de frecuencias de selección por rival

9.1.1. Valores de epsilon para rival con preferencia baja por resultados ventajosos

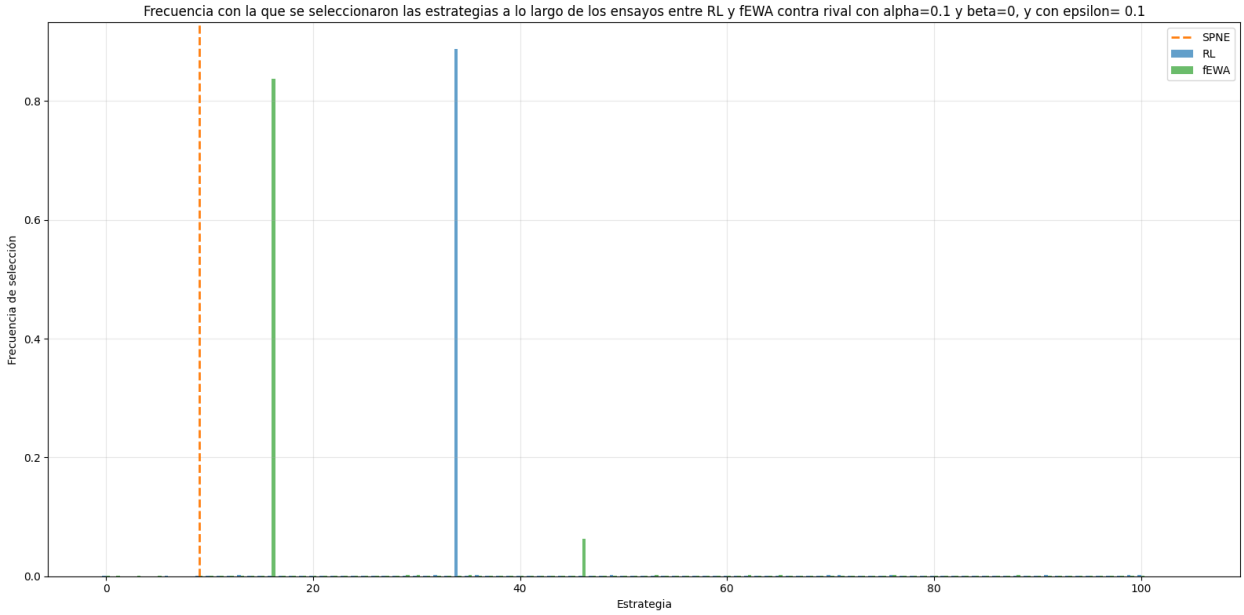

Figura 82: Frecuencia de selección de estrategias de RL y FEWA contra el rival con preferencia baja por resultados ventajosos cuando $\epsilon = 0.1$.

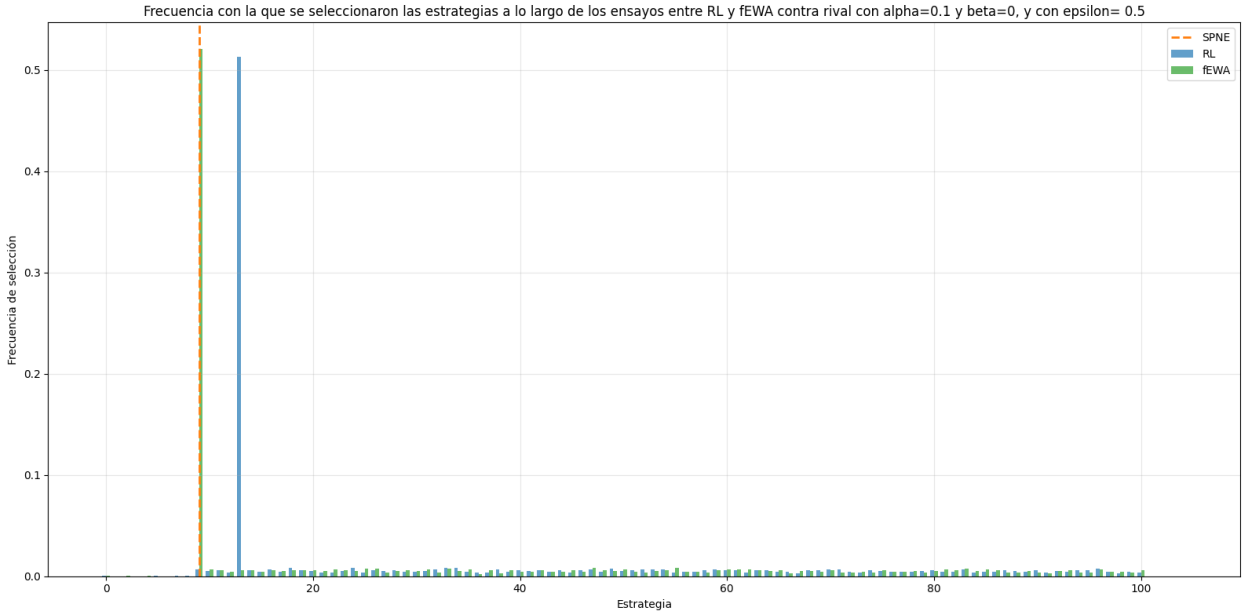

Figura 83: Frecuencia de selección de estrategias de RL y FEWA contra el rival con preferencia baja por resultados ventajosos cuando $\epsilon = 0{,}5$.

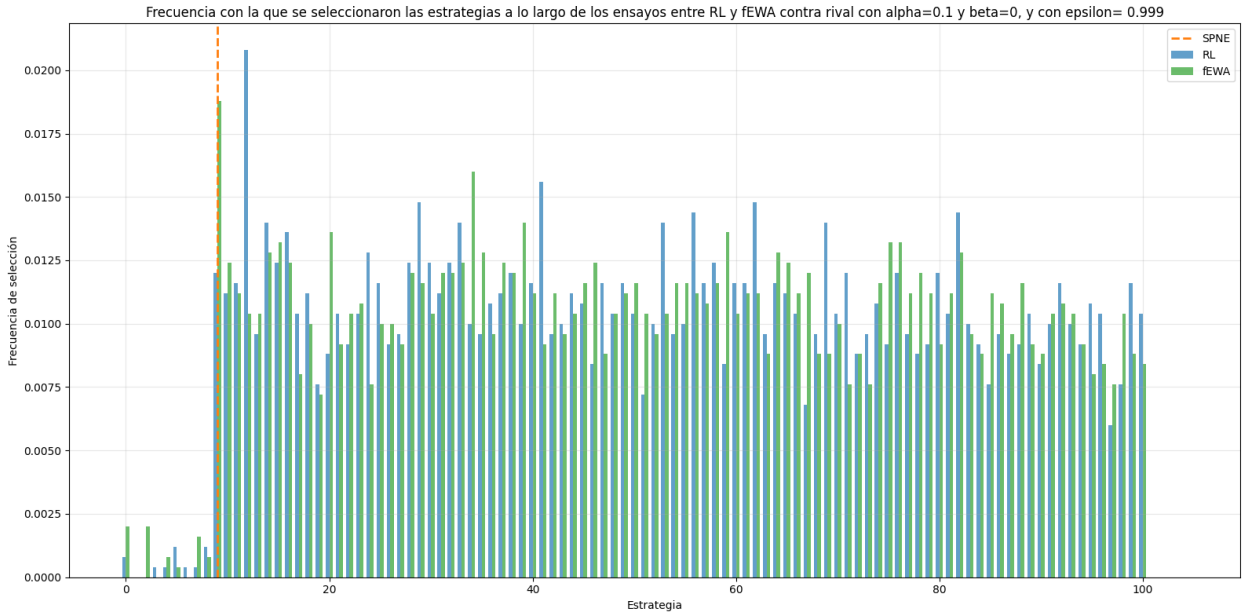

Figura 84: Frecuencia de selección de estrategias de RL y FEWA contra el rival con preferencia baja por resultados ventajosos cuando $\epsilon = 0{,}9$.

9.1.2. Valores de epsilon para rival con preferencia media por resultados justos

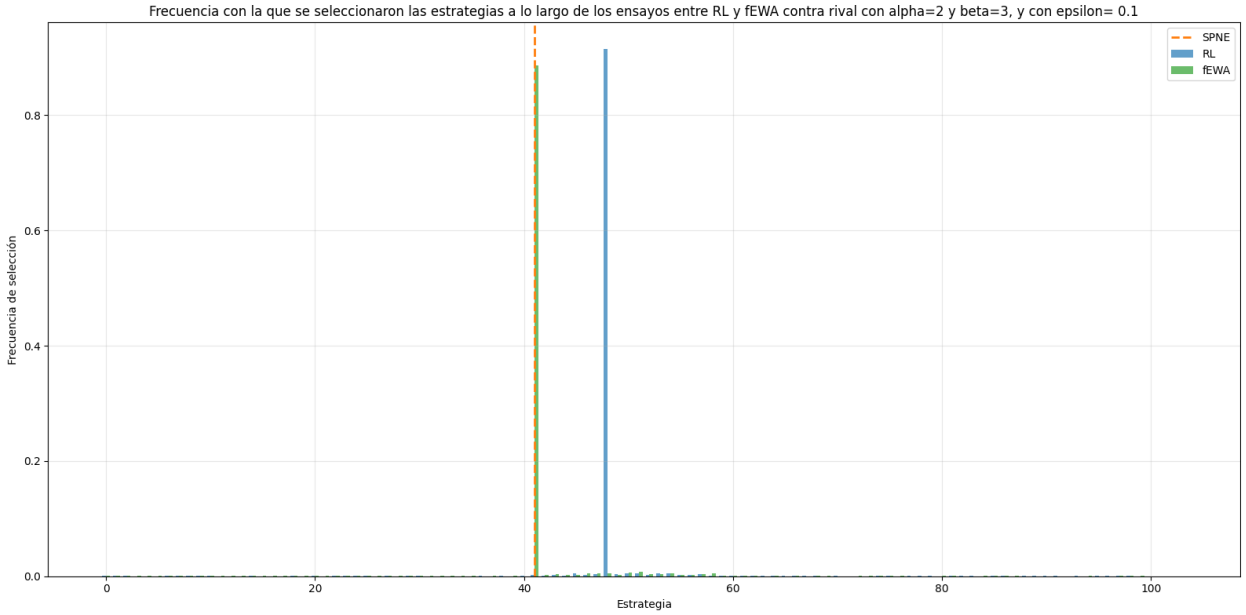

Figura 85: Frecuencia de selección de estrategias de RL y FEWA contra el rival con preferencia media por resultados justos cuando $\epsilon = 0{,}1$.

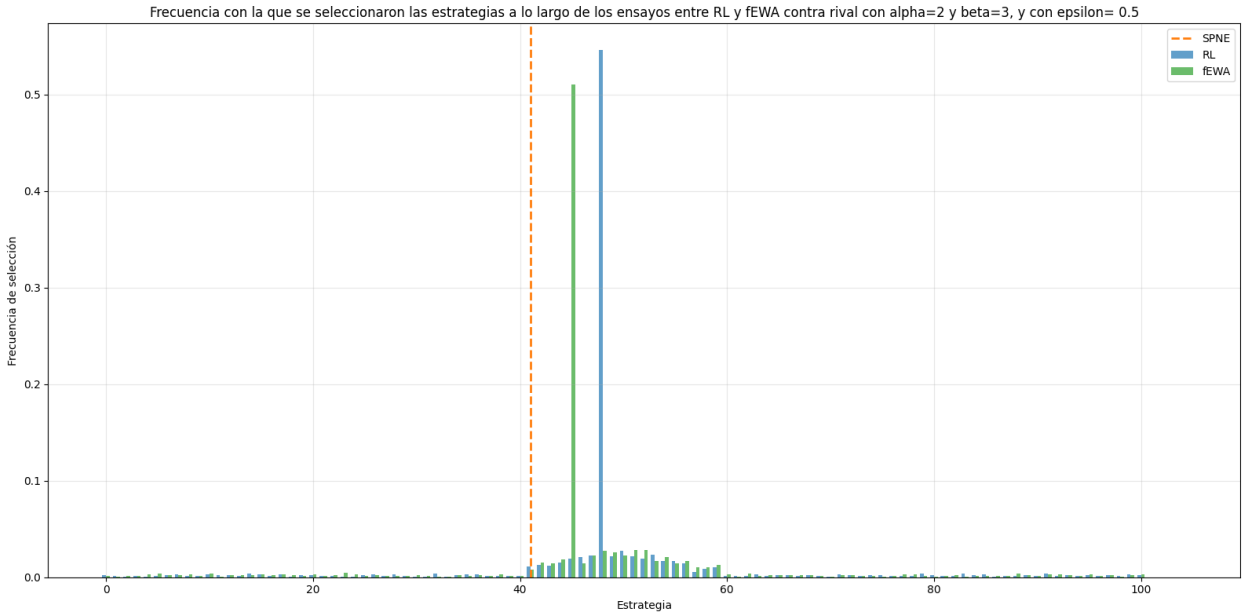

Figura 86: Frecuencia de selección de estrategias de RL y FEWA contra el rival con preferencia media por resultados justos cuando $\epsilon = 0{,}5$.

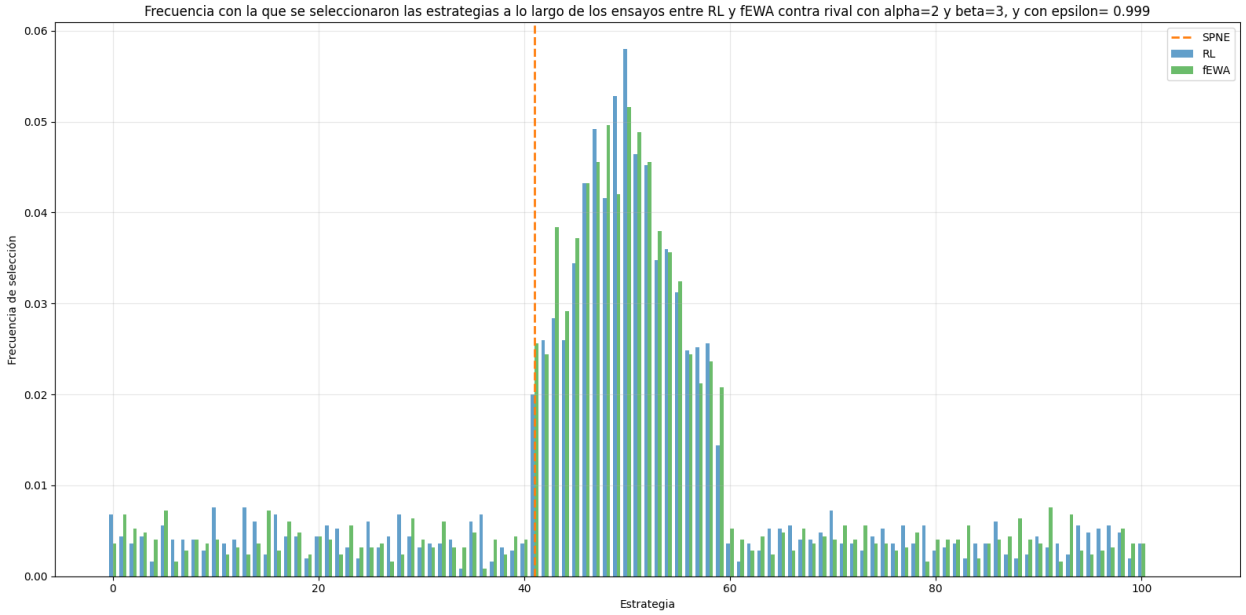

Figura 87: Frecuencia de selección de estrategias de RL y FEWA contra el rival con preferencia media por resultados justos cuando $\epsilon = 0{,}9$.

9.1.3. Valores de epsilon para rival con preferencia alta por resultados ventajosos

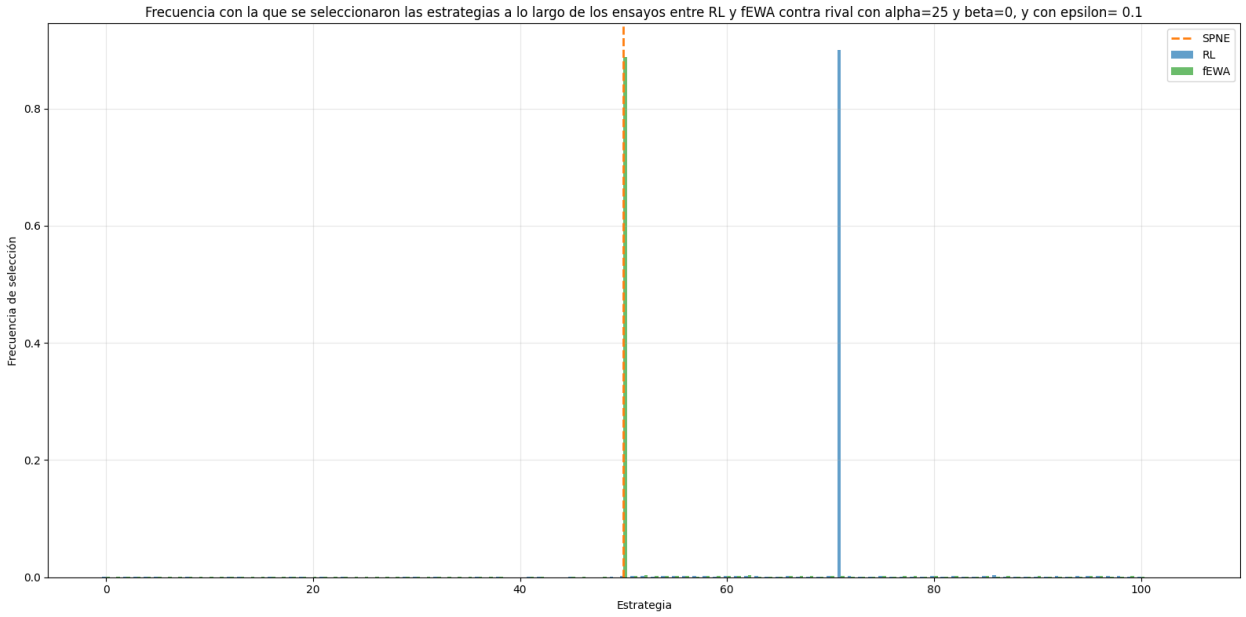

Figura 88: Frecuencia de selección de estrategias de RL y FEWA contra el rival con preferencia alta por resultados ventajosos cuando $\epsilon = 0{,}1$.

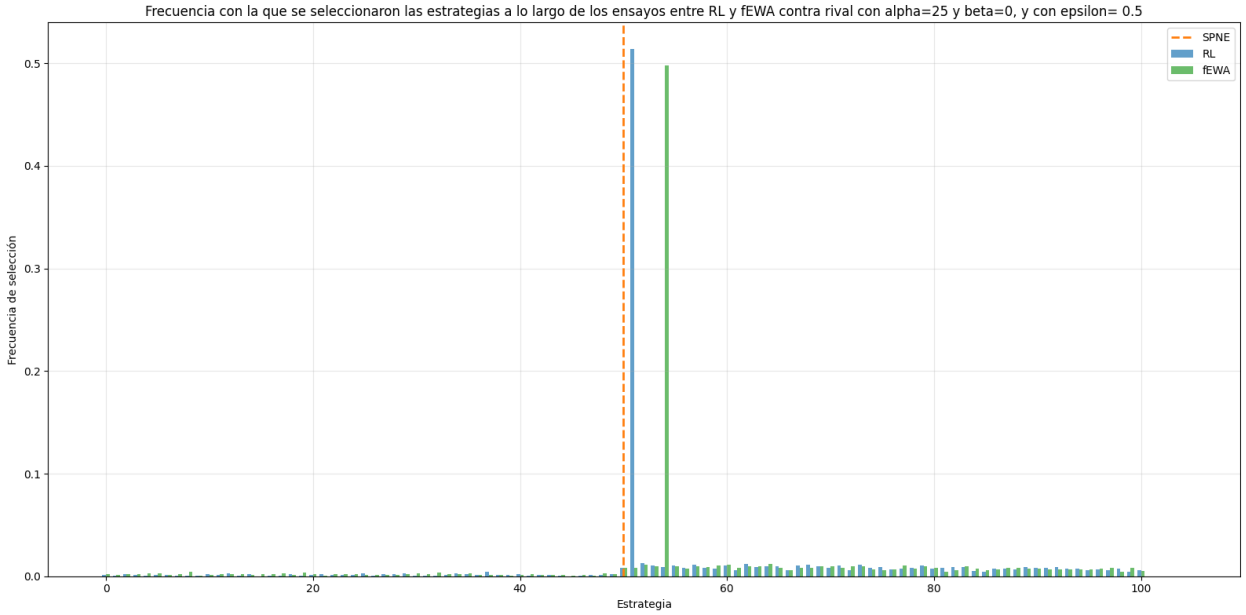

Figura 89: Frecuencia de selección de estrategias de RL y FEWA contra el rival con preferencia alta por resultados ventajosos cuando $\epsilon = 0{,}5$.

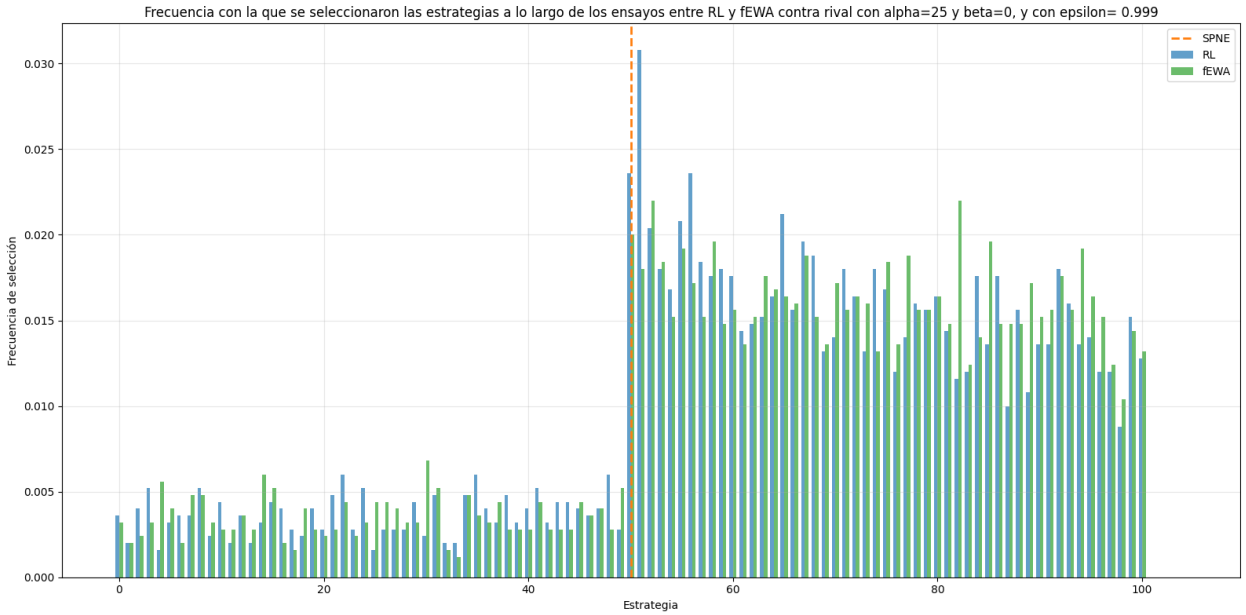

Figura 90: Frecuencia de selección de estrategias de RL y FEWA contra el rival con preferencia alta por resultados ventajosos cuando $\epsilon = 0{,}9$.

9.2. Gráficos de valor que asignaron RL y FEWA por rival

9.2.1. Valores de epsilon para rival con preferencia baja por resultados ventajosos

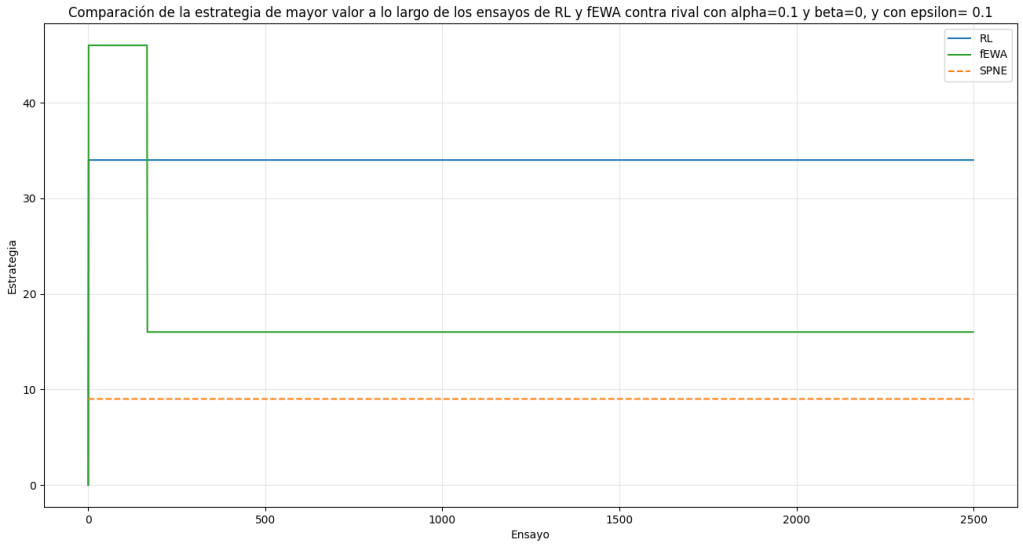

Figura 91: Valor normalizado que RL y FEWA le asignaron a cada estrategia contra el rival con preferencia baja por resultados ventajosos cuando $\epsilon = 0{,}1$.

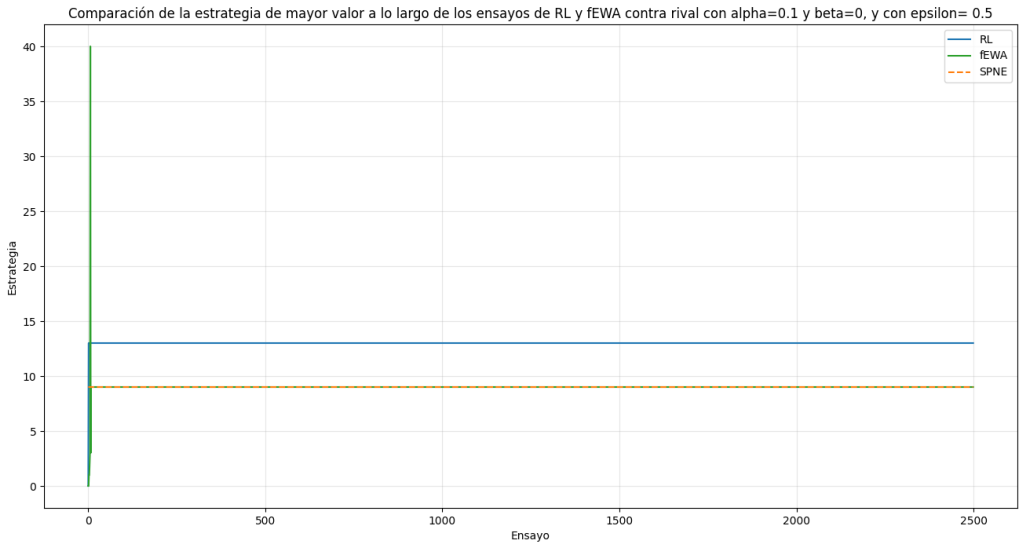

Figura 92: Valor normalizado que RL y FEWA le asignaron a cada estrategia contra el rival con preferencia baja por resultados ventajosos cuando $\epsilon = 0{,}5$.

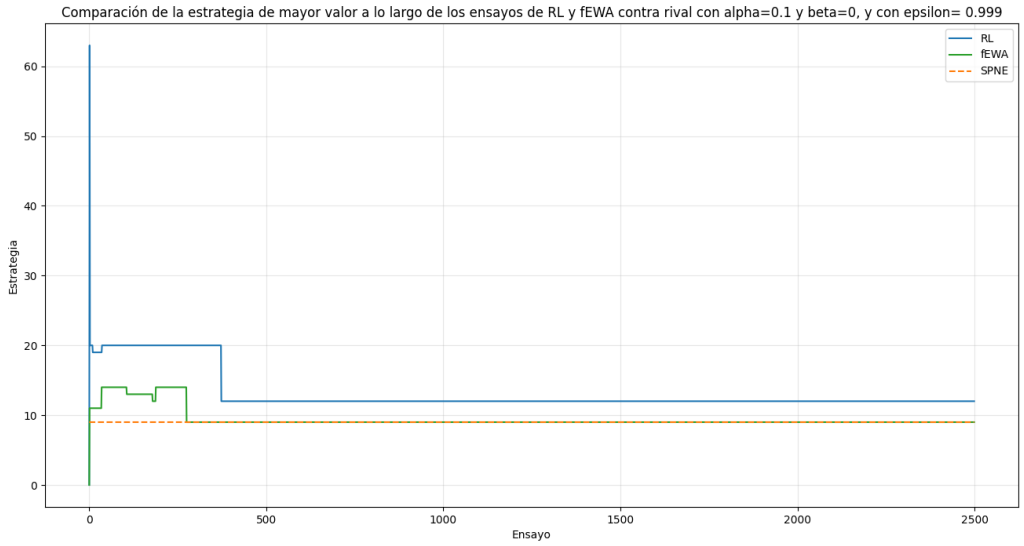

Figura 93: Valor normalizado que RL y FEWA le asignaron a cada estrategia contra el rival con preferencia baja por resultados ventajosos cuando $\epsilon = 0{,}9$.

9.2.2. Valores de epsilon para rival con preferencia media por resultados justos

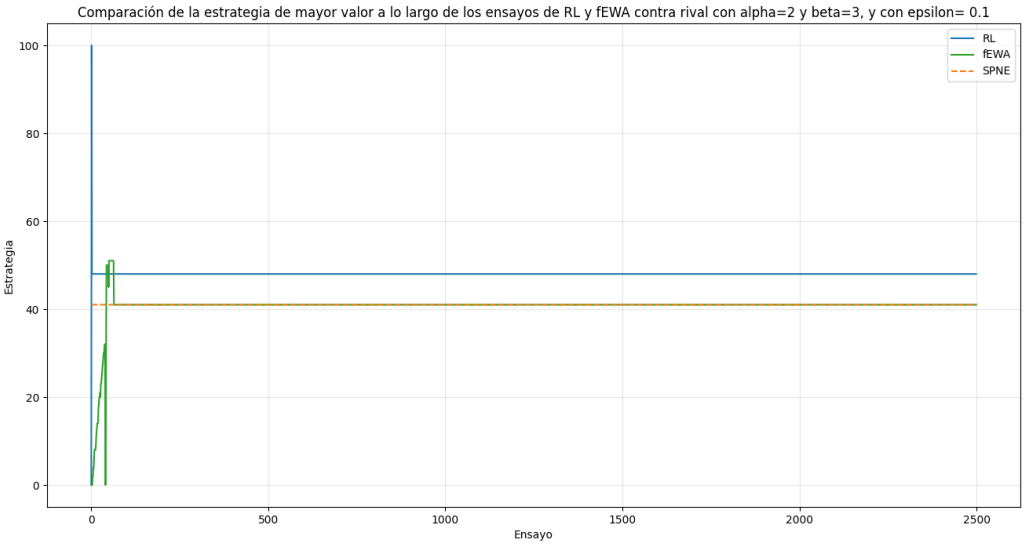

Figura 94: Valor normalizado que RL y FEWA le asignaron a cada estrategia contra el rival con preferencia media por resultados justos cuando $\epsilon = 0{,}1$.

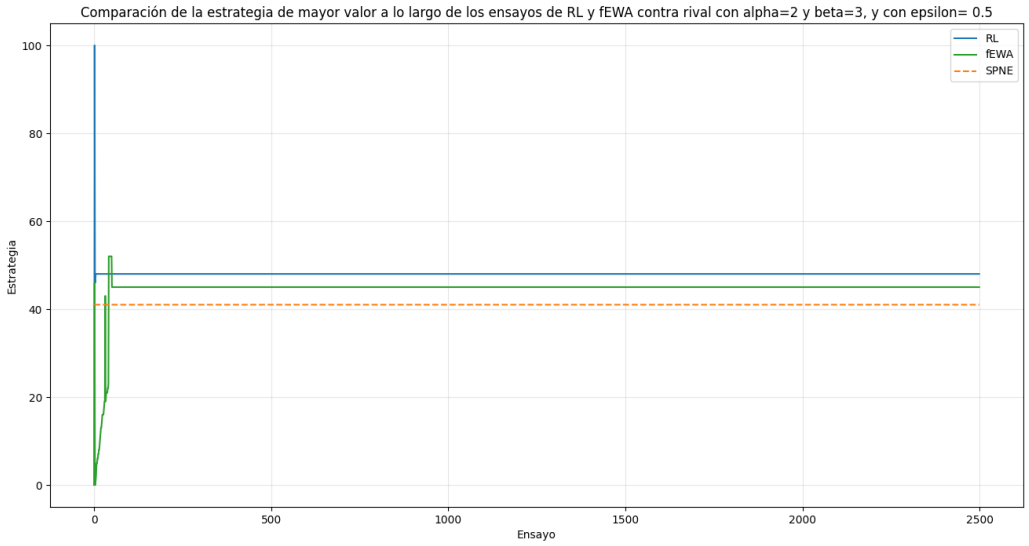

Figura 95: Valor normalizado que RL y FEWA le asignaron a cada estrategia contra el rival con preferencia media por resultados justos cuando $\epsilon = 0{,}5$.

Figura 96: Valor normalizado que RL y FEWA le asignaron a cada estrategia contra el rival con preferencia media por resultados justos cuando $\epsilon = 0{,}9$.

9.2.3. Valores de epsilon para rival con preferencia alta por resultados ventajosos

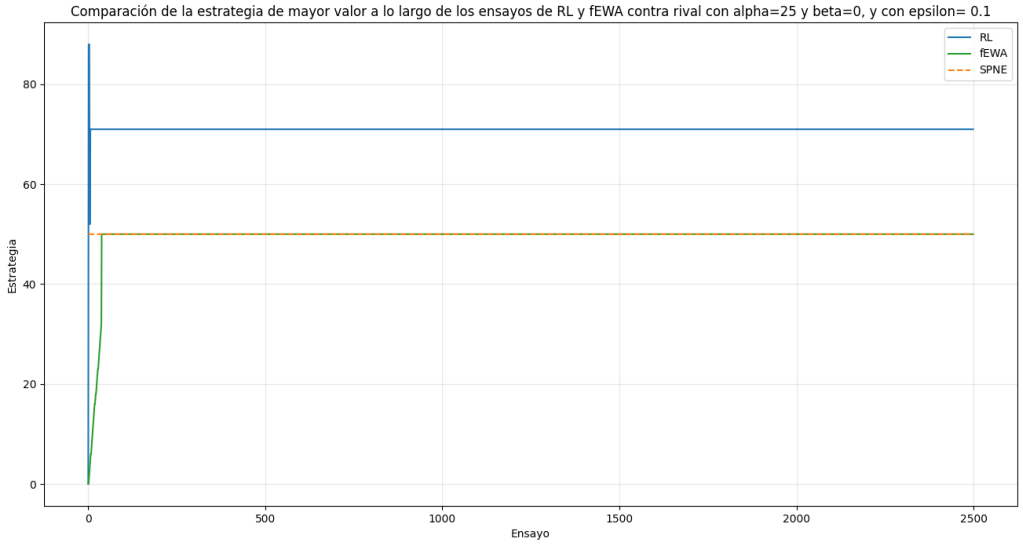

Figura 97: Valor normalizado que RL y FEWA le asignaron a cada estrategia contra el rival con preferencia alta por resultados ventajosos cuando $\epsilon = 0{,}1$.

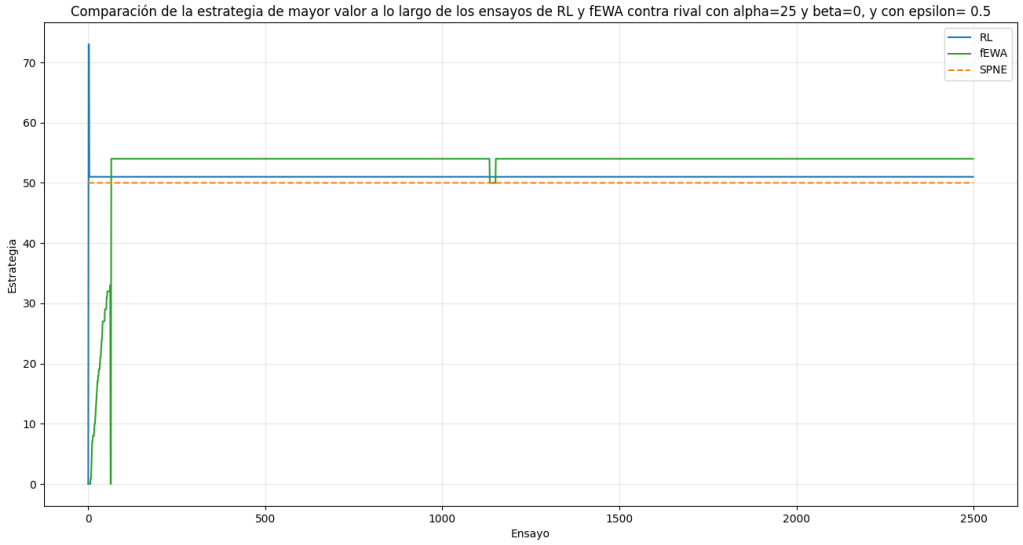

Figura 98: Valor normalizado que RL y FEWA le asignaron a cada estrategia contra el rival con preferencia alta por resultados ventajosos cuando $\epsilon = 0{,}5$.

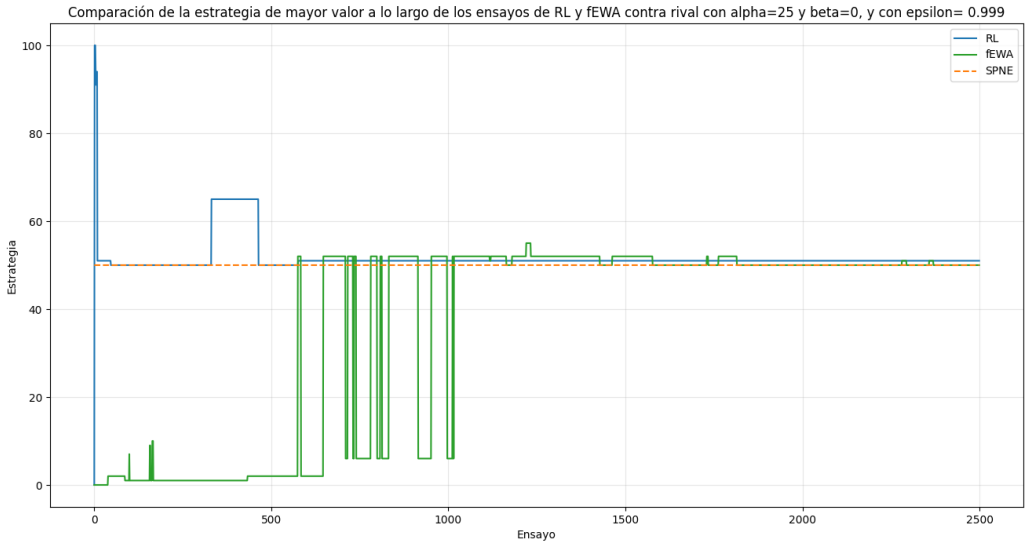

Figura 99: Valor normalizado que RL y FEWA le asignaron a cada estrategia contra el rival con preferencia alta por resultados ventajosos cuando $\epsilon = 0{,}9$.

9.3. Gráficos de acumulación de ganancias de RL y FEWA por rival y por tamaño de muestra

9.3.1. Valores de epsilon para rival con preferencia baja por resultados ventajosos

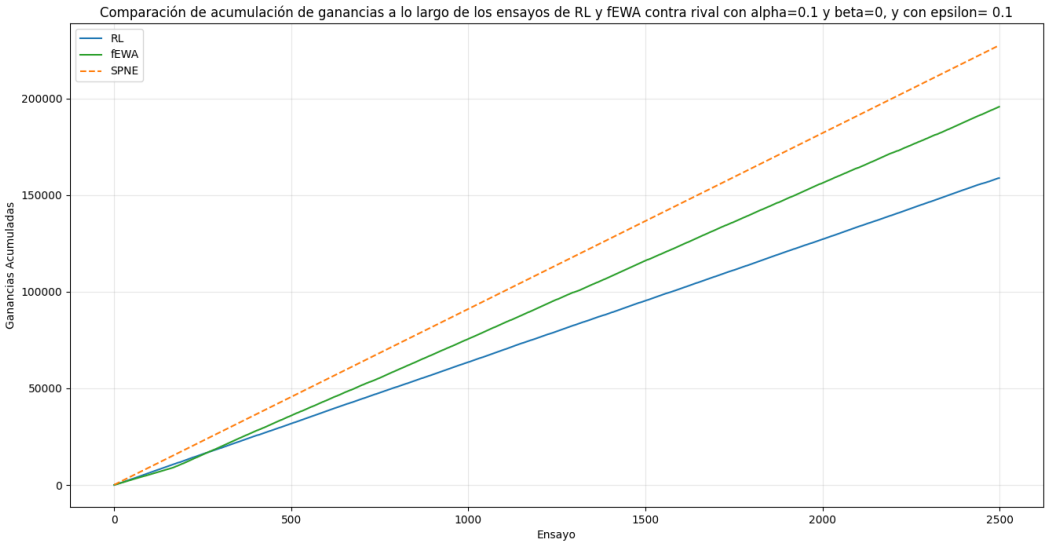

Figura 100: Ganancias acumuladas a lo largo de los ensayos de RL y FEWA contra el rival con preferencia baja por resultados ventajosos cuando $\epsilon = 0{,}1$.

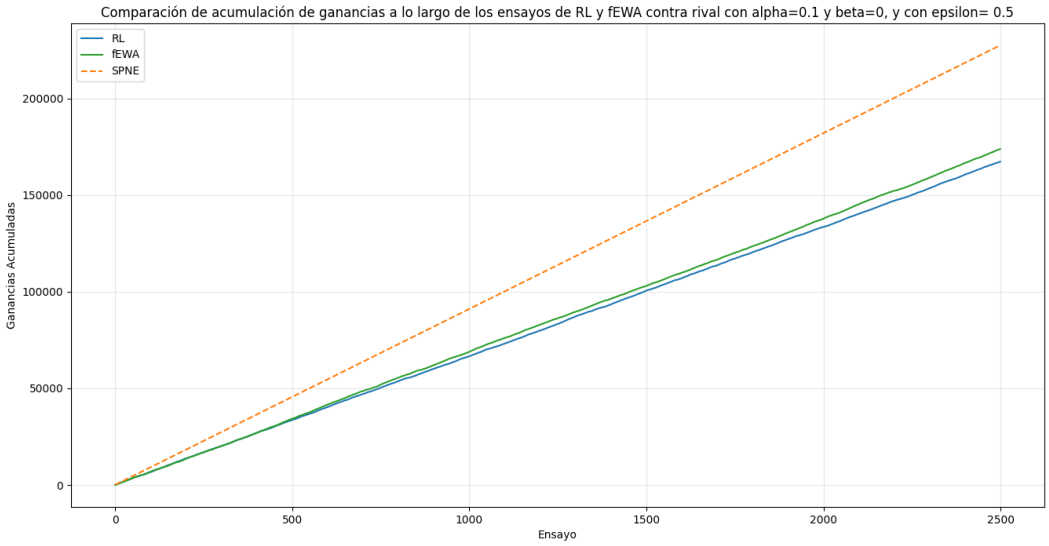

Figura 101: Ganancias acumuladas a lo largo de los ensayos de RL y FEWA contra el rival con preferencia baja por resultados ventajosos cuando $\epsilon = 0{,}5$.

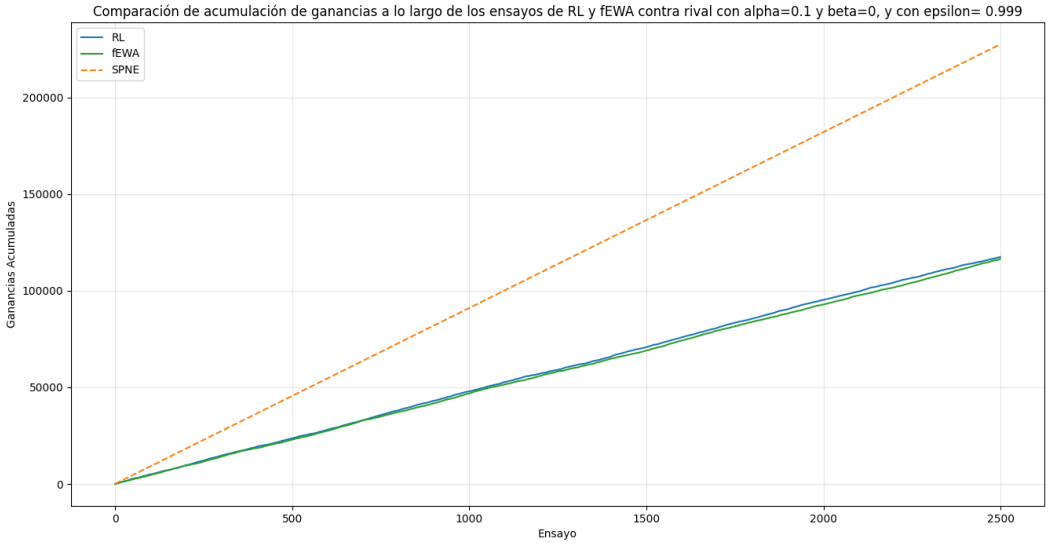

Figura 102: Ganancias acumuladas a lo largo de los ensayos de RL y FEWA contra el rival con preferencia baja por resultados ventajosos cuando $\epsilon = 0{,}9$.

9.3.2. Valores de epsilon para rival con preferencia media por resultados justos

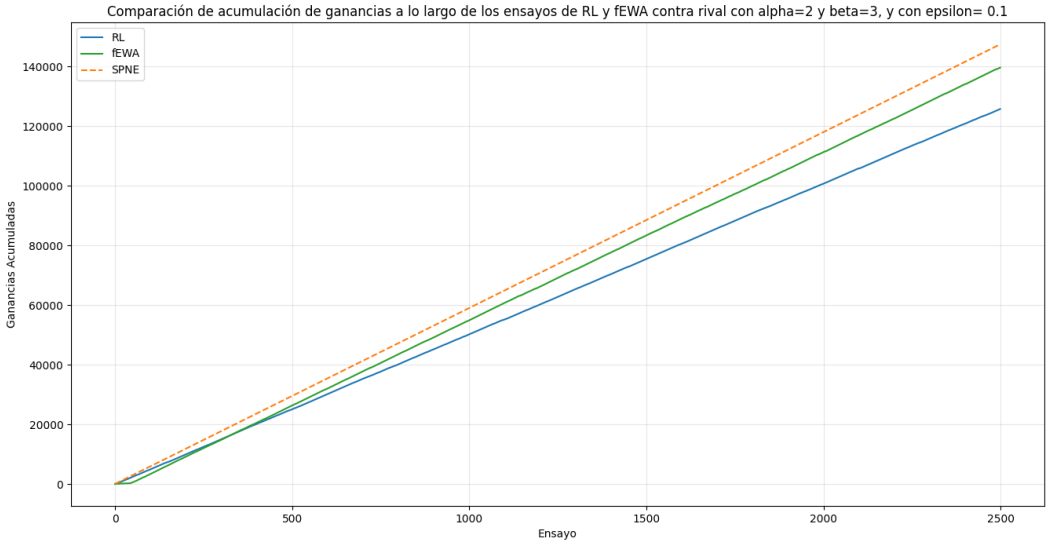

Figura 103: Ganancias acumuladas a lo largo de los ensayos de RL y FEWA contra el rival con preferencia media por resultados justos cuando $\epsilon = 0{,}1$.

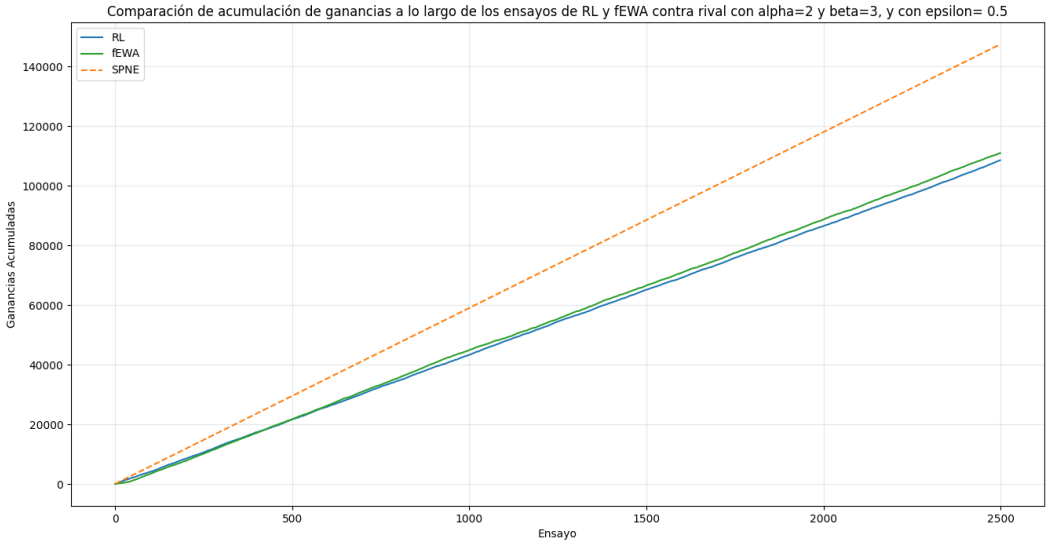

Figura 104: Ganancias acumuladas a lo largo de los ensayos de RL y FEWA contra el rival con preferencia media por resultados justos cuando $\epsilon = 0{,}5$.

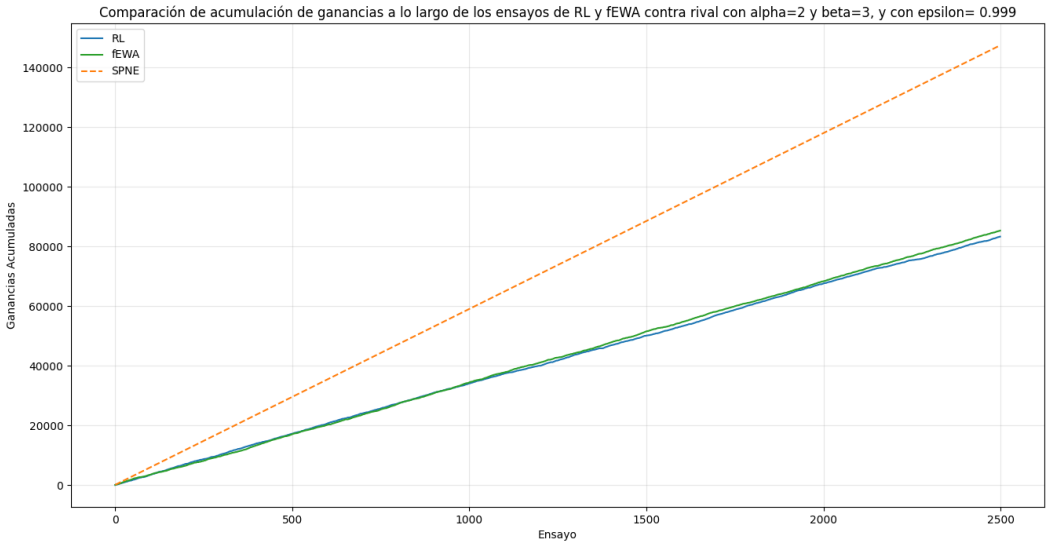

Figura 105: Ganancias acumuladas a lo largo de los ensayos de RL y FEWA contra el rival con preferencia media por resultados justos cuando $\epsilon = 0{,}9$.

9.3.3. Valores de epsilon para rival con preferencia alta por resultados ventajosos

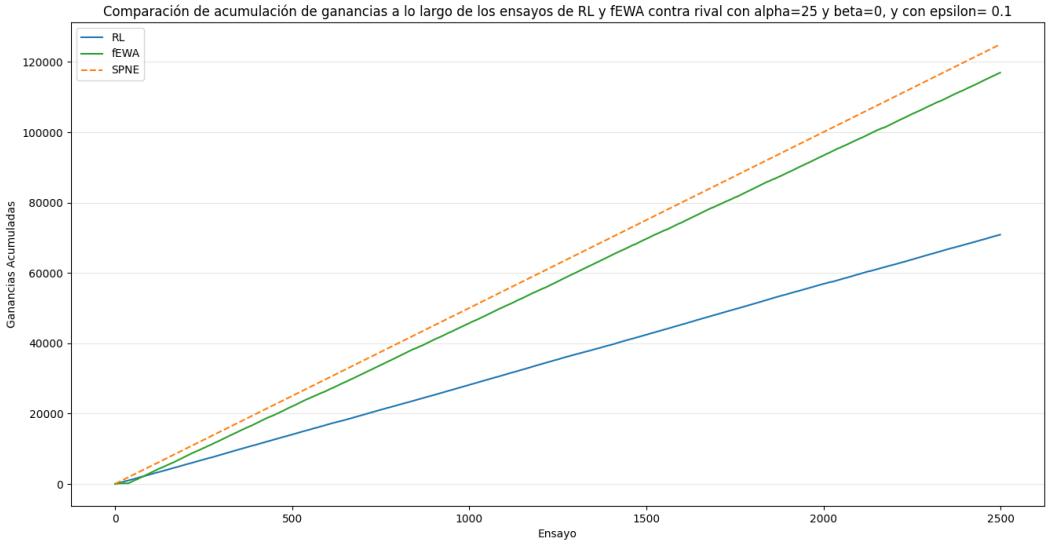

Figura 106: Ganancias acumuladas a lo largo de los ensayos de RL y FEWA contra el rival con preferencia alta por resultados ventajosos cuando $\epsilon = 0{,}1$.

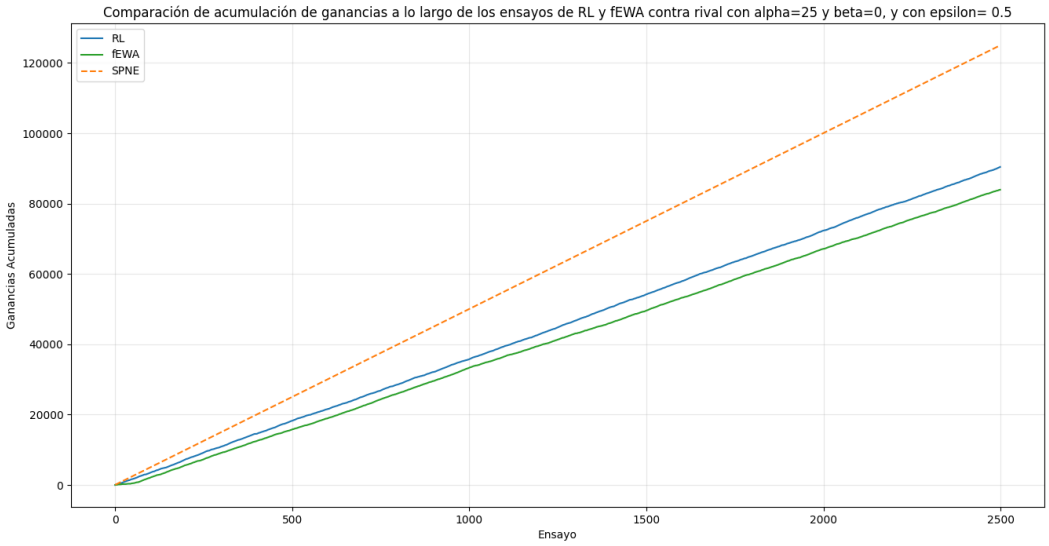

Figura 107: Ganancias acumuladas a lo largo de los ensayos de RL y FEWA contra el rival con preferencia alta por resultados ventajosos cuando $\epsilon = 0{,}5$.

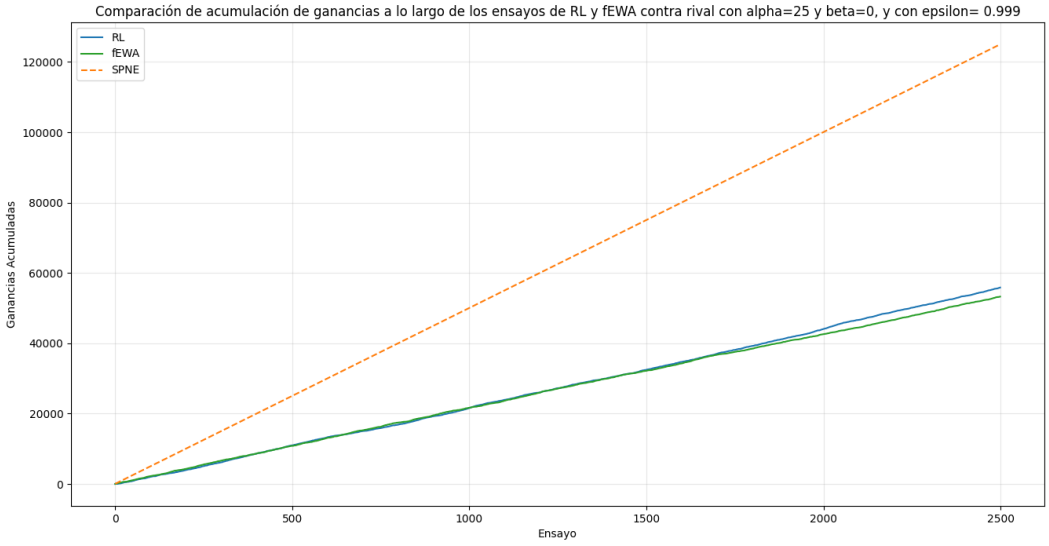

Figura 108: Ganancias acumuladas a lo largo de los ensayos de RL y FEWA contra el rival con preferencia alta por resultados ventajosos cuando $\epsilon = 0{,}9$.

9.4. Aprendizaje de RL y FEWA por rival, por epsilon y por ronda

9.4.1. Valores de epsilon para rival con preferencia baja por resultados ventajosos en ronda 1

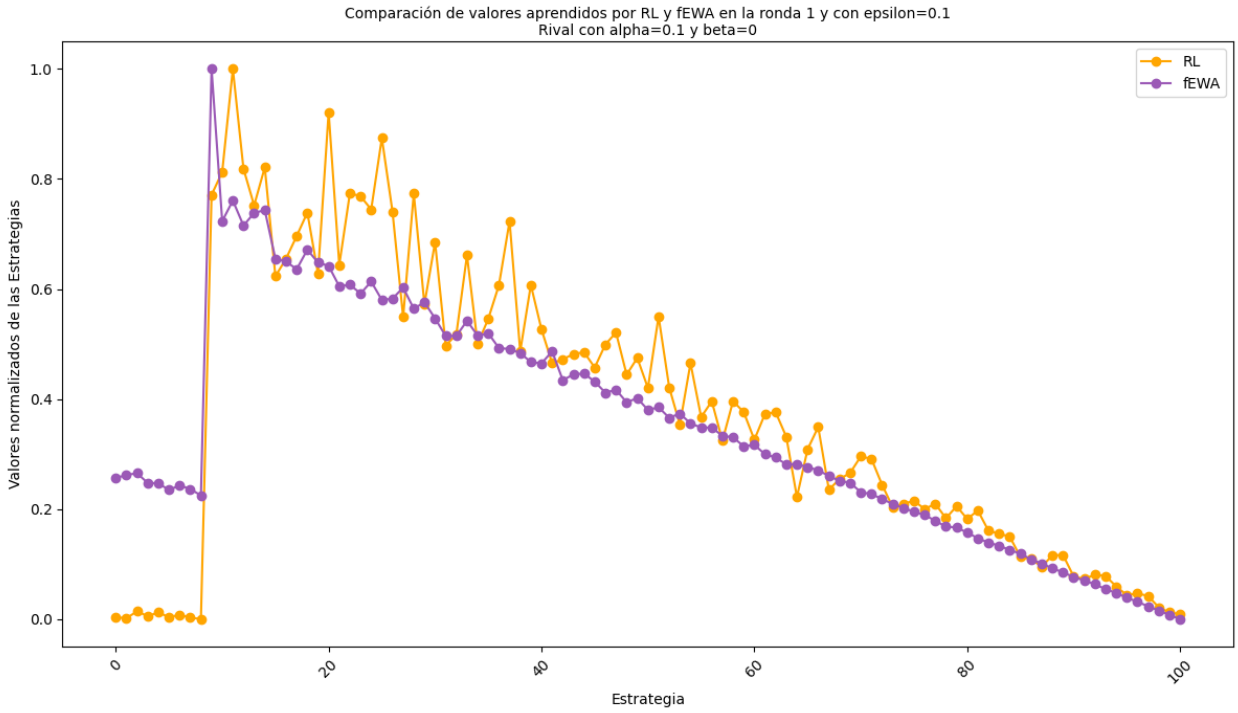

Figura 109: Valor normalizado del aprendizaje de estrategias de RL y FEWA en la ronda 1 contra el rival con preferencia baja por resultados ventajosos cuando $\epsilon = 0{,}1$.

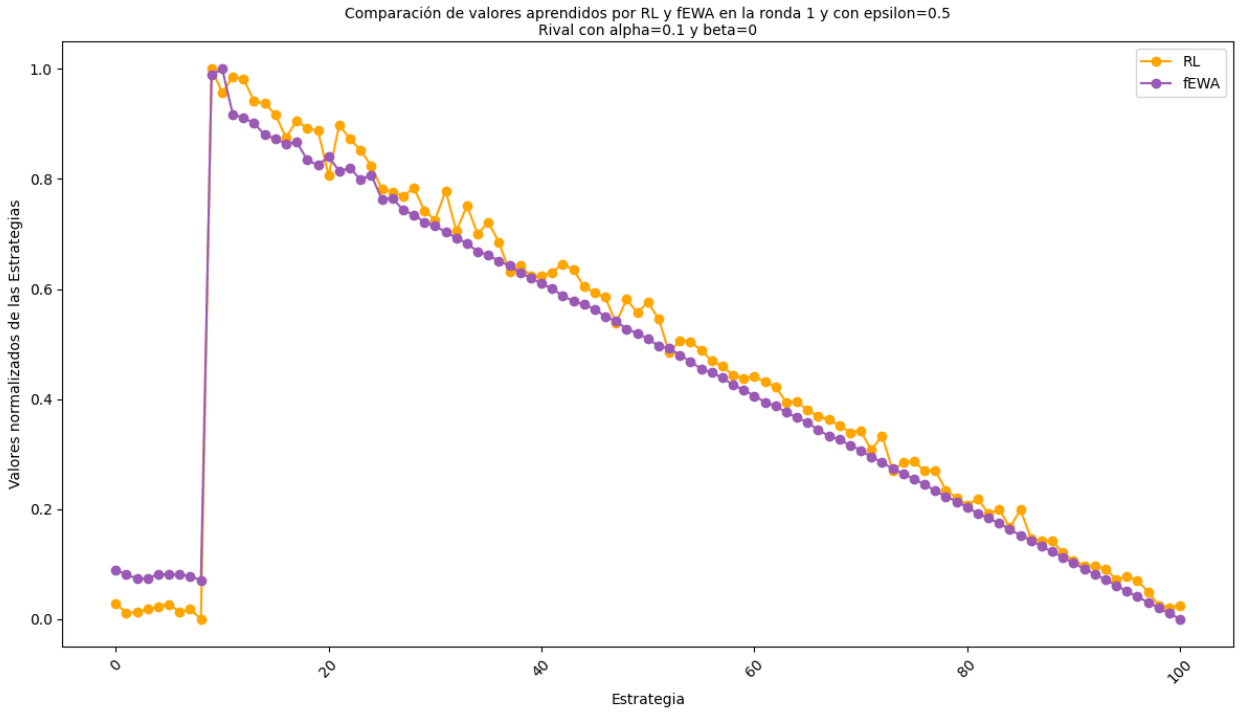

Figura 110: Valor normalizado del aprendizaje de estrategias de RL y FEWA en la ronda 1 contra el rival con preferencia baja por resultados ventajosos cuando $\epsilon = 0{,}5$.

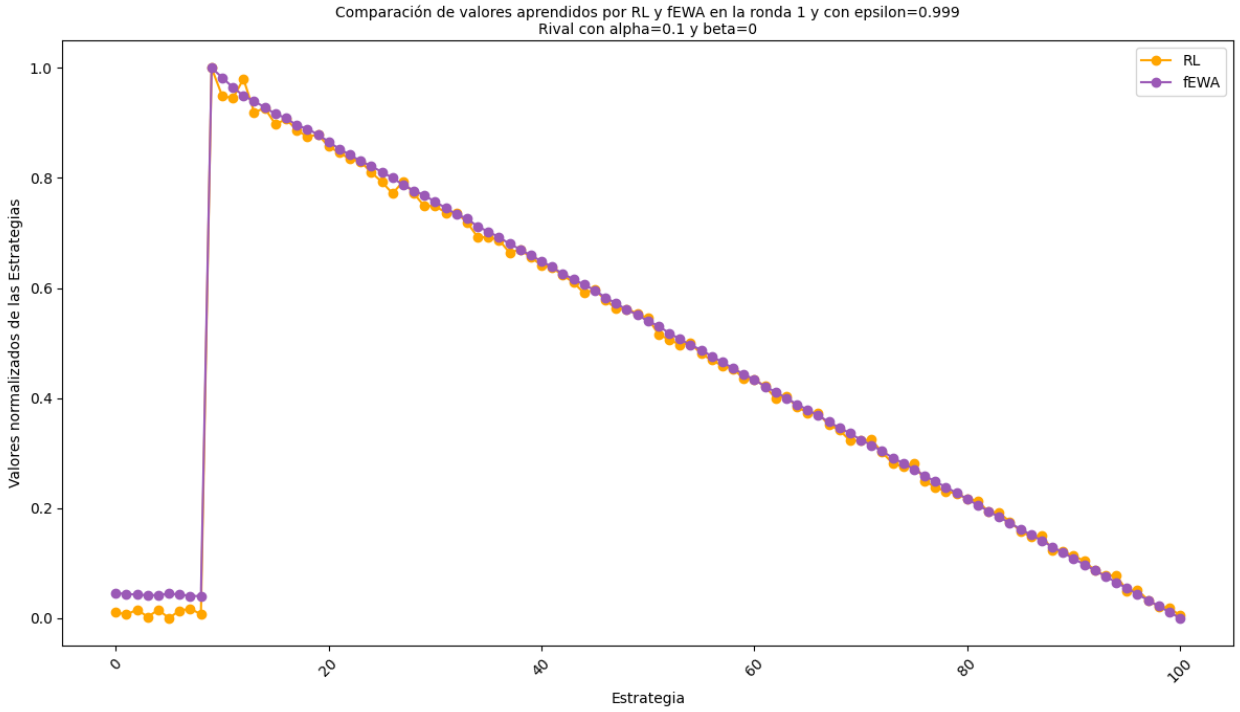

Figura 111: Valor normalizado del aprendizaje de estrategias de RL y FEWA en la ronda 1 contra el rival con preferencia baja por resultados ventajosos cuando $\epsilon = 0{,}9$.

9.4.2. Valores de epsilon para rival con preferencia media por resultados justos en ronda 1

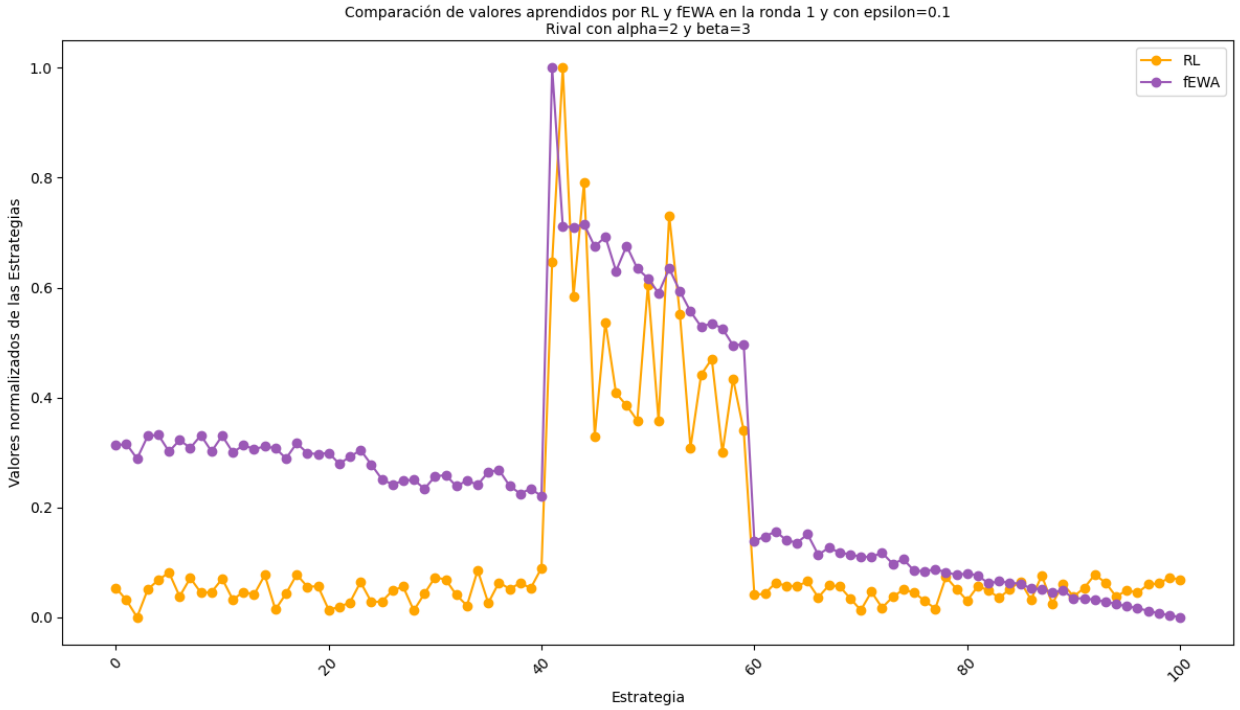

Figura 112: Valor normalizado del aprendizaje de estrategias de RL y FEWA en la ronda 1 contra el rival con preferencia media por resultados justos cuando $\epsilon = 0{,}1$.

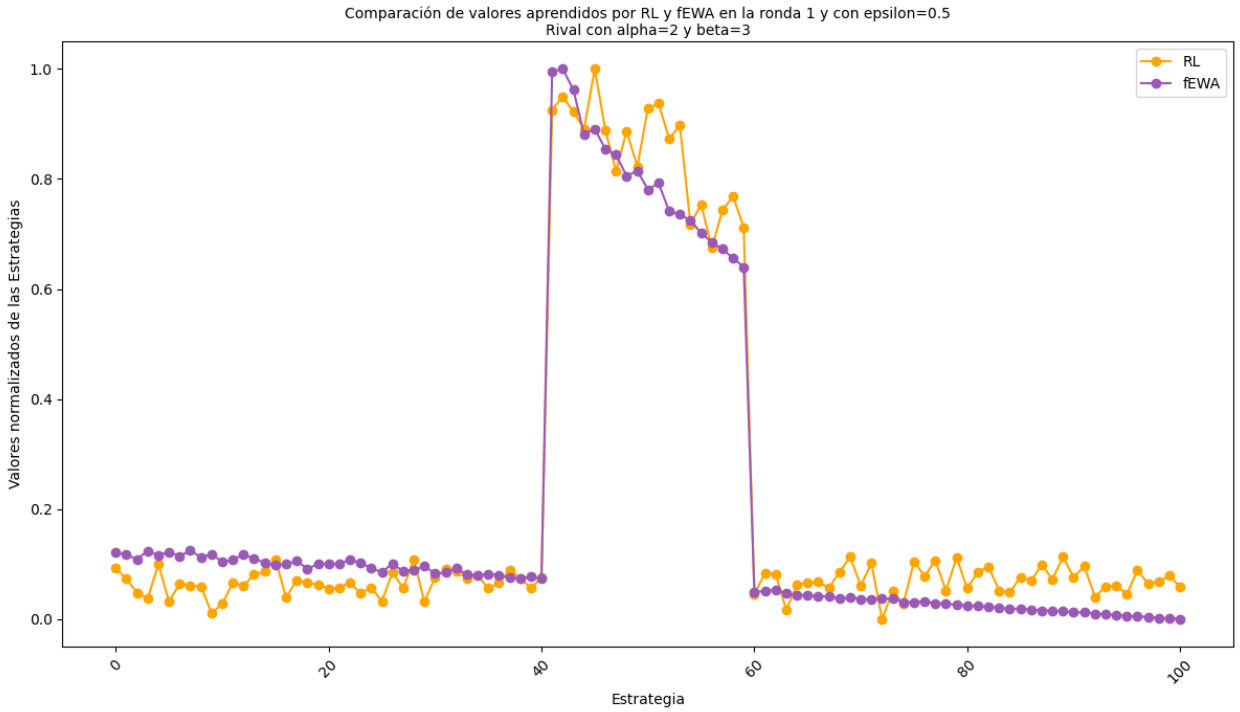

Figura 113: Valor normalizado del aprendizaje de estrategias de RL y FEWA en la ronda 1 contra el rival con preferencia media por resultados justos cuando $\epsilon = 0{,}5$.

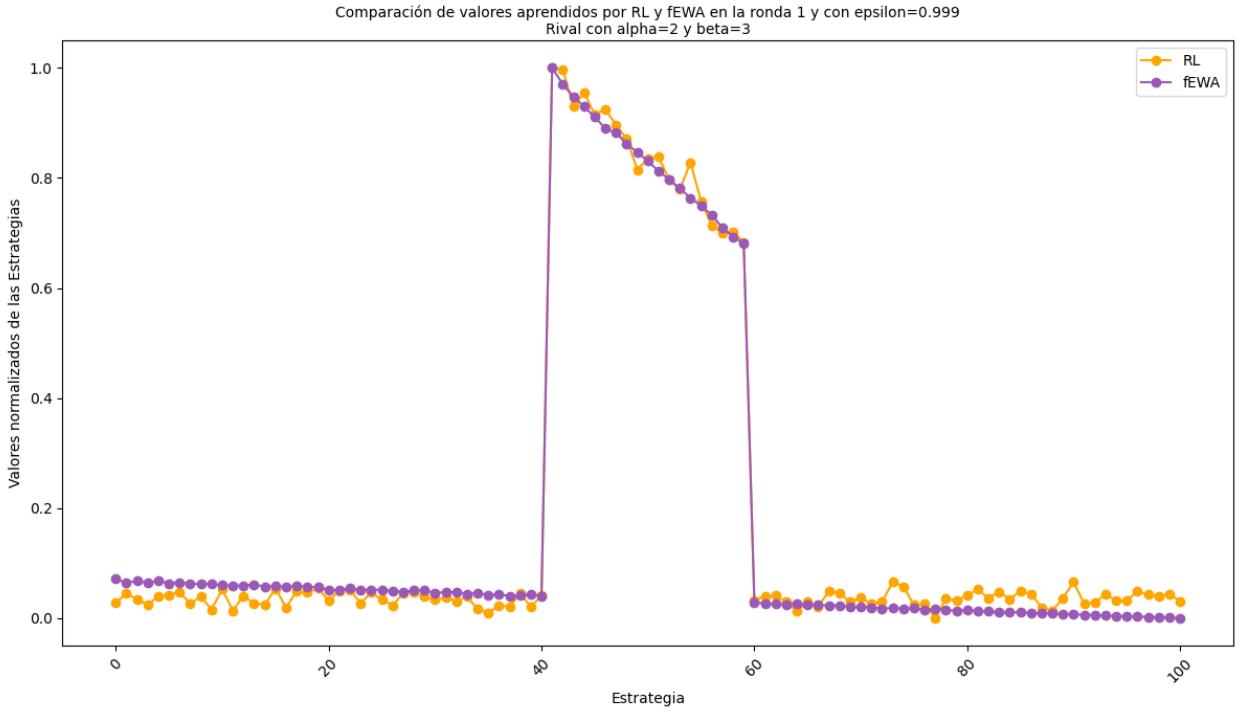

Figura 114: Valor normalizado del aprendizaje de estrategias de RL y FEWA en la ronda 1 contra el rival con preferencia media por resultados justos cuando $\epsilon = 0{,}9$.

9.4.3. Valores de epsilon para rival con preferencia alta por resultados ventajosos en ronda 1

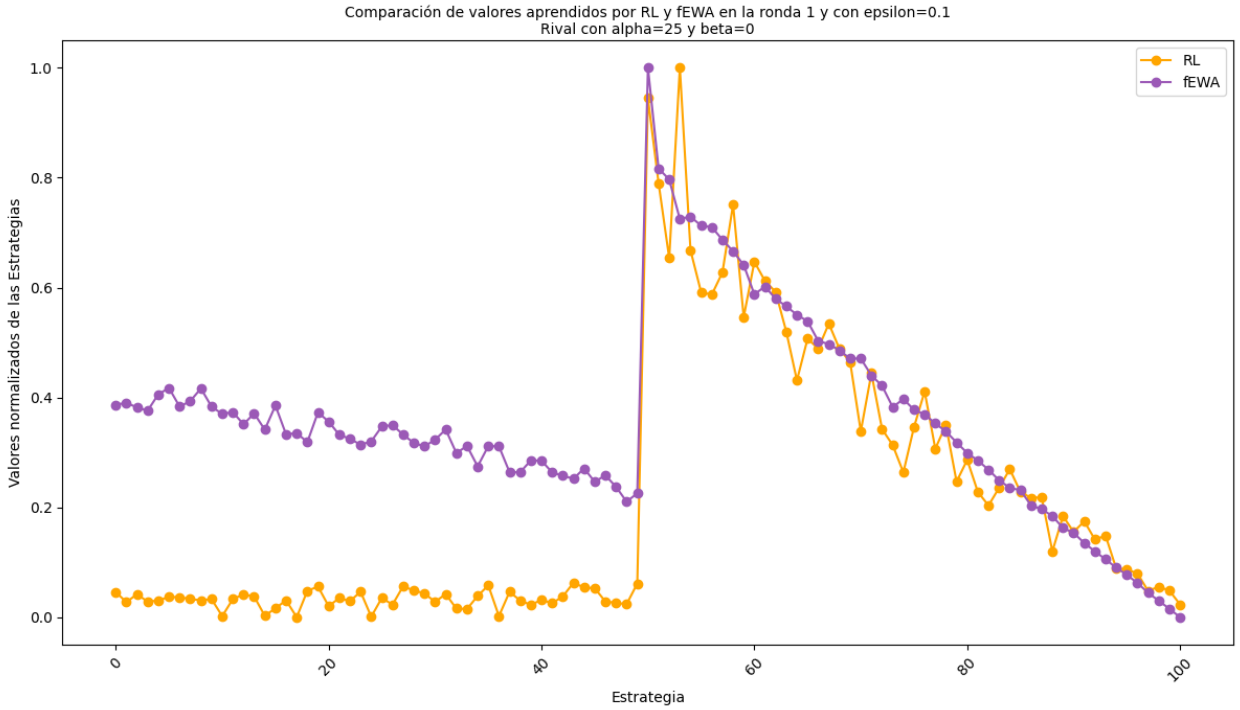

Figura 115: Valor normalizado del aprendizaje de estrategias de RL y FEWA en la ronda 1 contra el rival con preferencia alta por resultados ventajosos cuando $\epsilon = 0{,}1$.

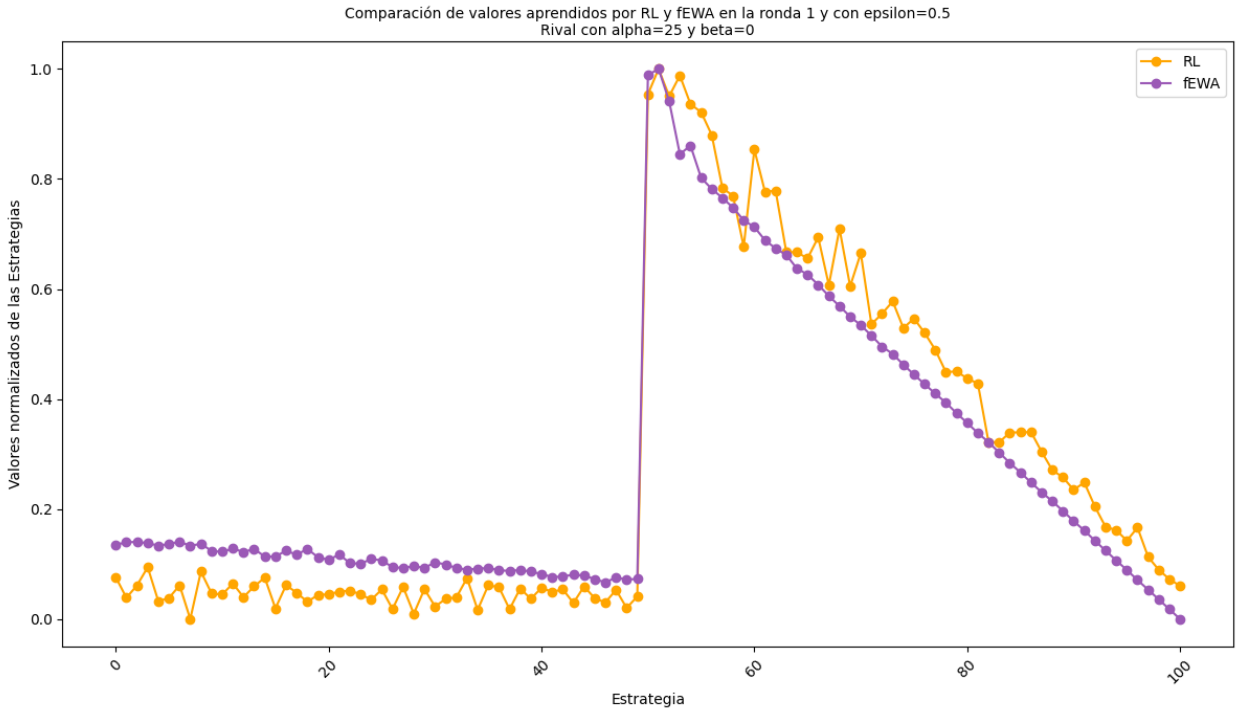

Figura 116: Valor normalizado del aprendizaje de estrategias de RL y FEWA en la ronda 1 contra el rival con preferencia alta por resultados ventajosos cuando $\epsilon = 0{,}5$.

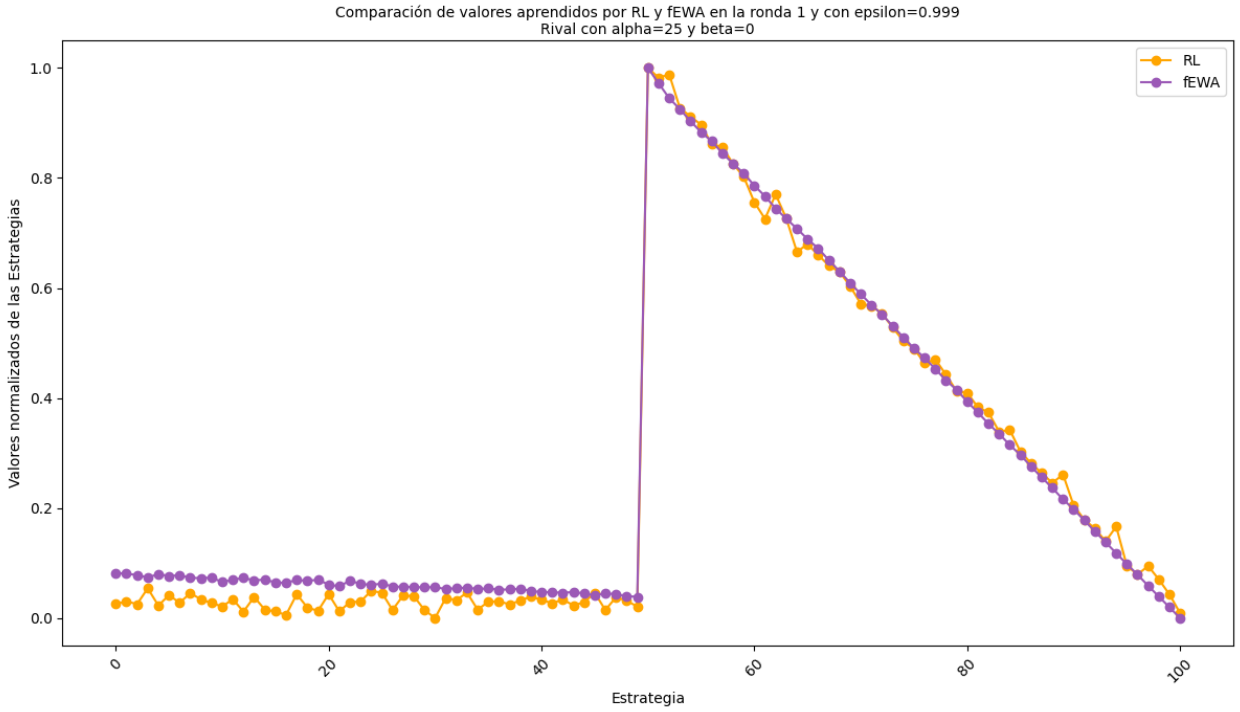

Figura 117: Valor normalizado del aprendizaje de estrategias de RL y FEWA en la ronda 1 contra el rival con preferencia alta por resultados ventajosos cuando $\epsilon = 0{,}9$.

9.4.4. Epsilon promediado para rival con preferencia baja por resultados ventajosos en ronda 2

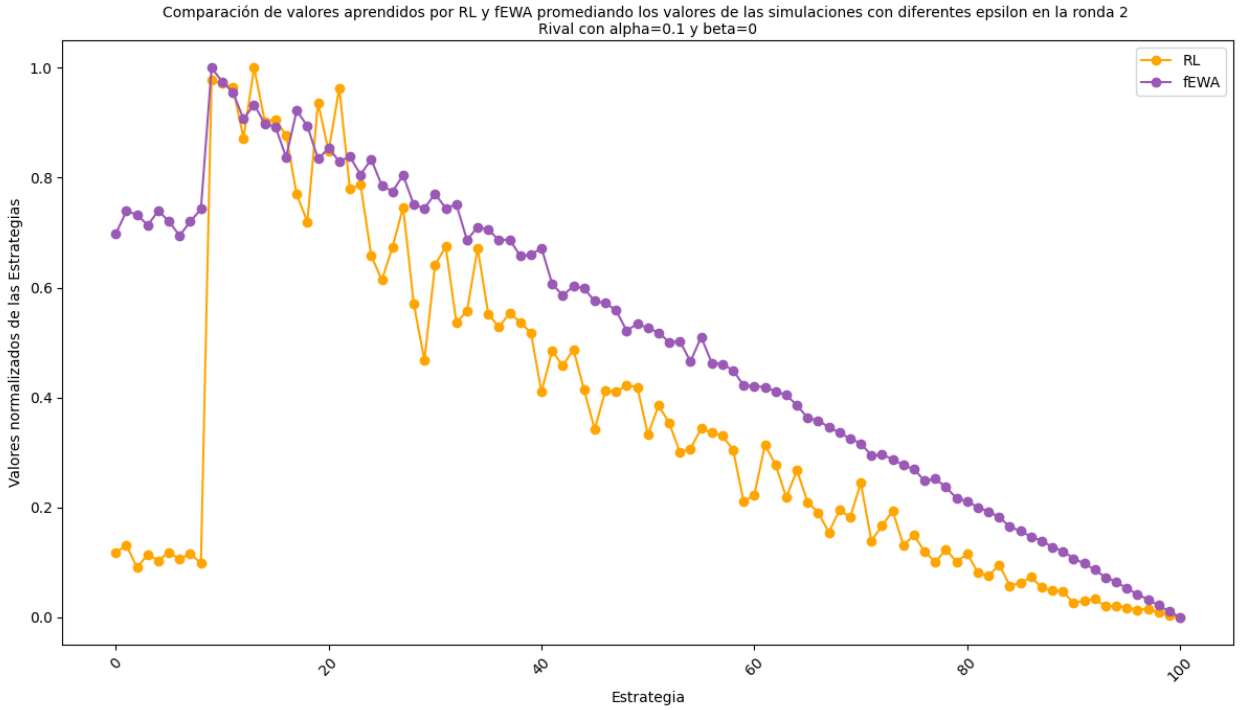

Figura 118: Valor normalizado del aprendizaje de estrategias de RL y FEWA en la ronda 2 contra el rival con preferencia baja a resultados ventajosos cuando se promedia ϵ.

9.4.5. Epsilon promediado para rival con preferencia media por resultados justos en ronda 2

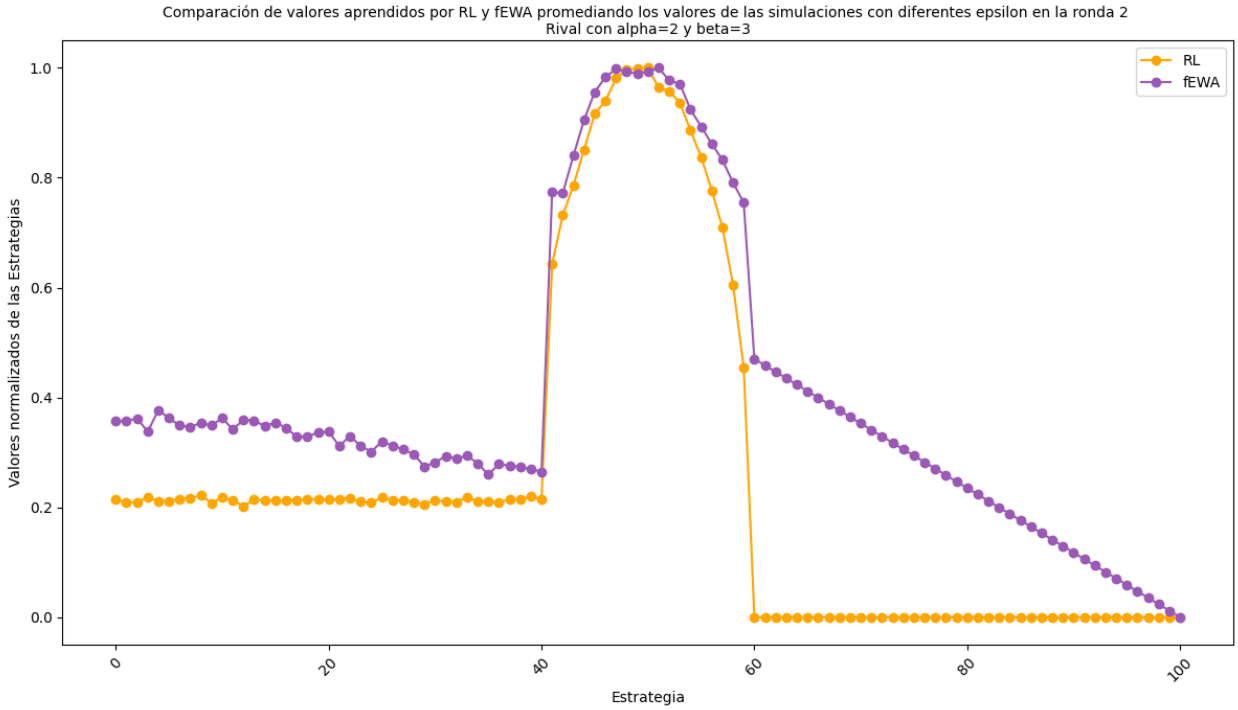

Figura 119: Valor normalizado del aprendizaje de estrategias de RL y FEWA en la ronda 2 contra el rival con preferencia media a resultados justos cuando se promedia ϵ.

9.4.6. Epsilon promediado para rival con preferencia alta por resultados ventajosos en ronda 2

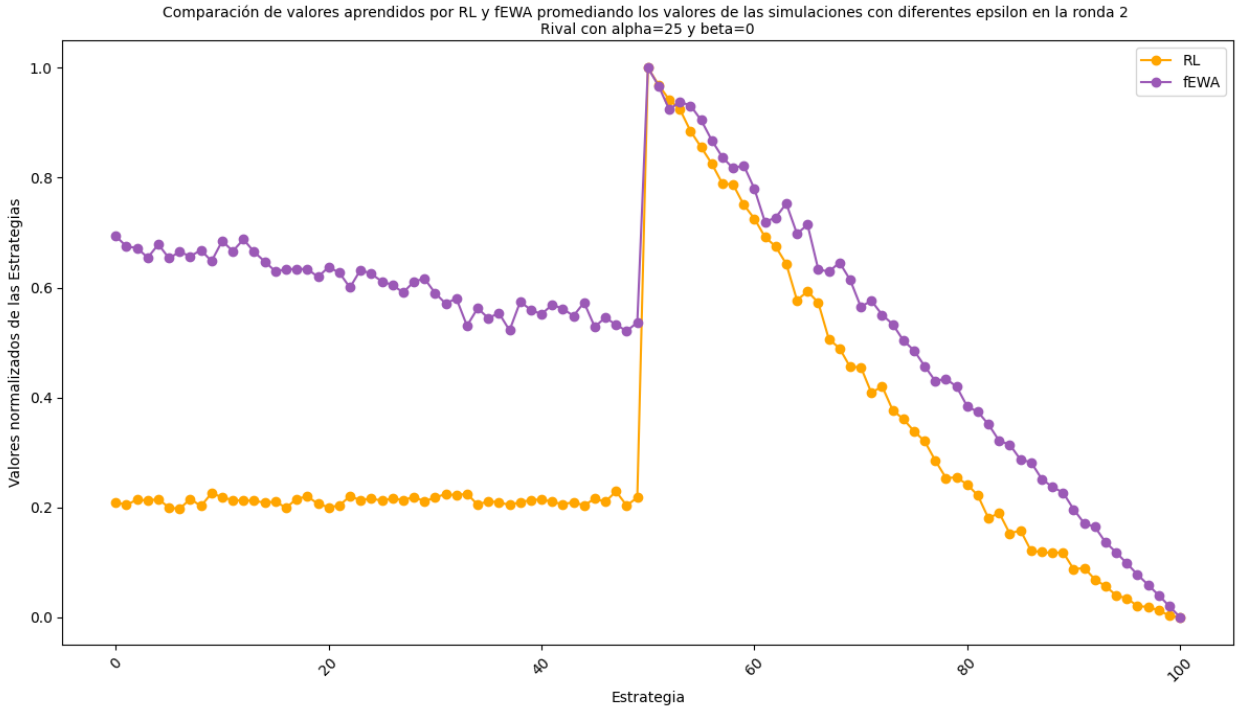

Figura 120: Valor normalizado del aprendizaje de estrategias de RL y FEWA en la ronda 2 contra el rival con preferencia alta por resultados ventajosos cuando se promedia ϵ.

9.4.7. Epsilon promediado para rival con preferencia baja por resultados ventajosos en ronda 3

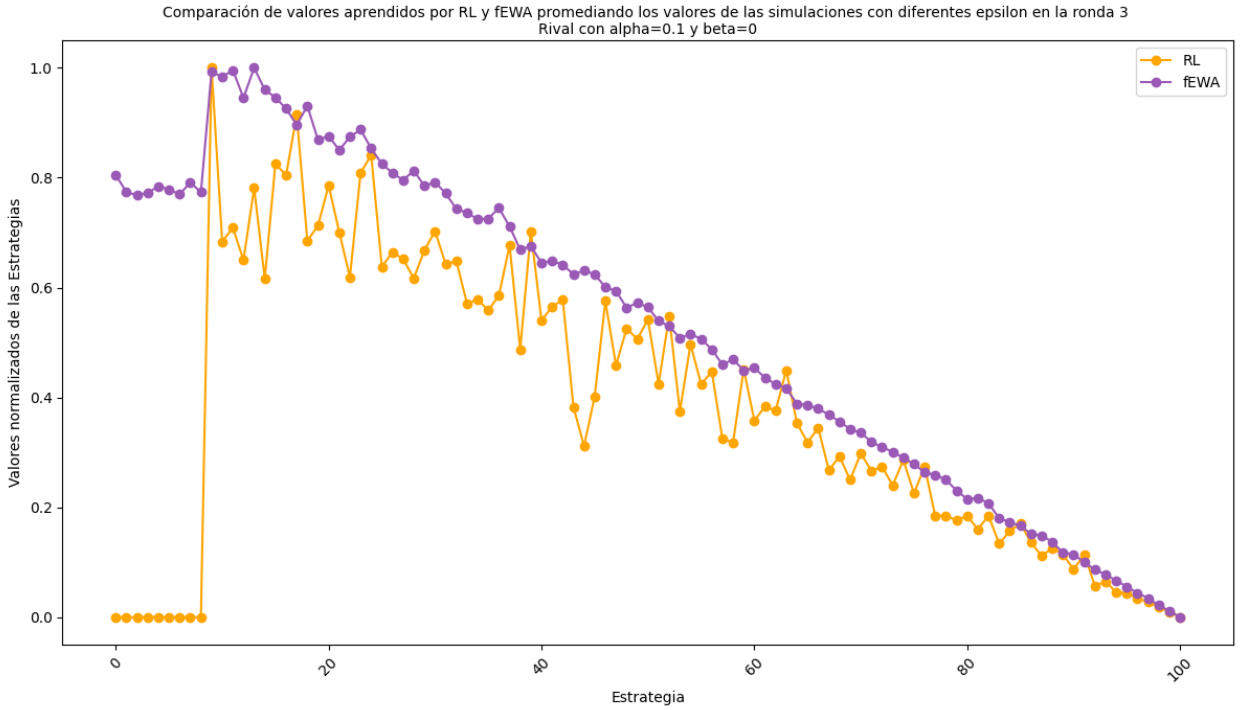

Figura 121: Valor normalizado del aprendizaje de estrategias de RL y FEWA en la ronda 3 contra el rival con preferencia baja por resultados ventajosos cuando se promedia ϵ.

9.4.8. Epsilon promediado para rival con preferencia media por resultados justos en ronda 3

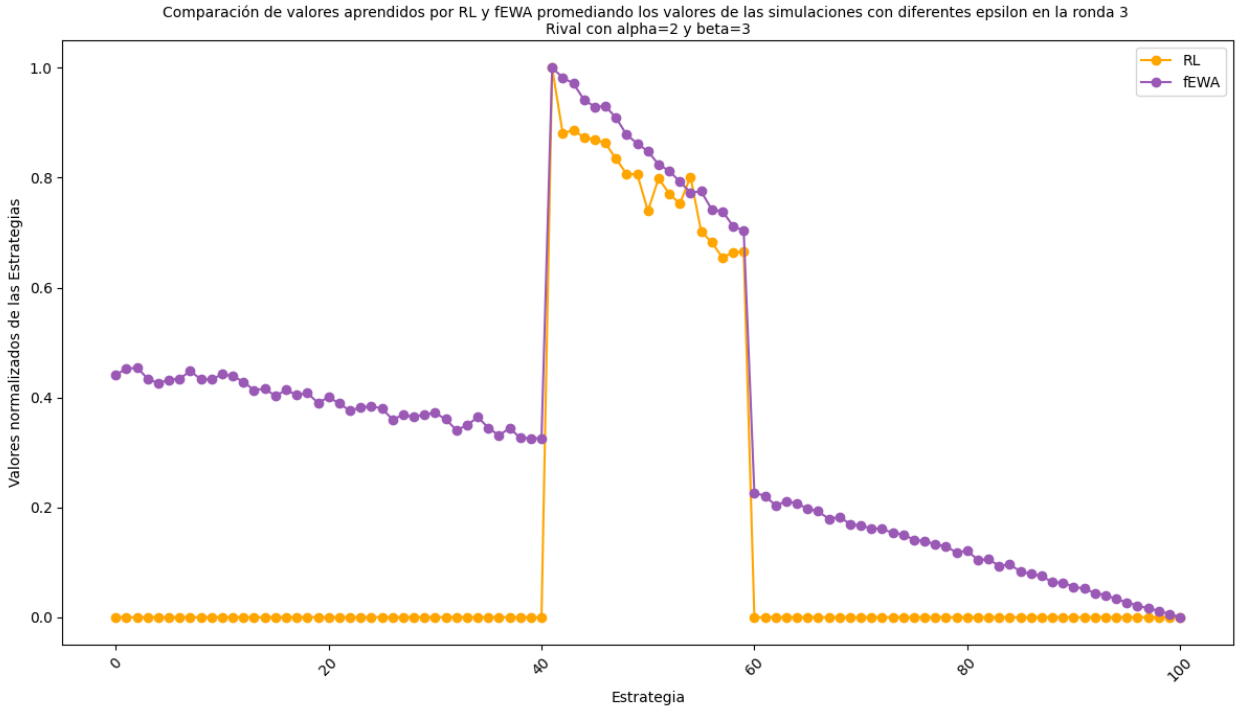

Figura 122: Valor normalizado del aprendizaje de estrategias de RL y FEWA en la ronda 3 contra el rival con preferencia media a resultados justos cuando se promedia ϵ.

9.4.9. Epsilon promediado para rival con preferencia alta por resultados ventajosos en ronda 3

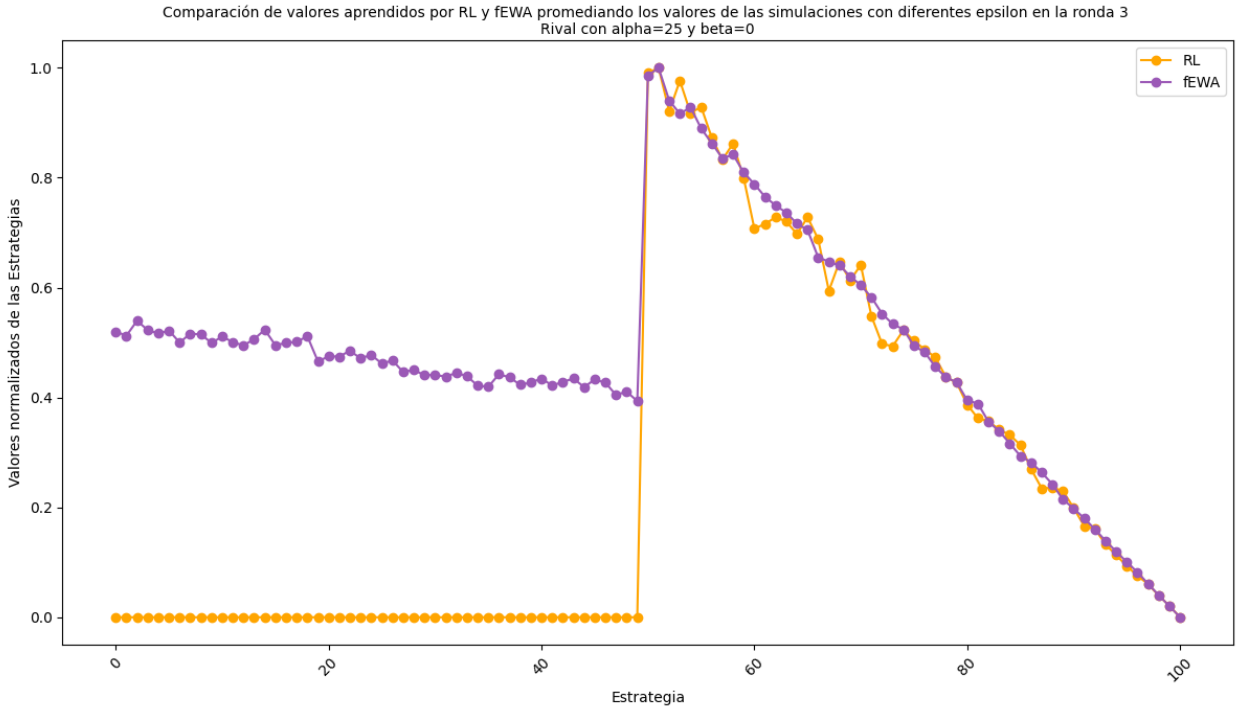

Figura 123: Valor normalizado del aprendizaje de estrategias de RL y FEWA en la ronda 3 contra el rival con preferencia alta por resultados ventajosos cuando se promedia ϵ.

9.5. Gráficos de Recall y F1-score de RL y FEWA

9.5.1. Recall

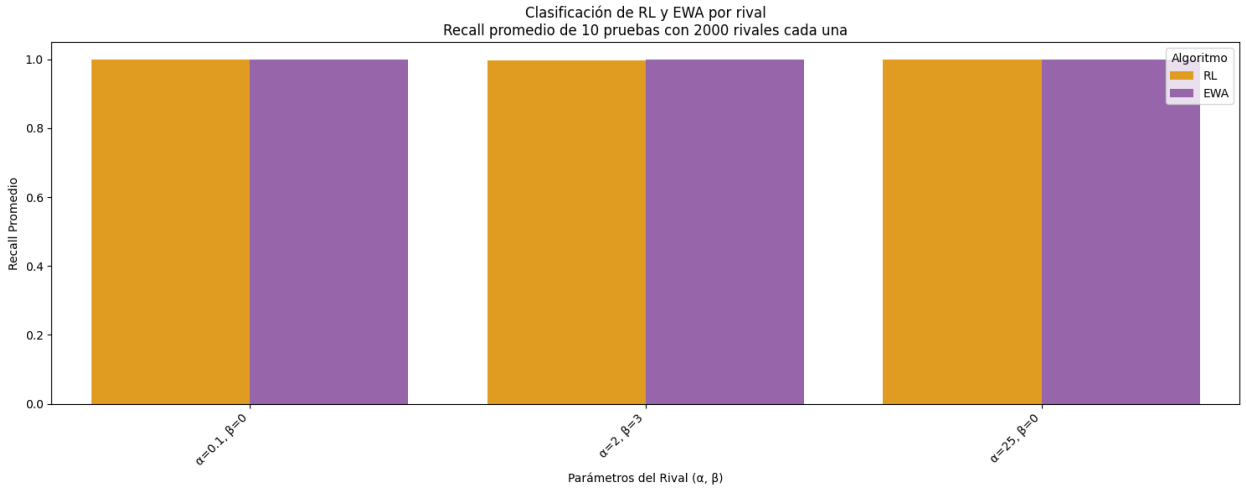

Figura 124: Recall al clasificar 3 rivales cuando ambos agentes aprendieron 3 rivales.

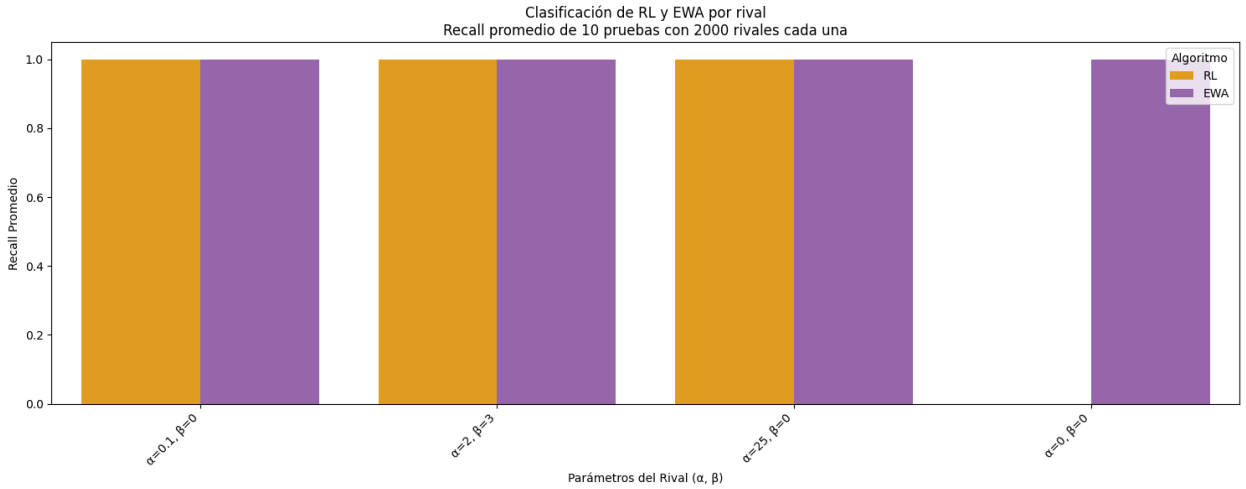

Figura 125: Recall al clasificar 4 rivales cuando ambos agentes aprendieron 3 rivales y se agrega un rival nuevo con un rango de ofertas aceptable muy distinguible.

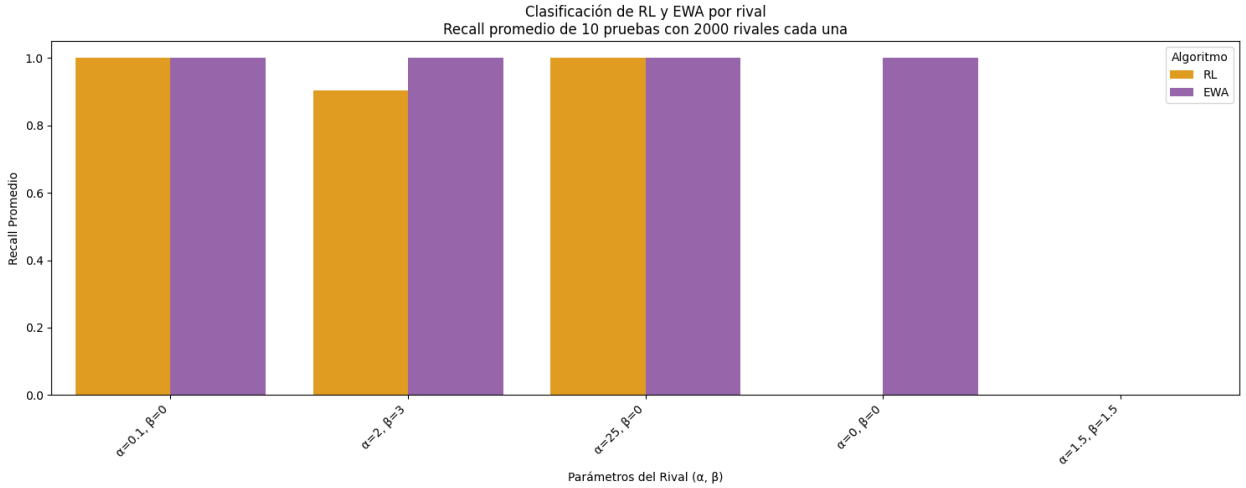

Figura 126: Recall al clasificar 5 rivales cuando ambos agentes aprendieron 3 rivales y se agrega dos rivales nuevos, uno con un rango de ofertas aceptables muy distinguible y otro poco distinguible.

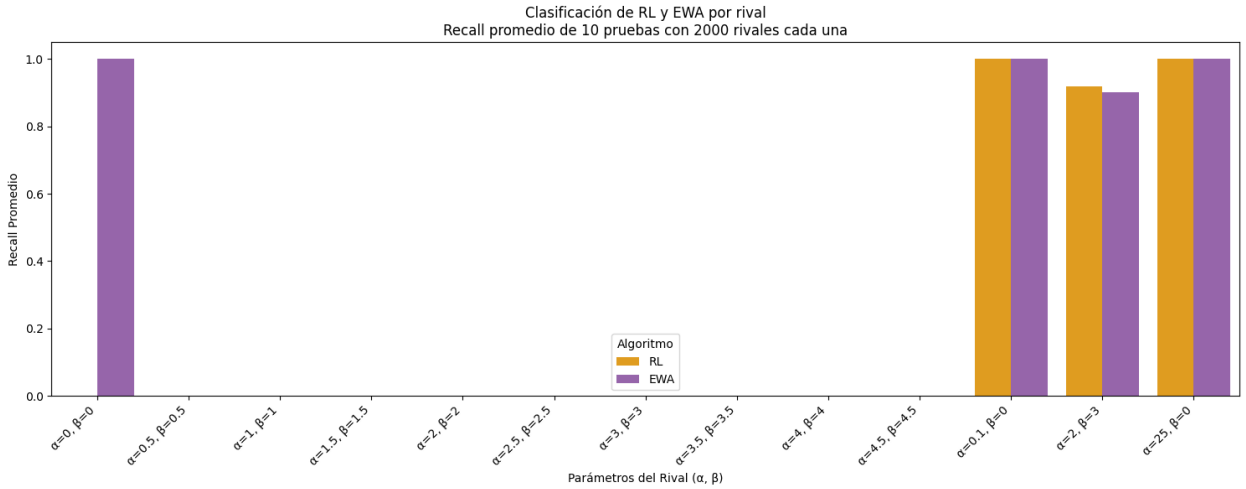

Figura 127: Recall al clasificar 13 rivales cuando ambos agentes aprendieron 3 rivales.

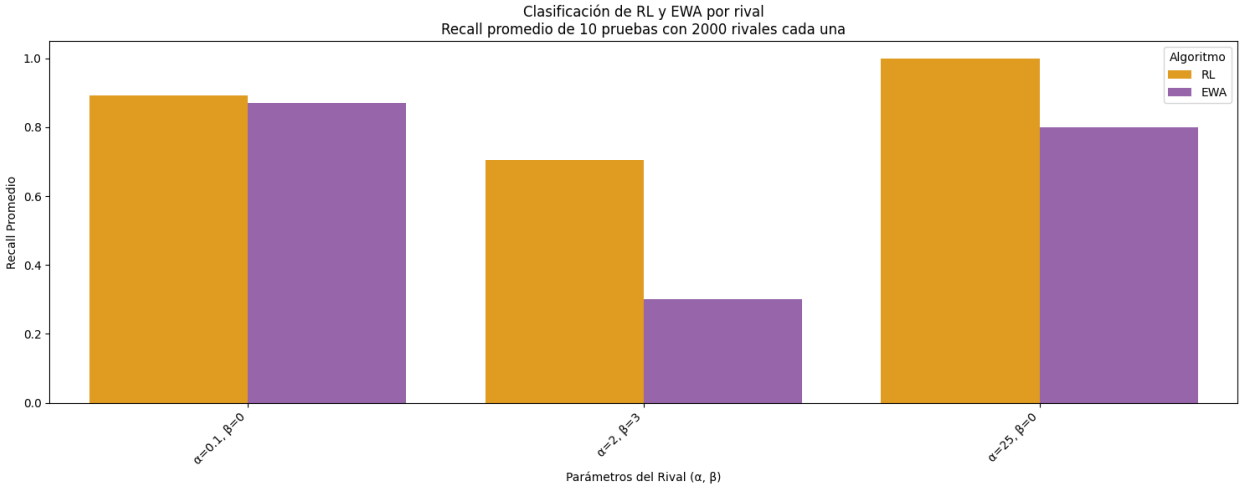

Figura 128: Recall al clasificar 3 rivales cuando ambos agentes aprendieron 13 rivales.

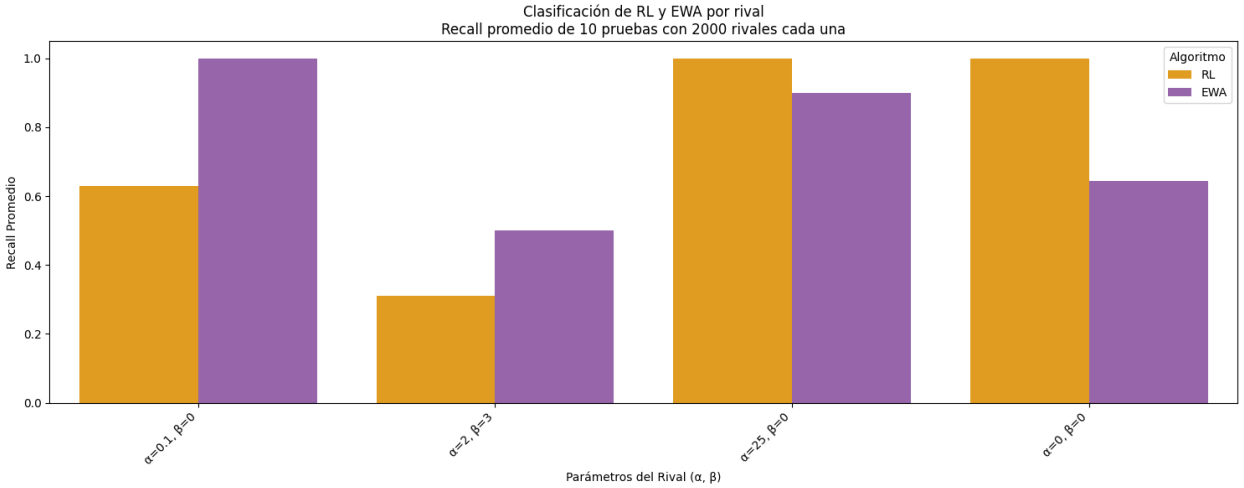

Figura 129: Recall al clasificar 4 rivales cuando ambos agentes aprendieron 13 rivales y se agrega un rival nuevo con un rango de ofertas aceptable muy distinguible.

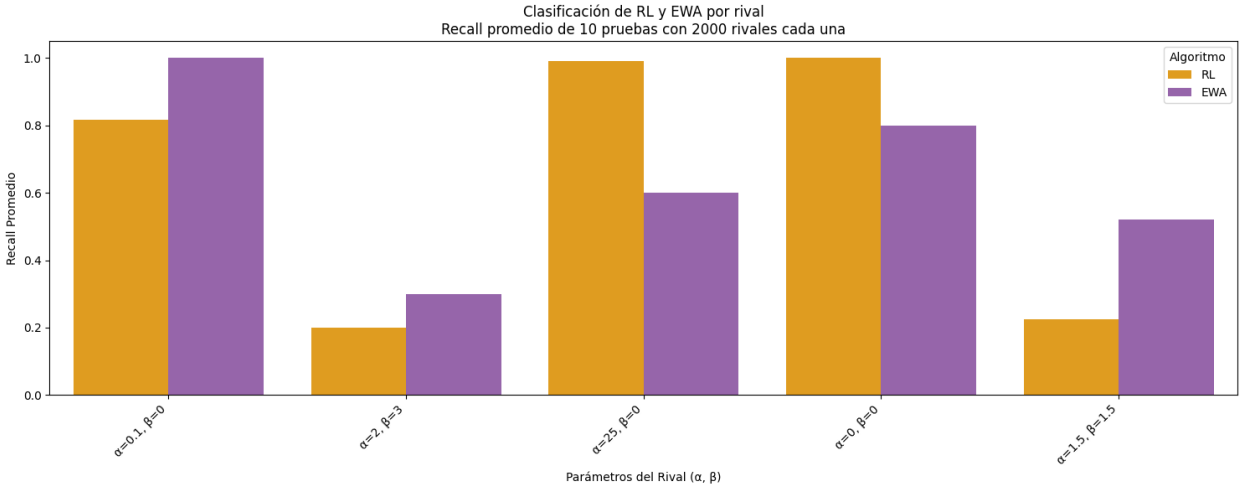

Figura 130: Recall al clasificar 5 rivales cuando ambos agentes aprendieron 13 rivales y se agrega dos rivales nuevos, uno con un rango de ofertas aceptables muy distinguible y otro poco distinguible.

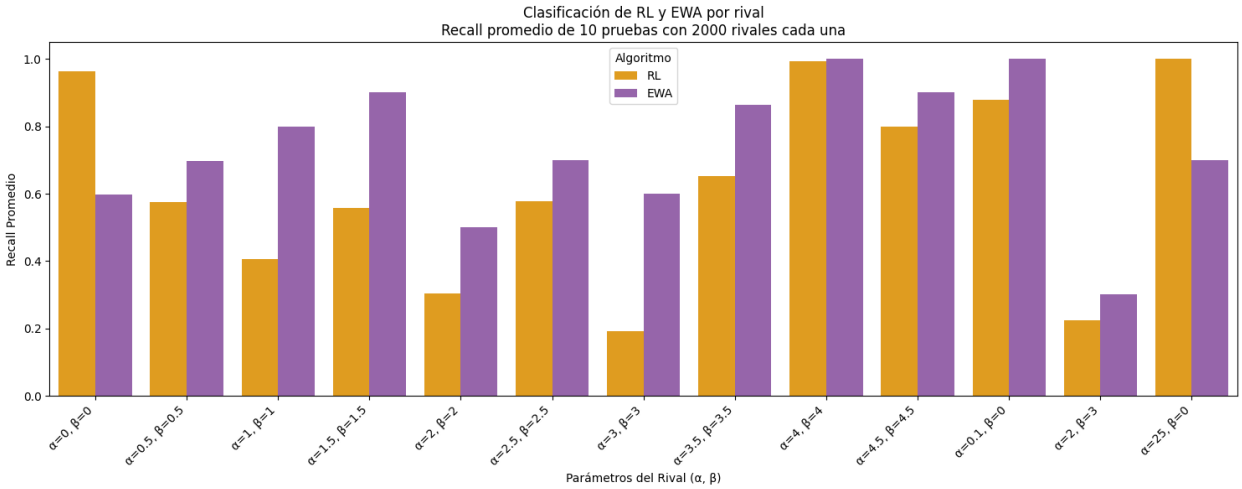

Figura 131: Recall al clasificar 13 rivales cuando ambos agentes aprendieron 13 rivales.

9.5.2. F1-Score

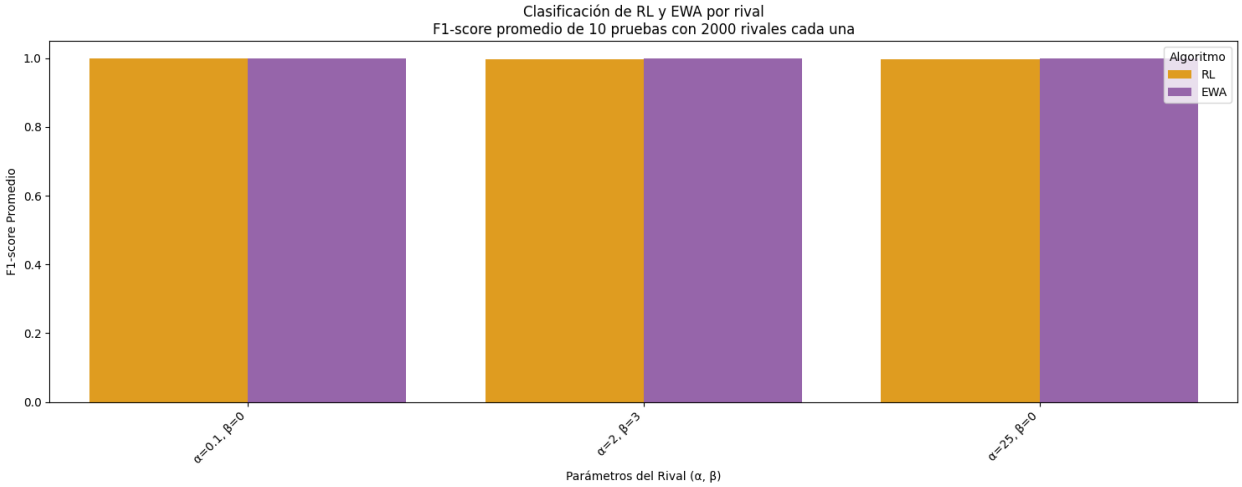

Figura 132: F1-Score al clasificar 3 rivales cuando ambos agentes aprendieron 3 rivales.

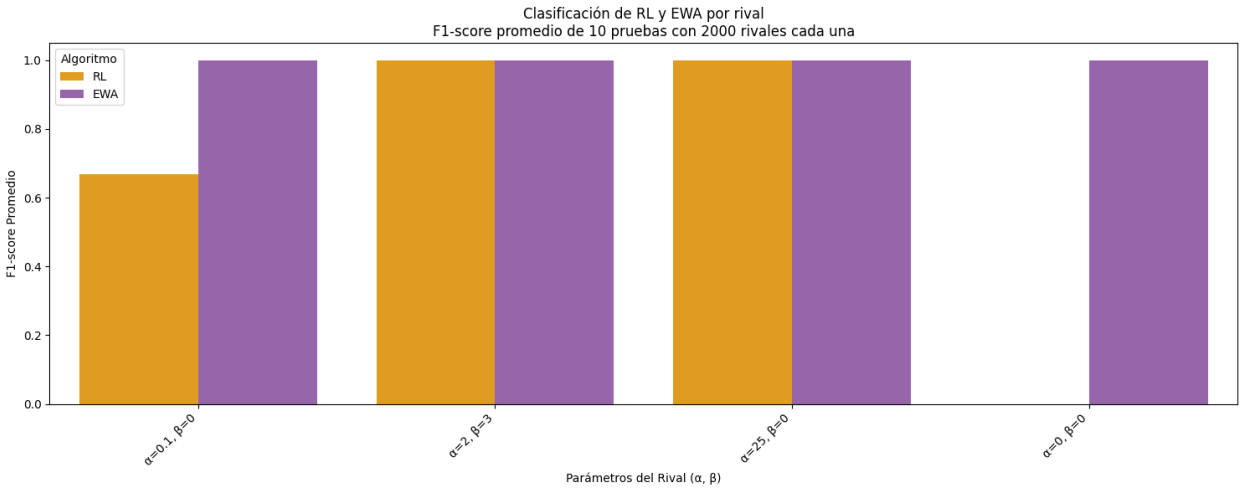

Figura 133: F1-Score al clasificar 4 rivales cuando ambos agentes aprendieron 3 rivales y se agrega un rival nuevo con un rango de ofertas aceptable muy distinguible.

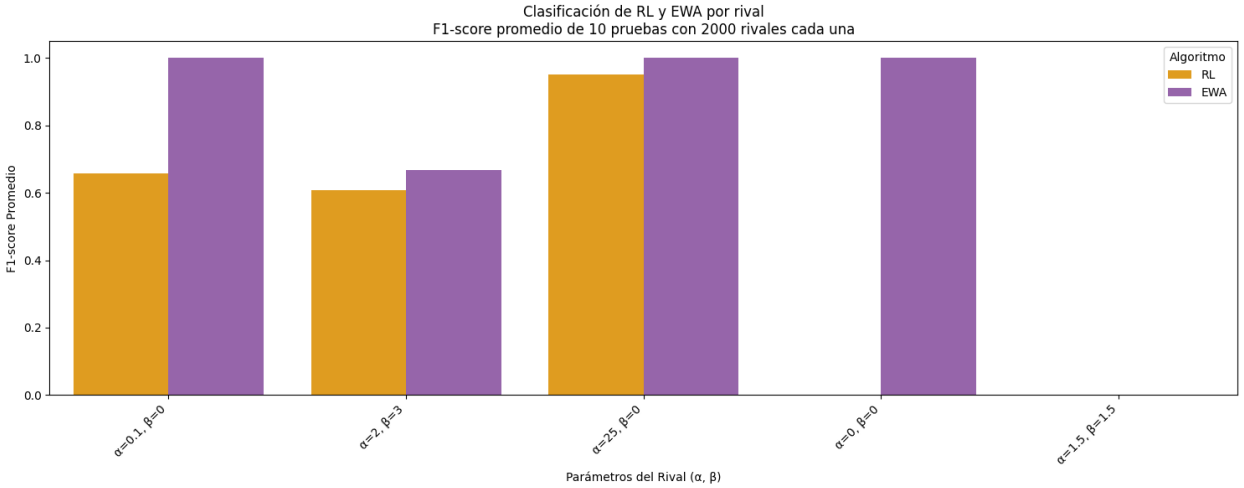

Figura 134: F1-Score al clasificar 5 rivales cuando ambos agentes aprendieron 3 rivales y se agrega dos rivales nuevos, uno con un rango de ofertas aceptables muy distinguible y otro poco distinguible.

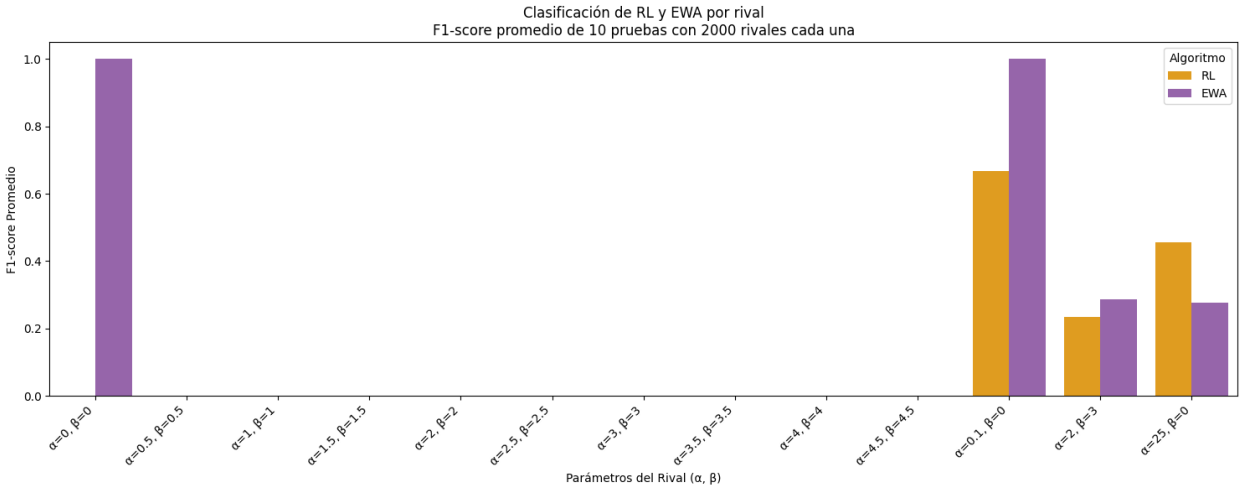

Figura 135: F1-Score al clasificar 13 rivales cuando ambos agentes aprendieron 3 rivales.

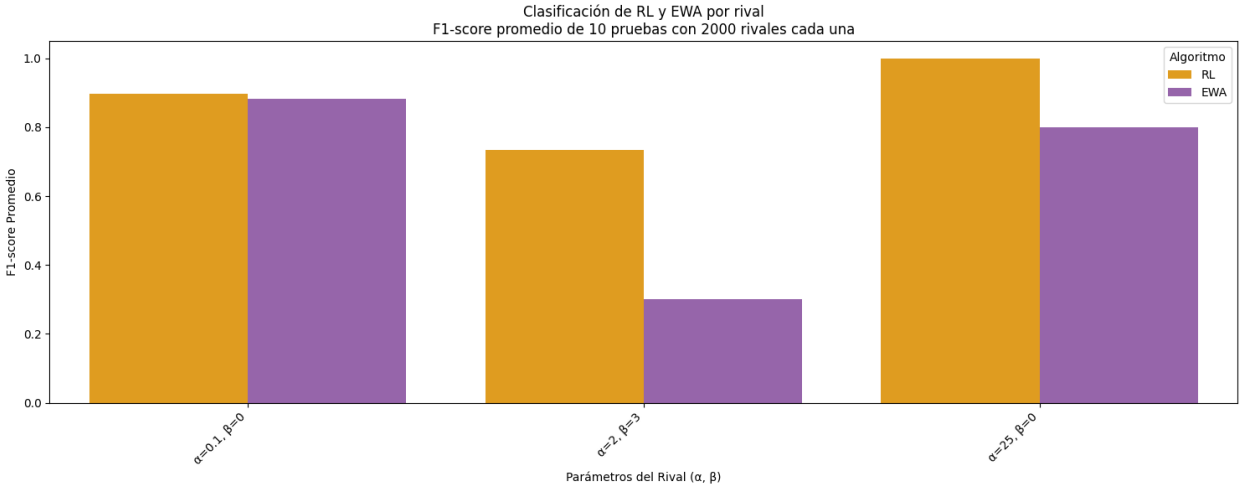

Figura 136: F1-Score al clasificar 3 rivales cuando ambos agentes aprendieron 13 rivales.

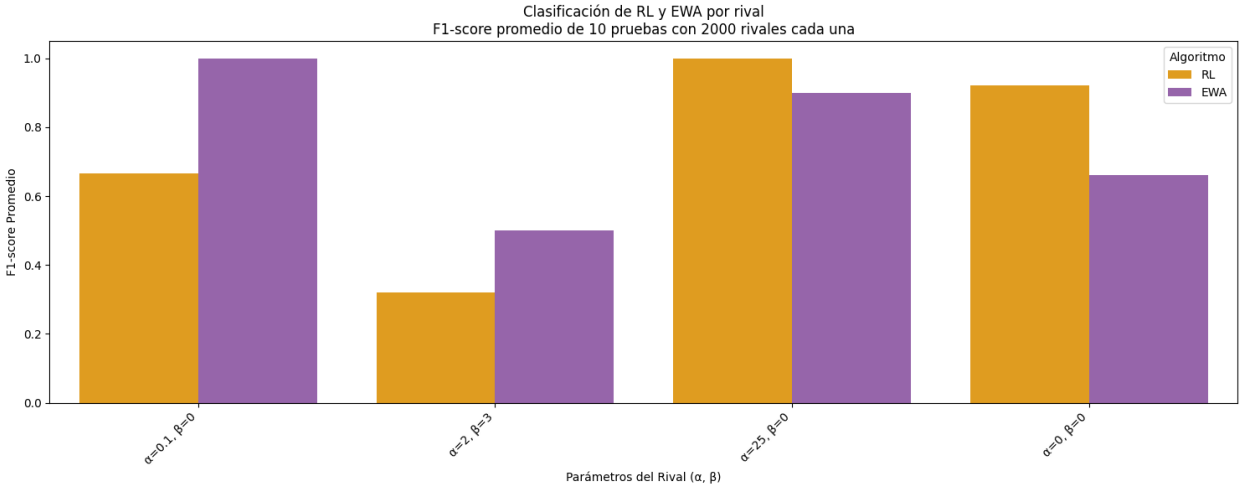

Figura 137: F1-Score al clasificar 4 rivales cuando ambos agentes aprendieron 13 rivales y se agrega un rival nuevo con un rango de ofertas aceptable muy distinguible.

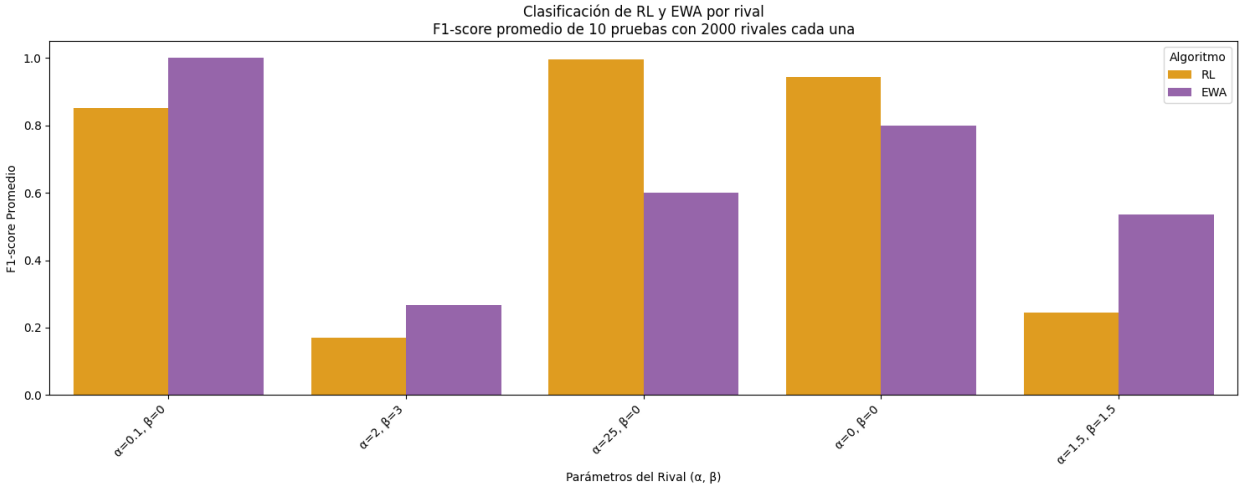

Figura 138: F1-Score al clasificar 5 rivales cuando ambos agentes aprendieron 13 rivales y se agrega dos rivales nuevos, uno con un rango de ofertas aceptables muy distinguible y otro poco distinguible.

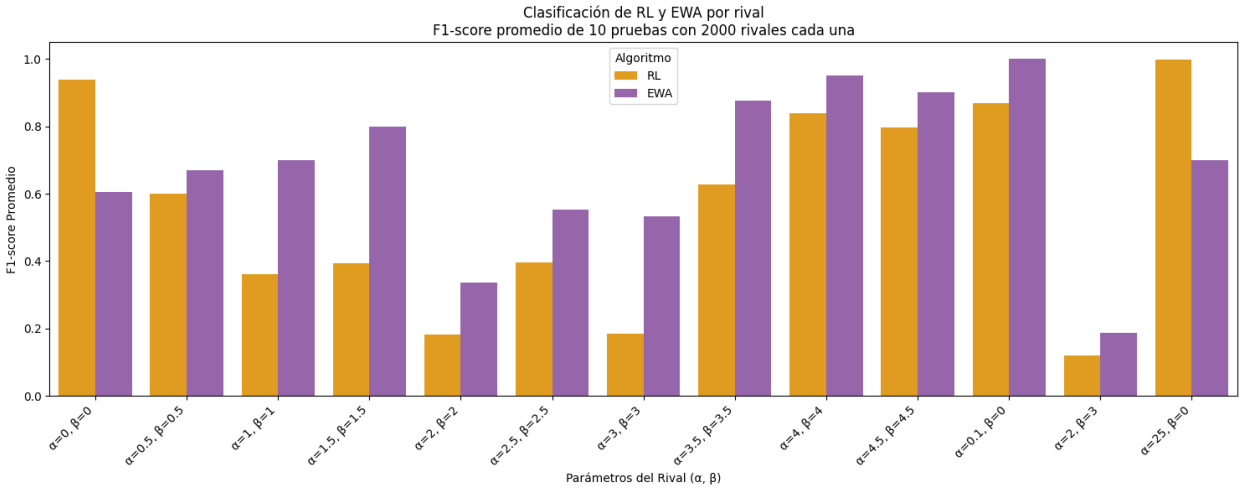

Figura 139: F1-Score al clasificar 13 rivales cuando ambos agentes aprendieron 13 rivales.

9.6. Gráficos de desempeño de RL y FEWA en el sistema AB por distribución y por tamaño

9.6.1. Distribuciones para tamaño de muestra de 300 agentes

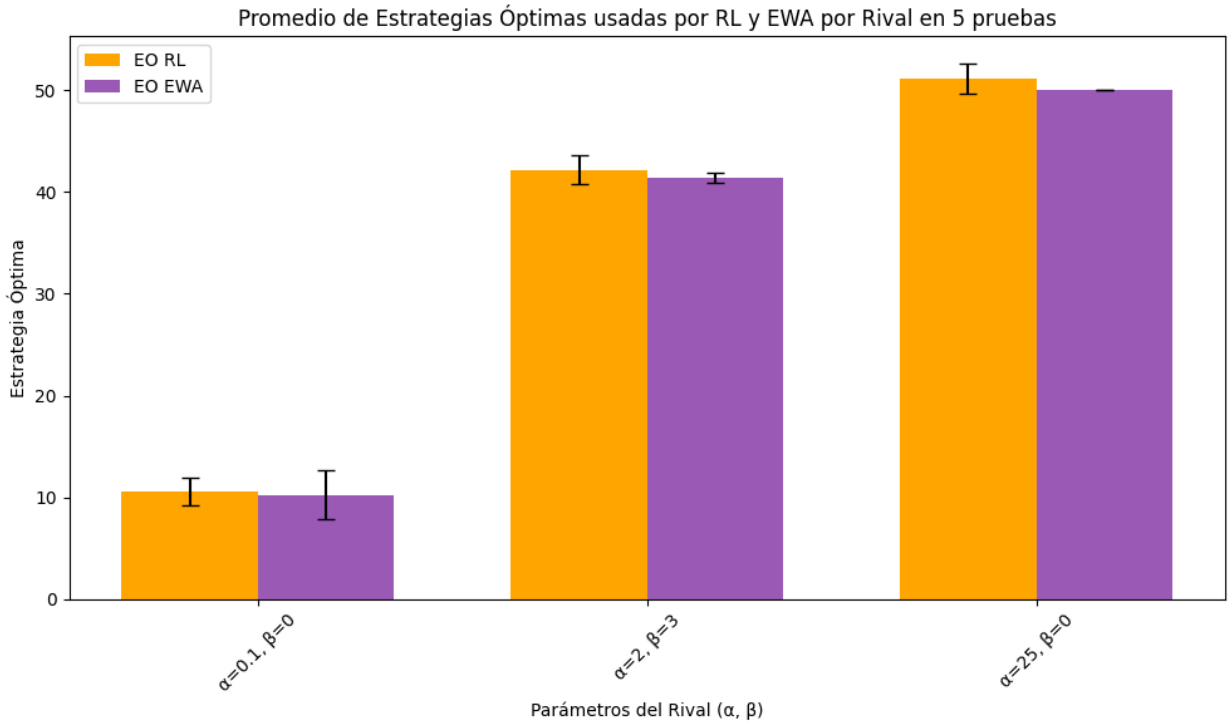

Figura 140: Proporción de estrategias usadas por los agentes RL y FEWA cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA donde todos tienen una proporción uniforme.

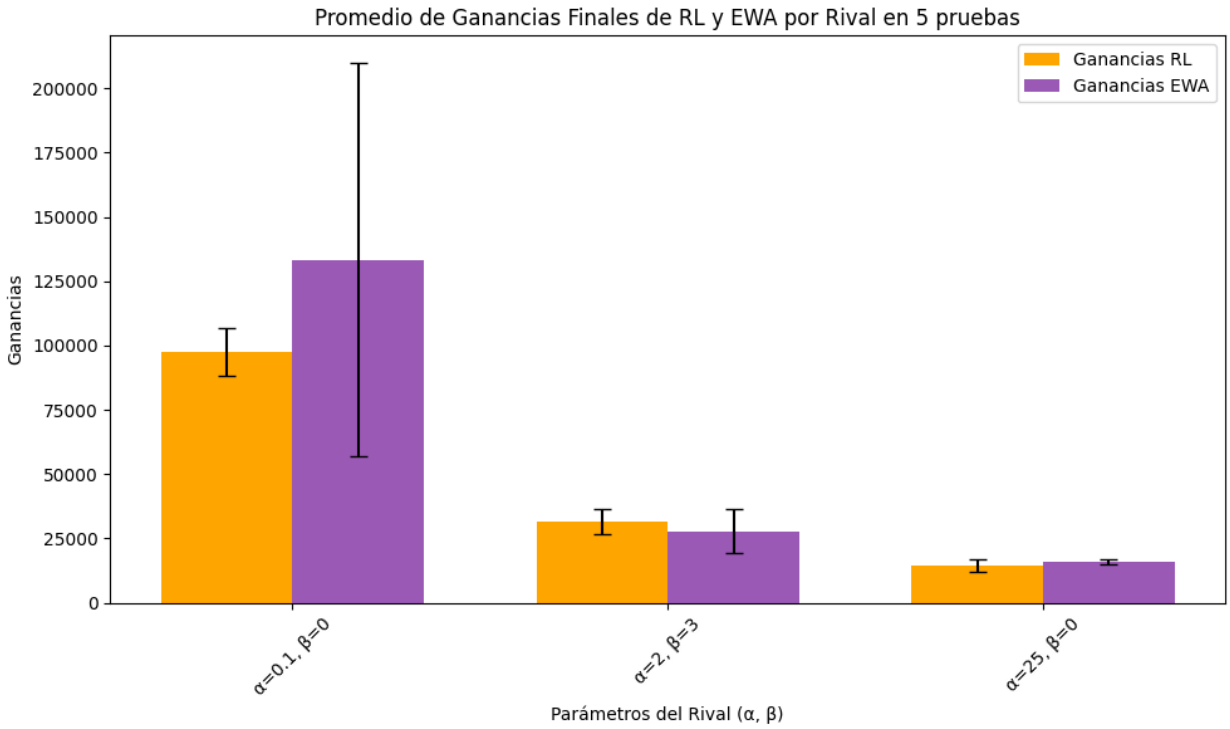

Figura 141: Ganancias que obtivieron RL y FEWA por cada grupo cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA donde todos tienen una proporción uniforme.

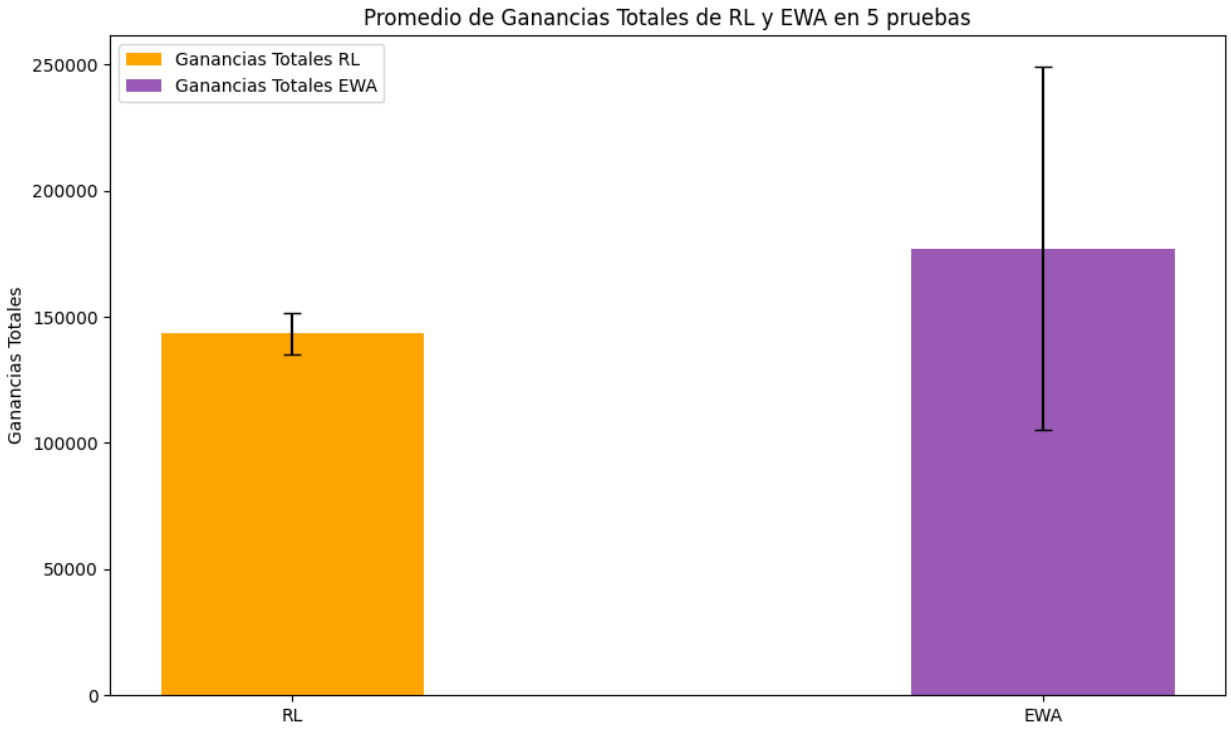

Figura 142: Ganancias totales que obtivieron RL y FEWA cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA donde todos tienen una proporción uniforme.

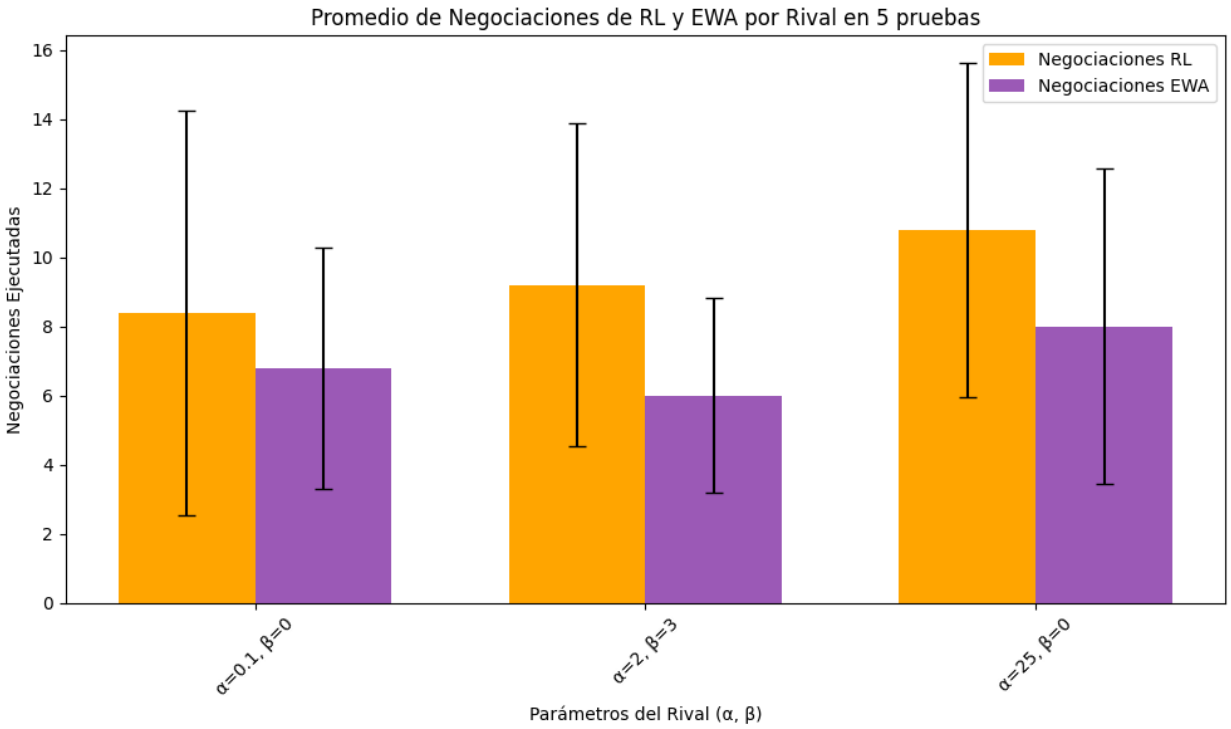

Figura 143: Negociaciones que llevaron a cabo RL y FEWA por cada grupo cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA donde todos tienen una proporción uniforme.

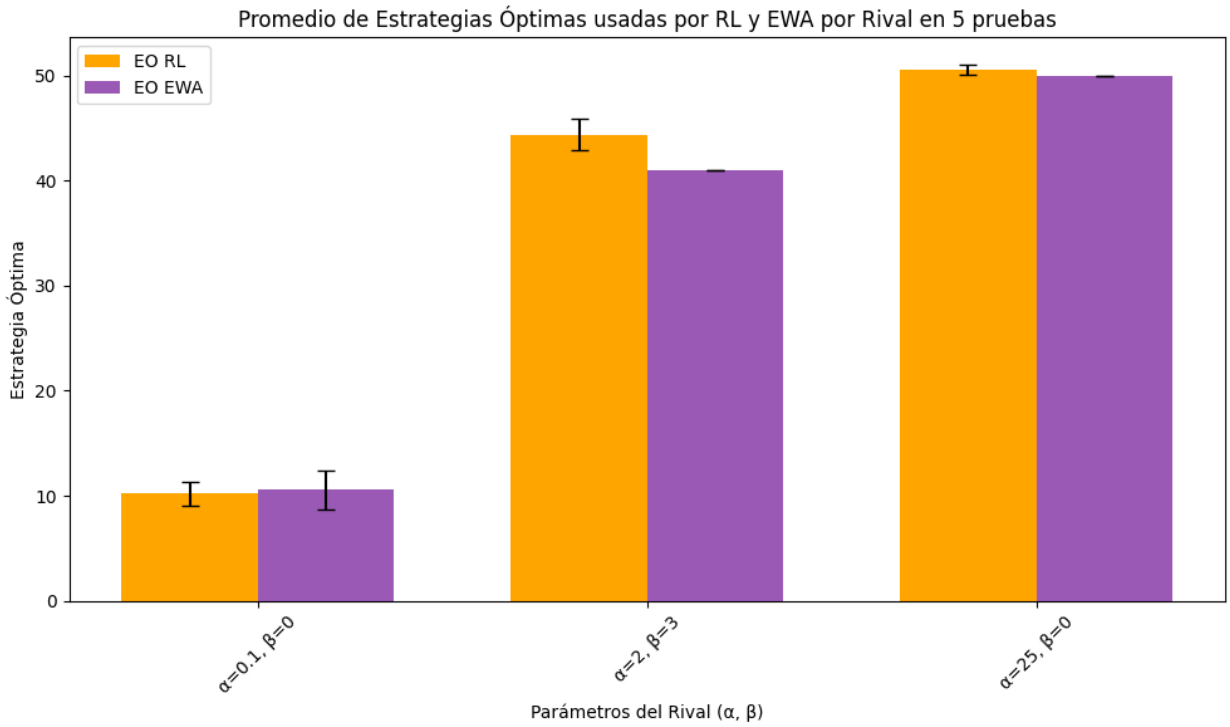

Figura 144: Estrategias usadas por los agentes RL y FEWA cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia alta a resultados ventajosos tienen mayoría.

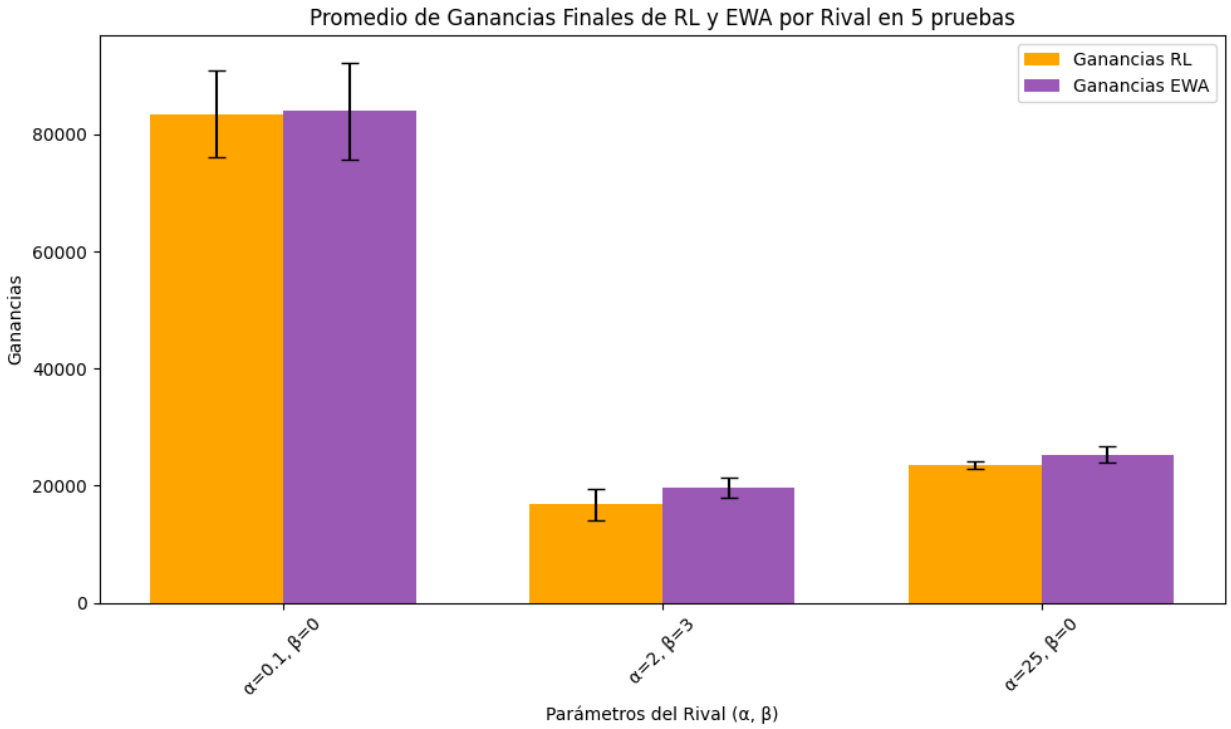

Figura 145: Ganancias que obtivieron RL y FEWA por cada grupo cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia alta a resultados ventajosos tienen mayoría.

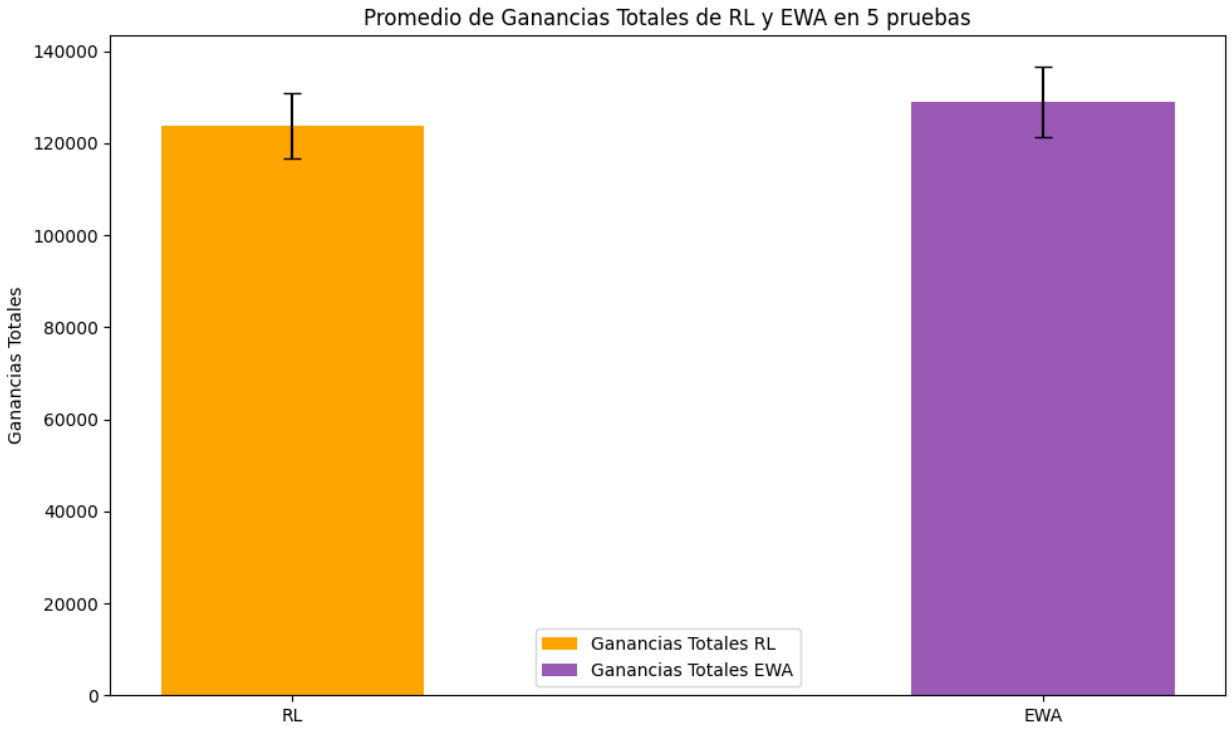

Figura 146: Ganancias totales que obtivieron RL y FEWA cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia alta a resultados ventajosos tienen mayoría.

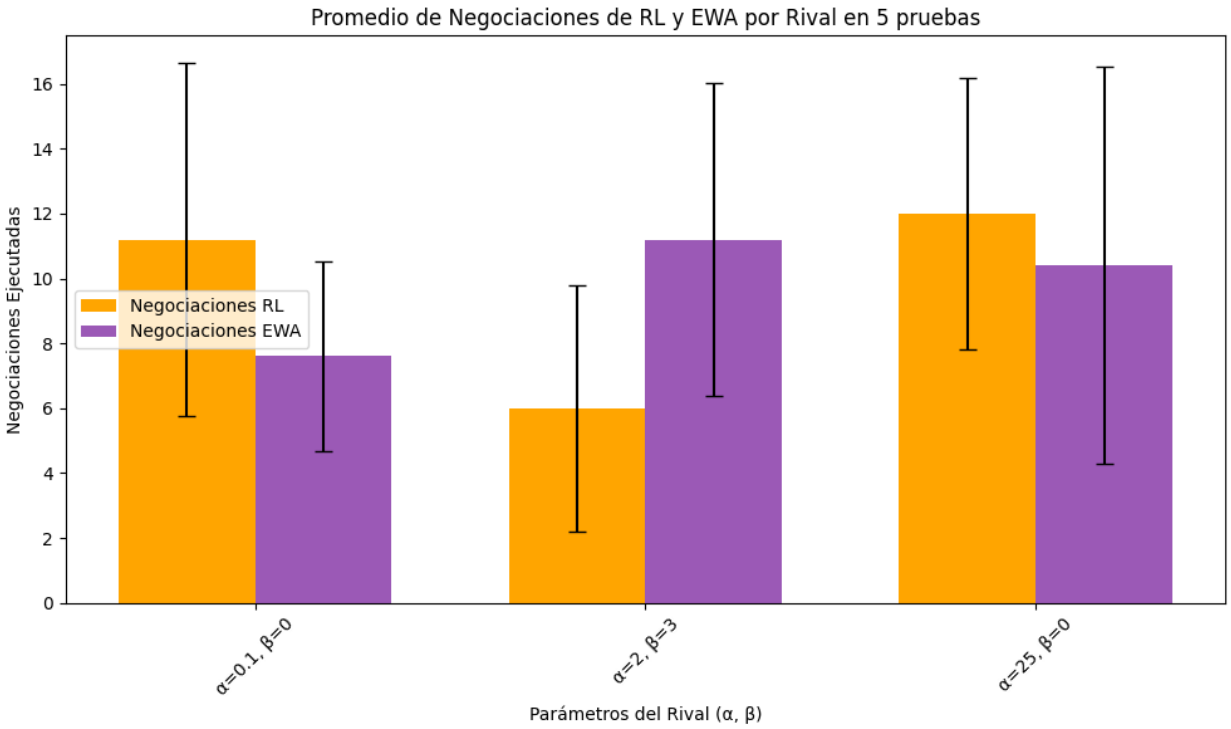

Figura 147: Negociaciones que llevaron a cabo RL y FEWA por cada grupo cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia alta a resultados ventajosos tienen mayoría.

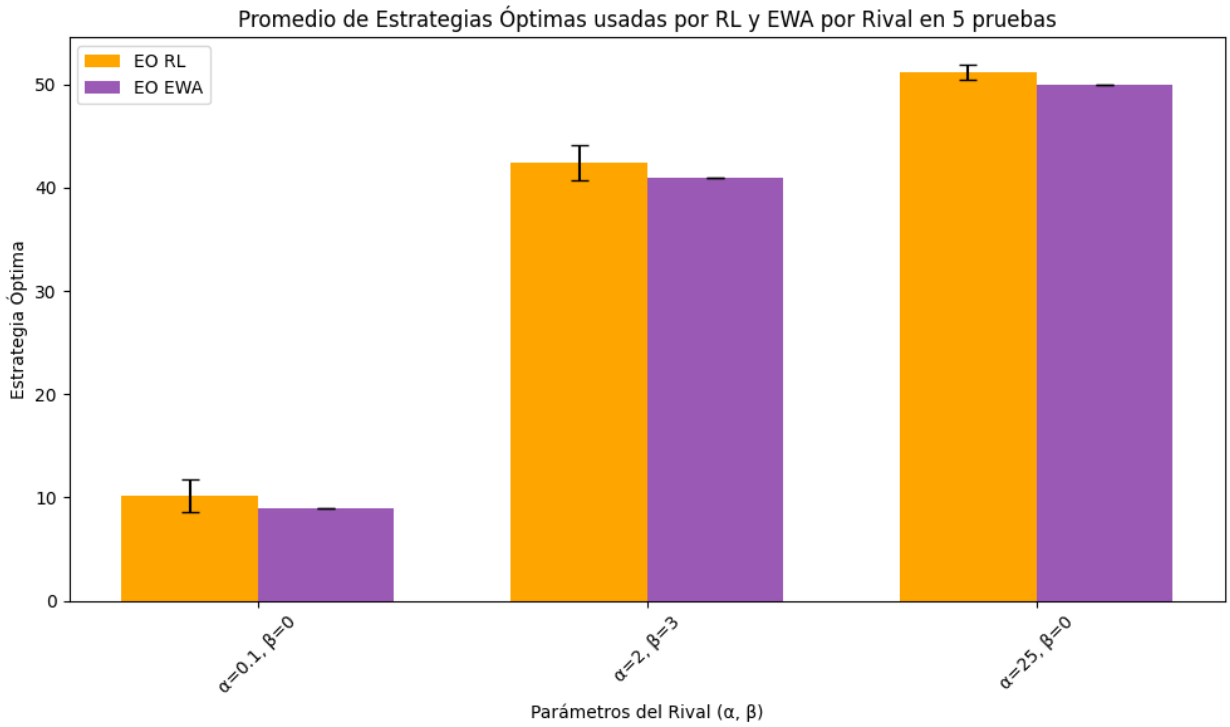

Figura 148: Estrategias usadas por los agentes RL y FEWA cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia baja a resultados ventajosos tienen mayoría.

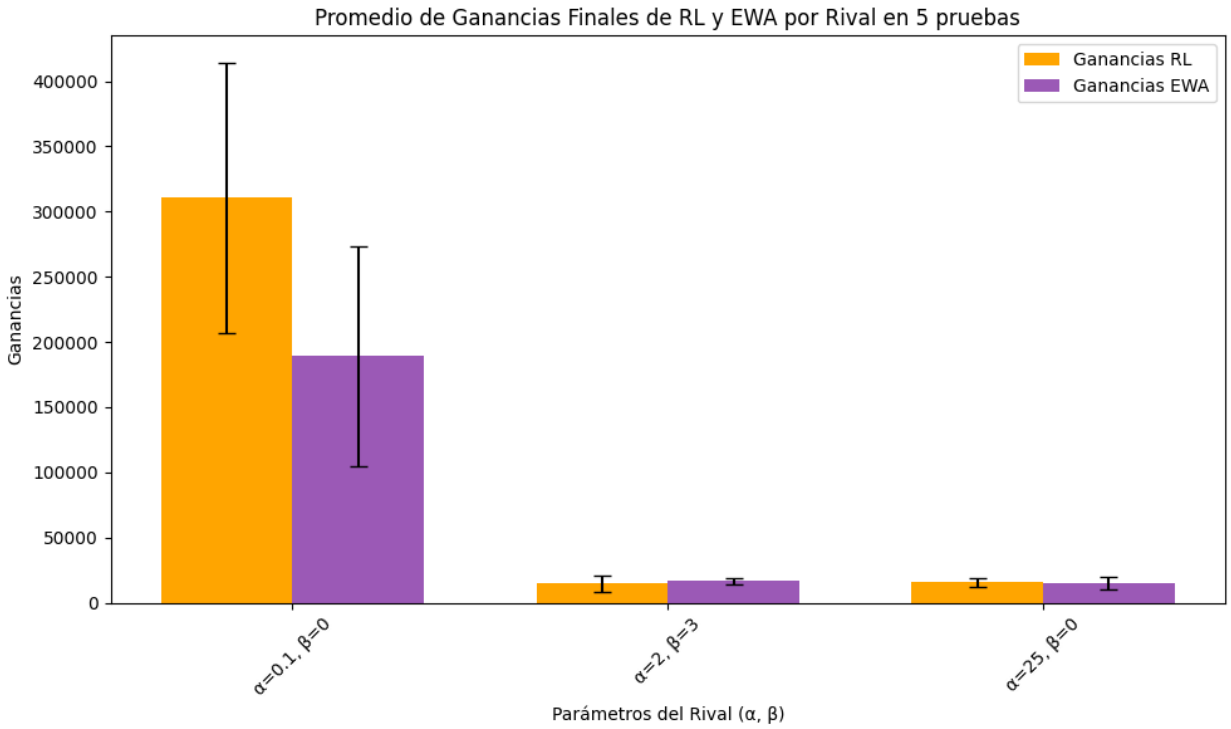

Figura 149: Ganancias que obtivieron RL y FEWA por cada grupo cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia baja a resultados ventajosos tienen mayoría.

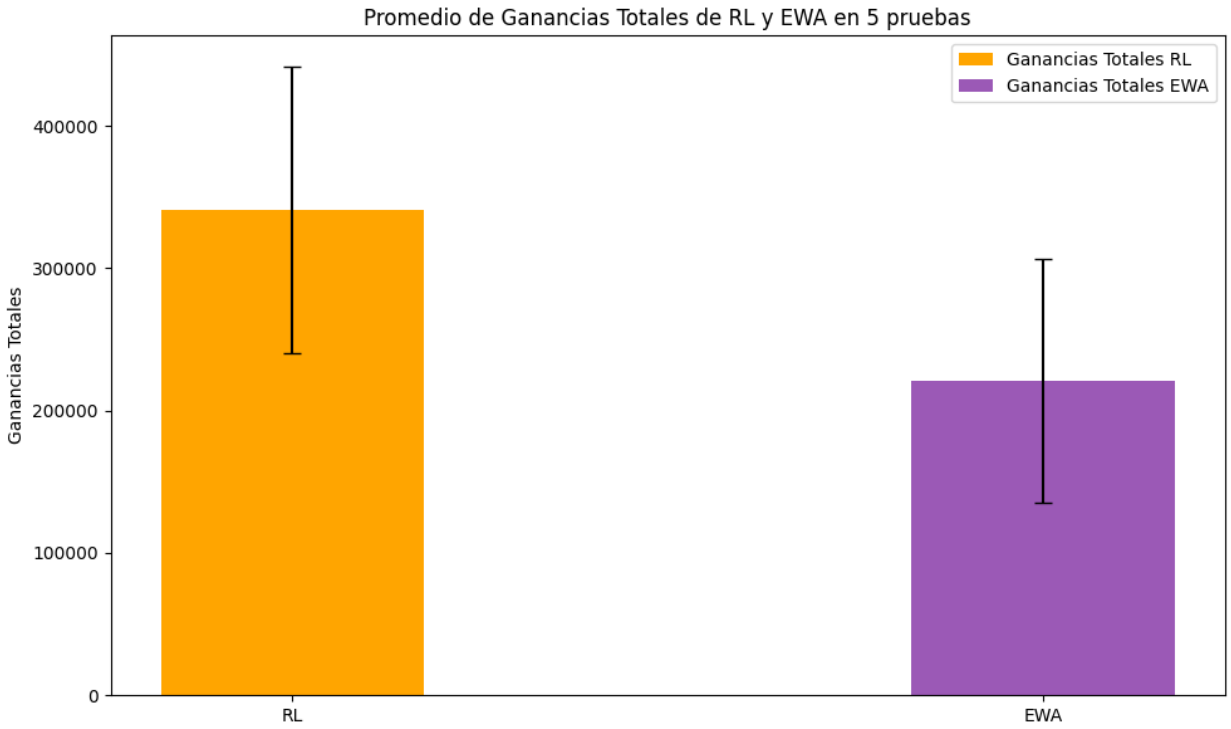

Figura 150: Ganancias totales que obtivieron RL y FEWA por cada grupo cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia baja a resultados ventajosos tienen mayoría.

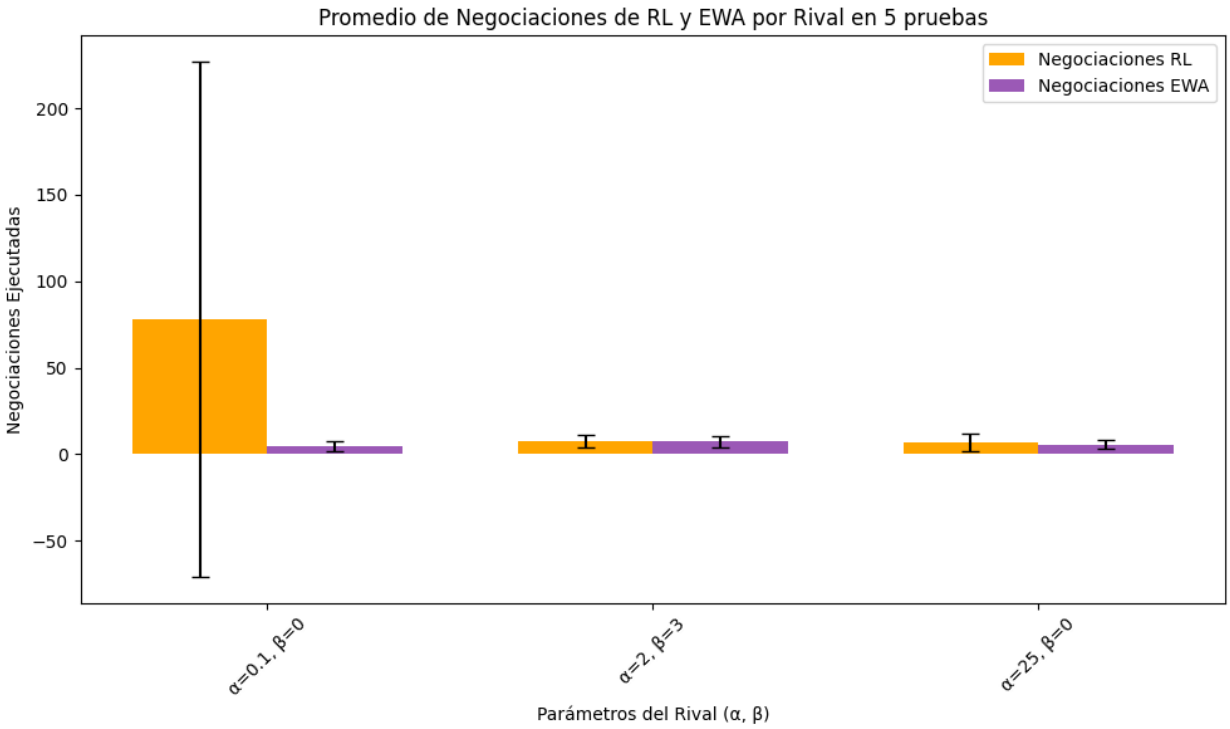

Figura 151: Negociaciones que llevaron a cabo RL y FEWA por cada grupo cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia baja a resultados ventajosos tienen mayoría.

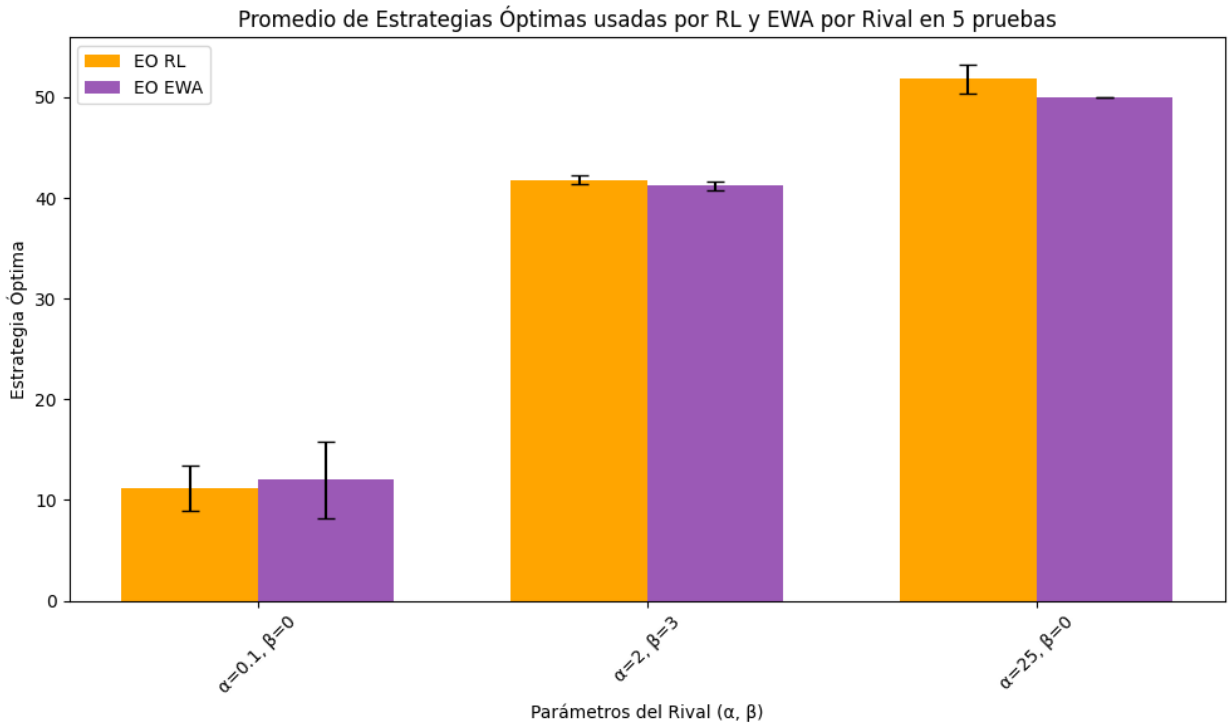

Figura 152: Estrategias usadas por los agentes RL y FEWA cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia media a resultados justos tienen mayoría.

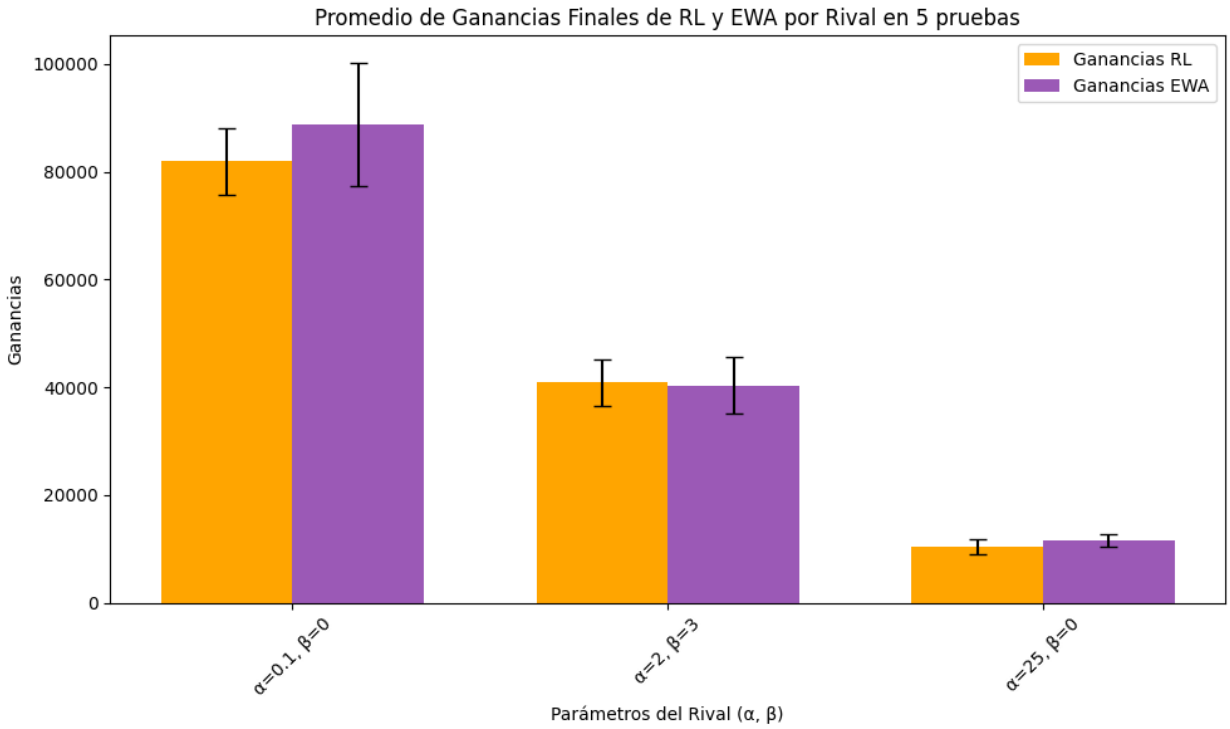

Figura 153: Ganancias que obtivieron RL y FEWA por cada grupo cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia media a resultados justos tienen mayoría.

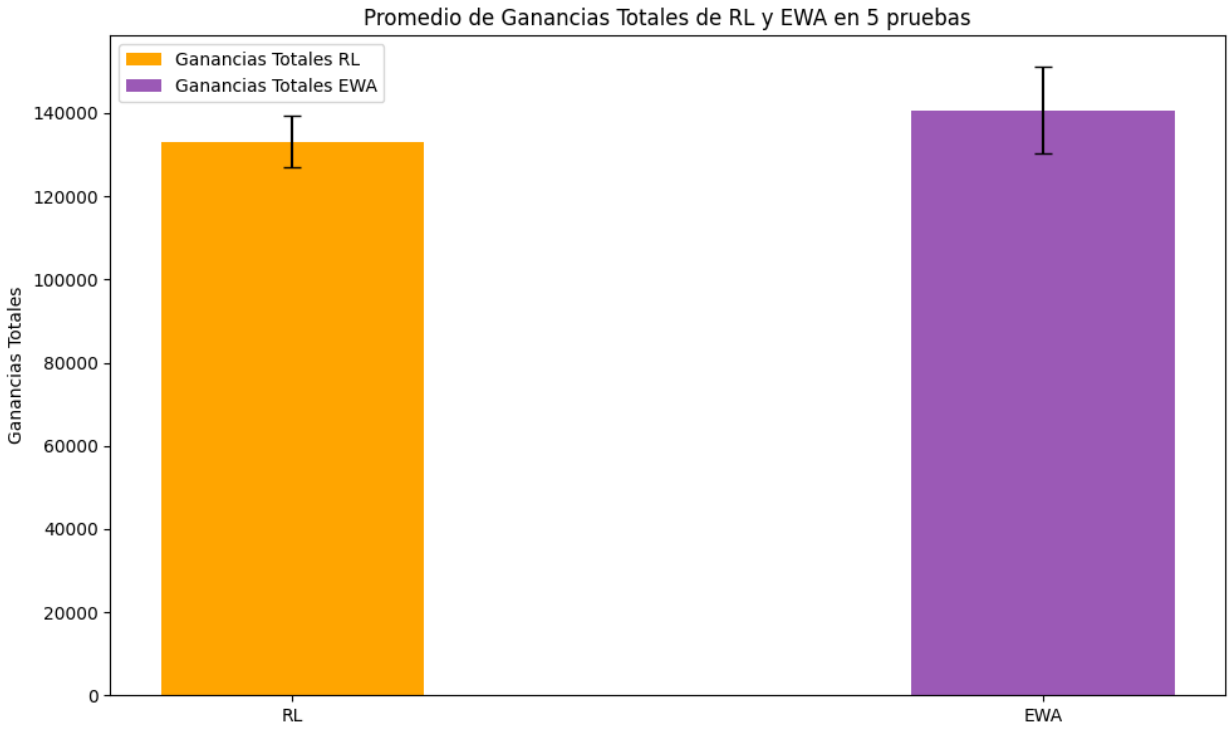

Figura 154: Ganancias totales que obtivieron RL y FEWA cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia media a resultados justos tienen mayoría.

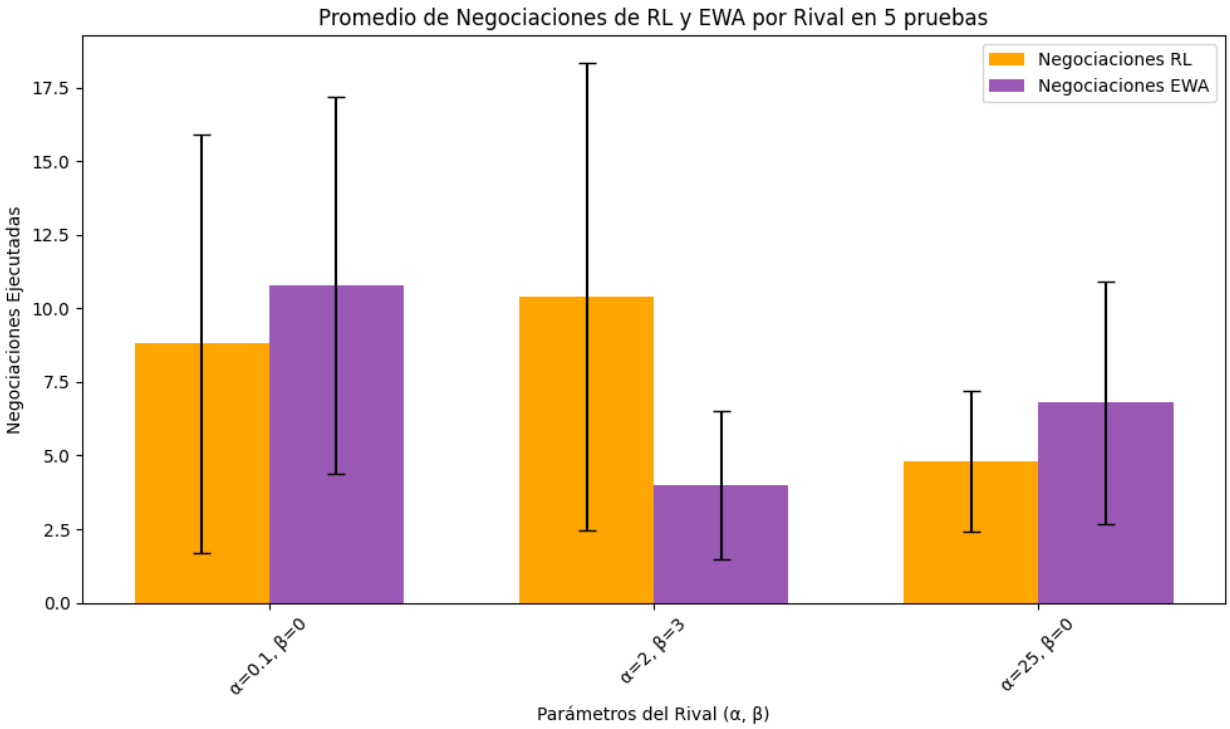

Figura 155: Negociaciones que llevaron a cabo RL y FEWA por cada grupo cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia media a resultados justos tienen mayoría.

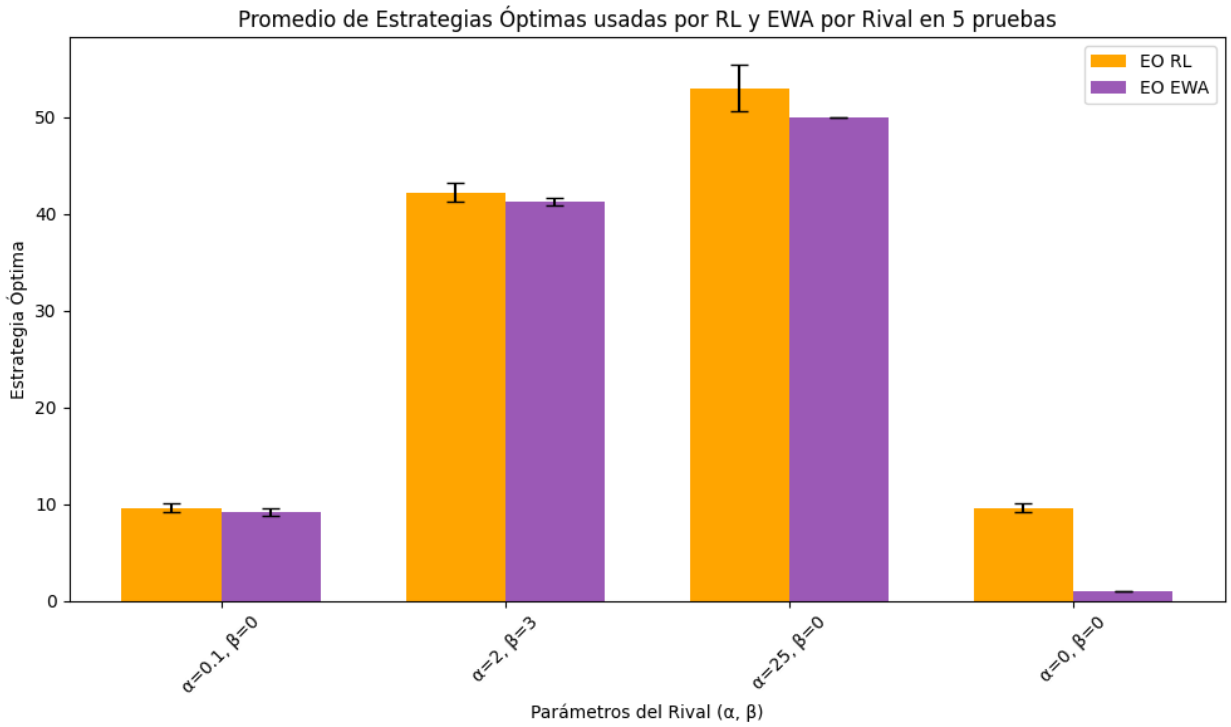

Figura 156: Estrategias usadas por los agentes RL y FEWA cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde todos tienen una proporción uniforme.

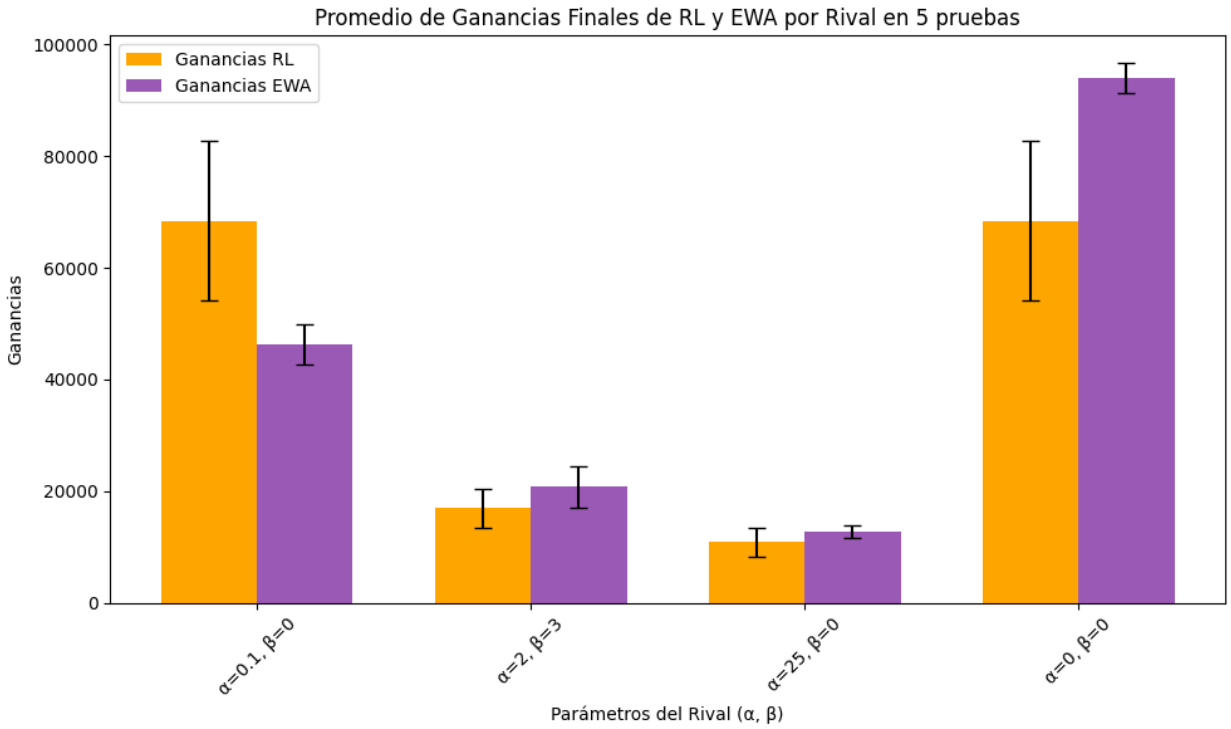

Figura 157: Ganancias que obtivieron RL y FEWA por cada grupo cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde todos tienen una proporción uniforme.

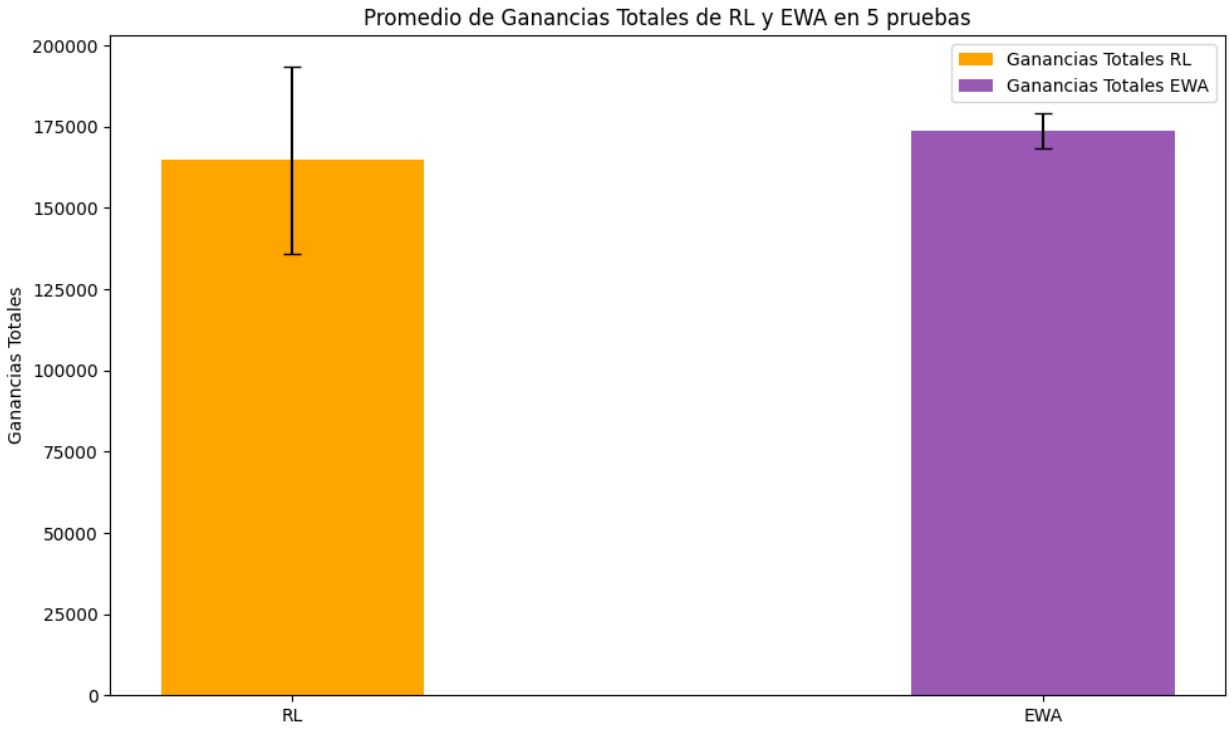

Figura 158: Ganancias totales que obtivieron RL y FEWA cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde todos tienen una proporción uniforme.

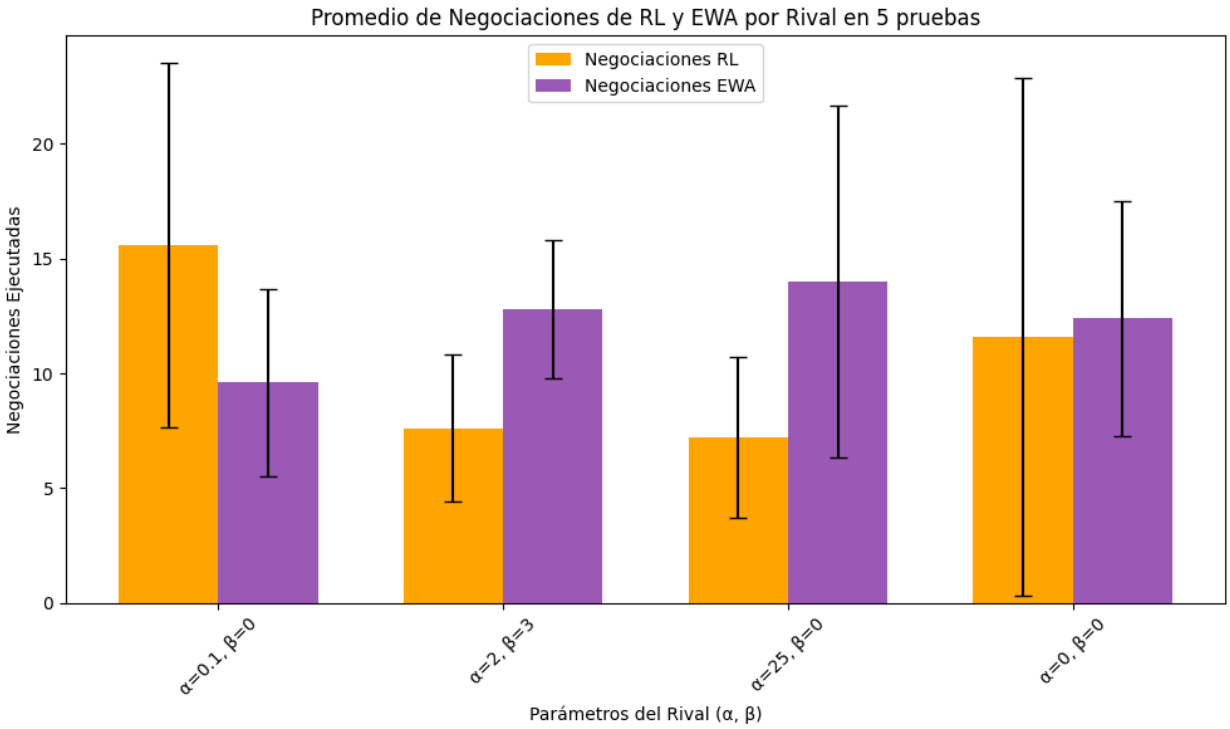

Figura 159: Negociaciones que llevaron a cabo RL y FEWA por cada grupo cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde todos tienen una proporción uniforme.

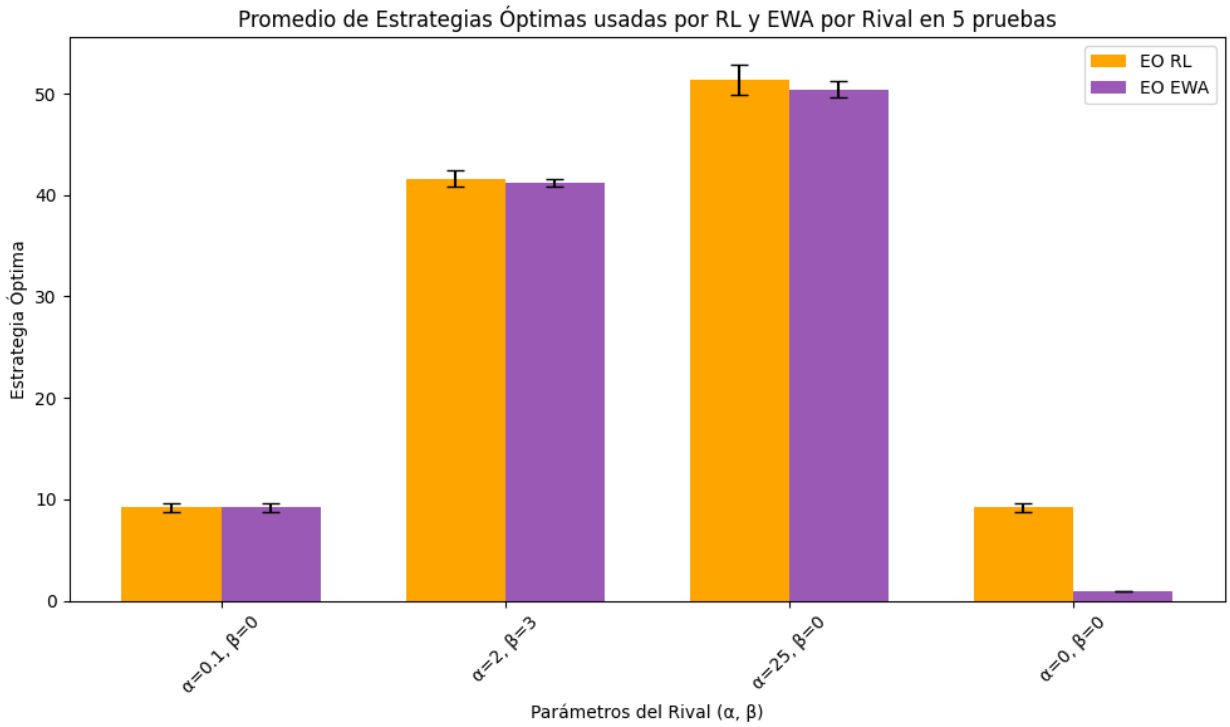

Figura 160: Estrategias usadas por los agentes RL y FEWA cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene minoría.

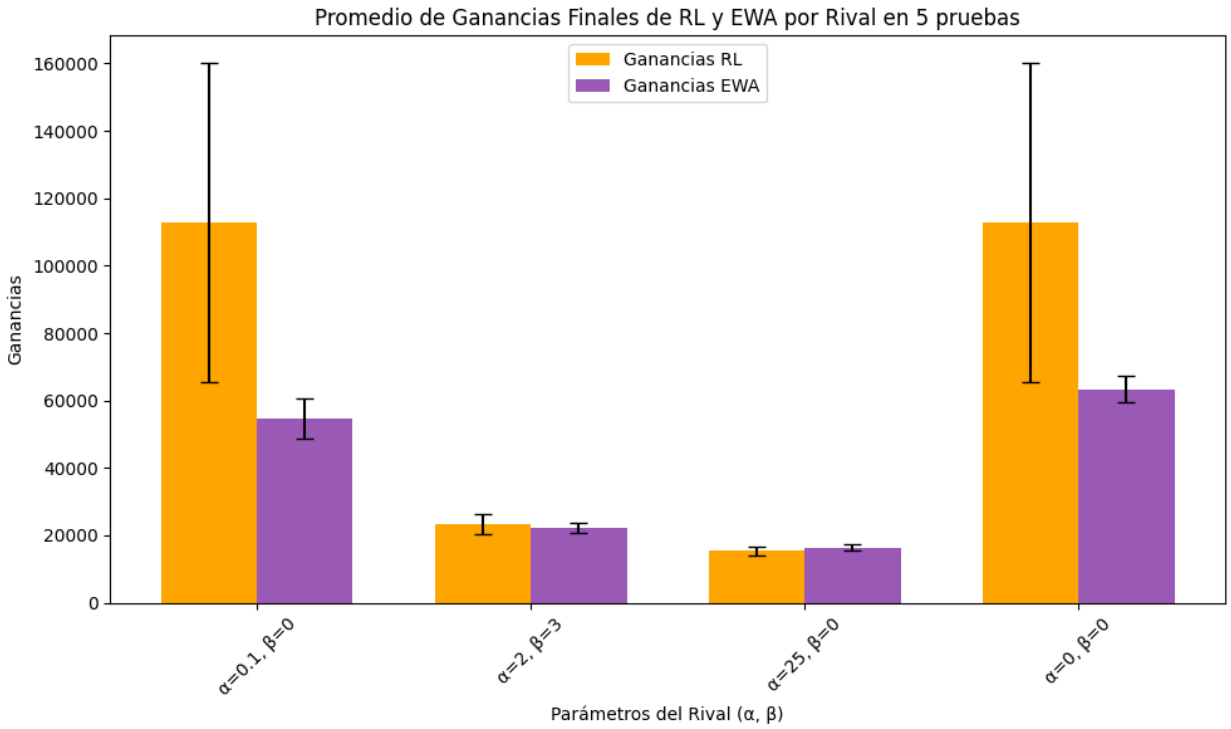

Figura 161: Ganancias que obtivieron RL y FEWA por cada grupo cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene minoría.

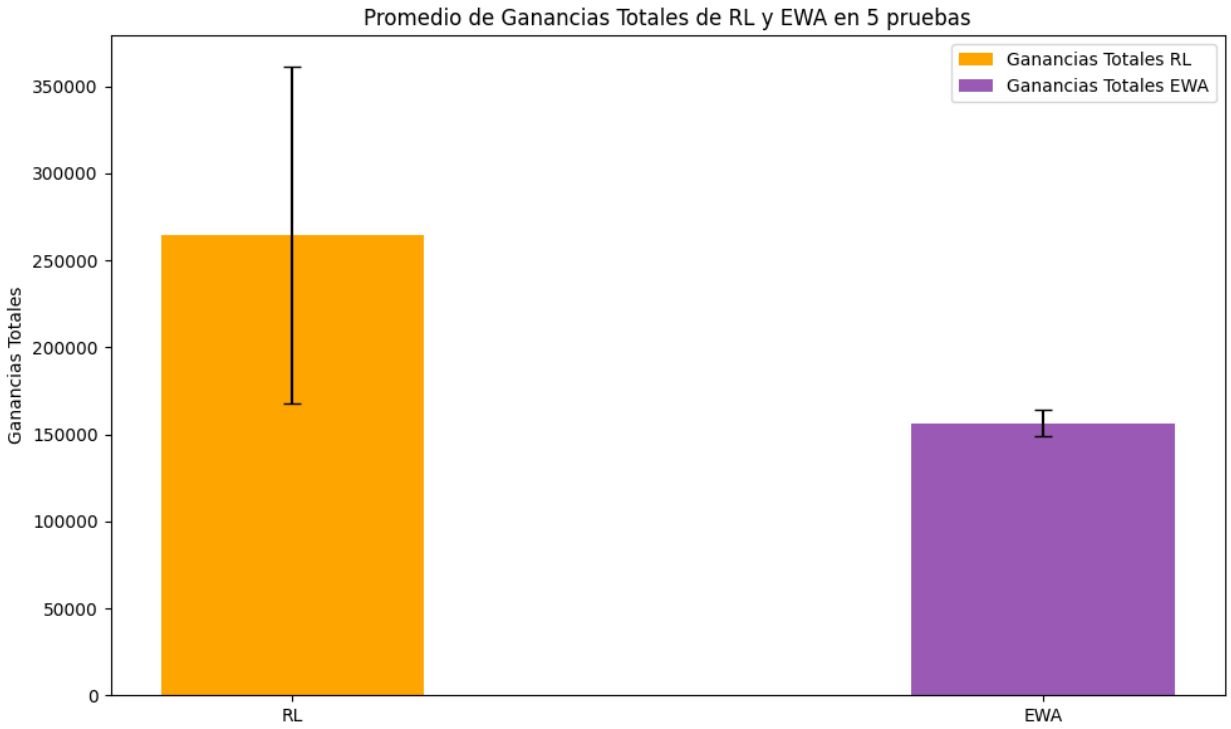

Figura 162: Ganancias totales que obtivieron RL y FEWA cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene minoría.

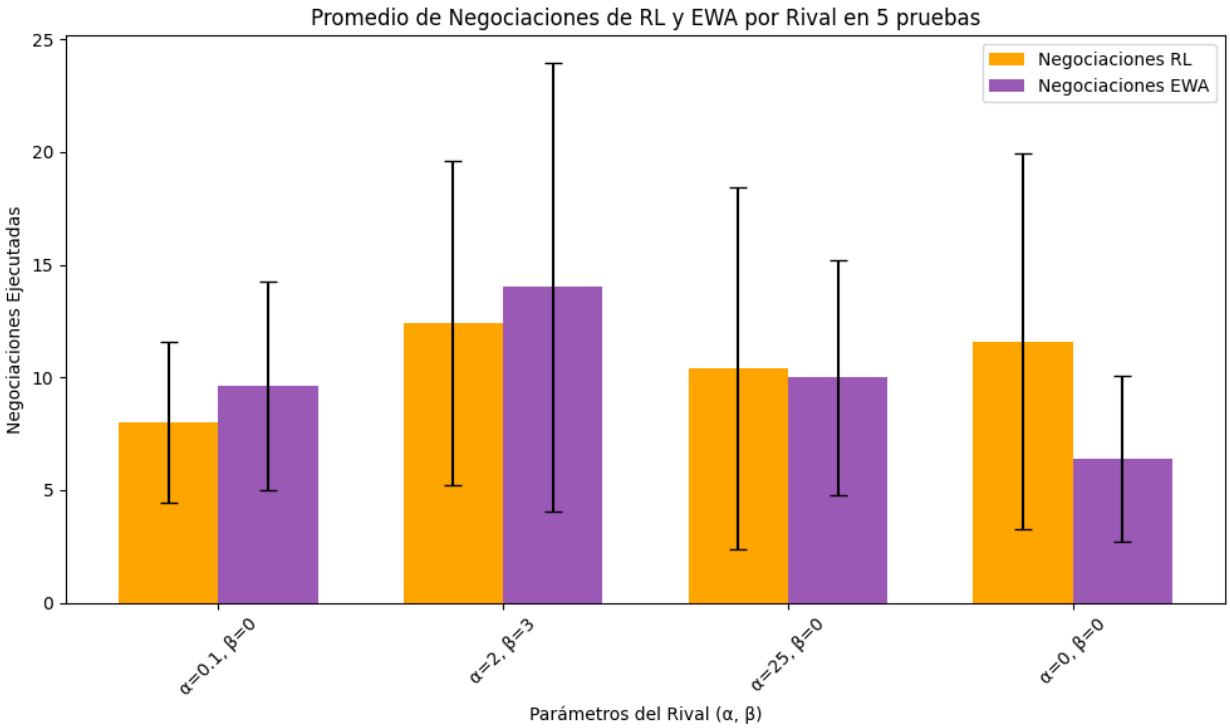

Figura 163: Negociaciones que llevaron a cabo RL y FEWA por cada grupo cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene minoría.

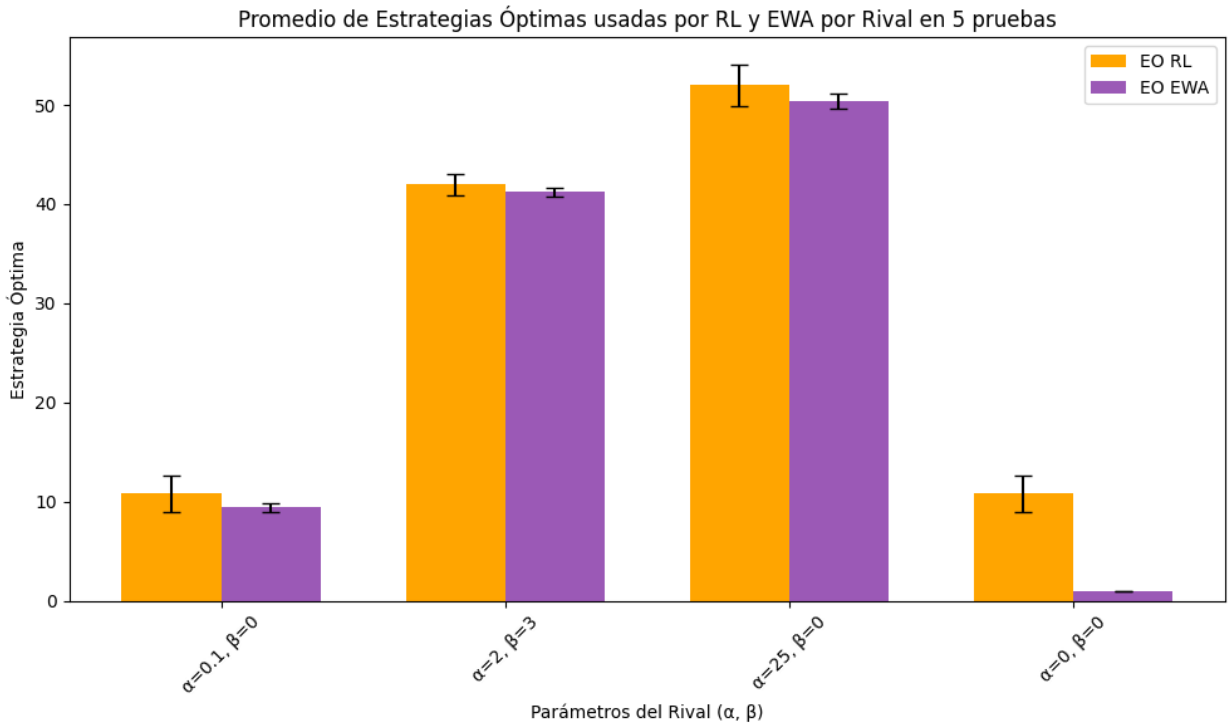

Figura 164: Estrategias usadas por los agentes RL y FEWA cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene mayoría.

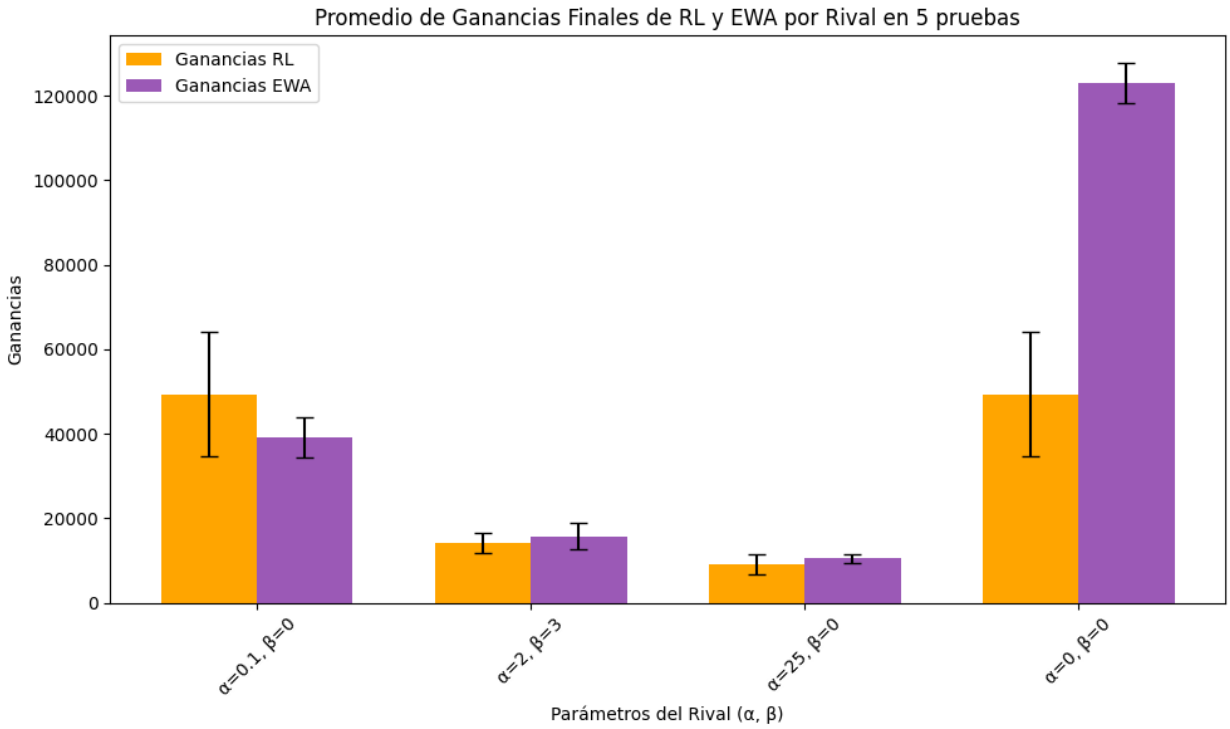

Figura 165: Ganancias que obtivieron RL y FEWA por cada grupo cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene mayoría.

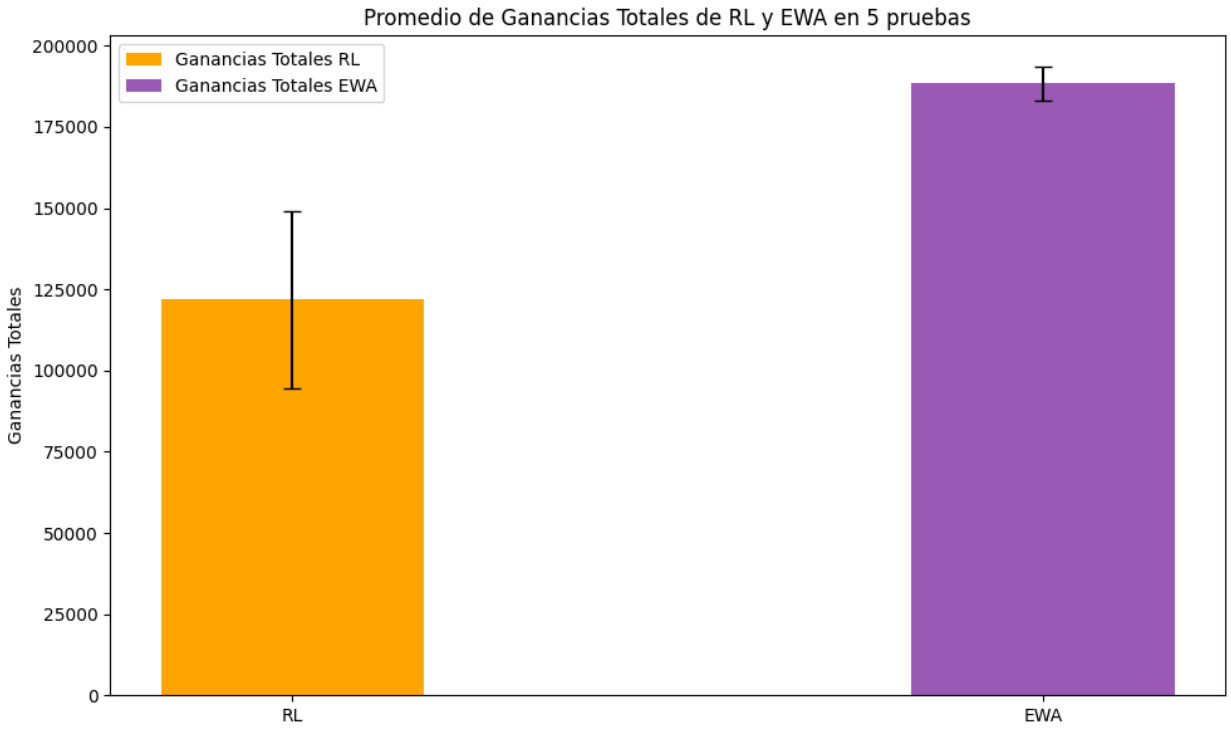

Figura 166: Ganancias totales que obtivieron RL y FEWA cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene mayoría.

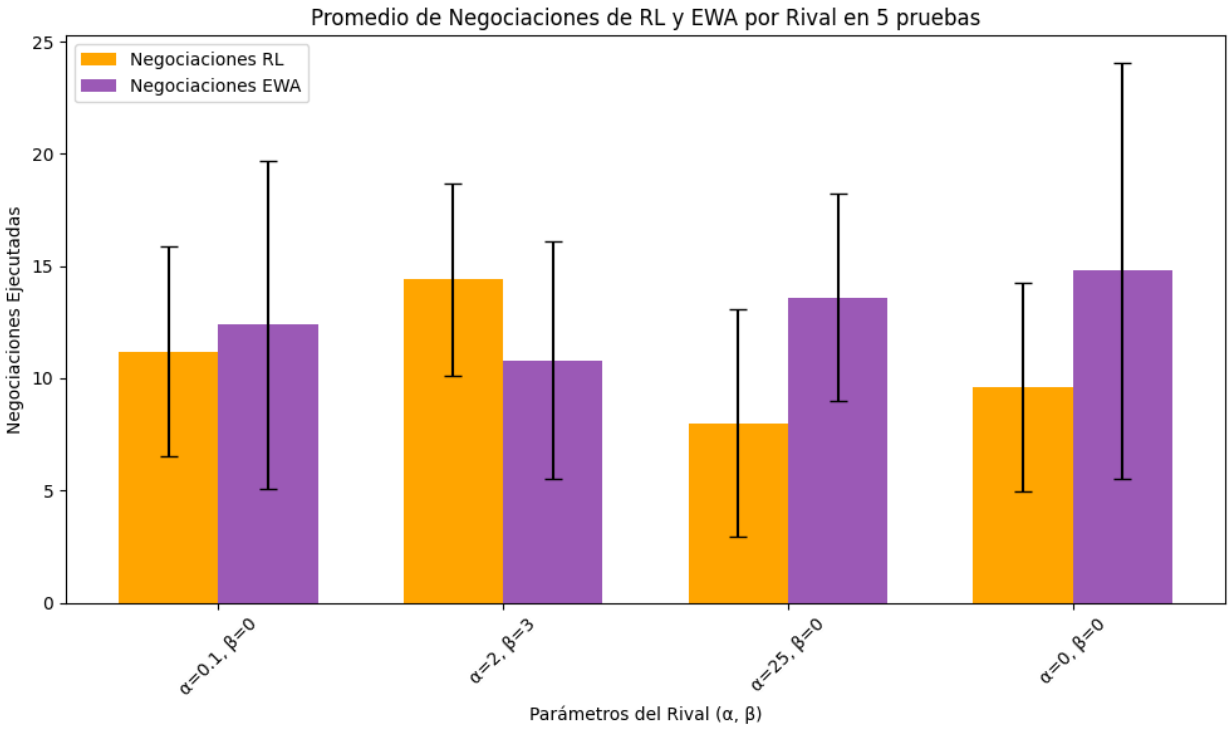

Figura 167: Negociaciones que llevaron a cabo RL y FEWA por cada grupo cuando cuando la población es de 300 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene mayoría.

9.6.2. Distribuciones para tamaño de muestra de 50 agentes

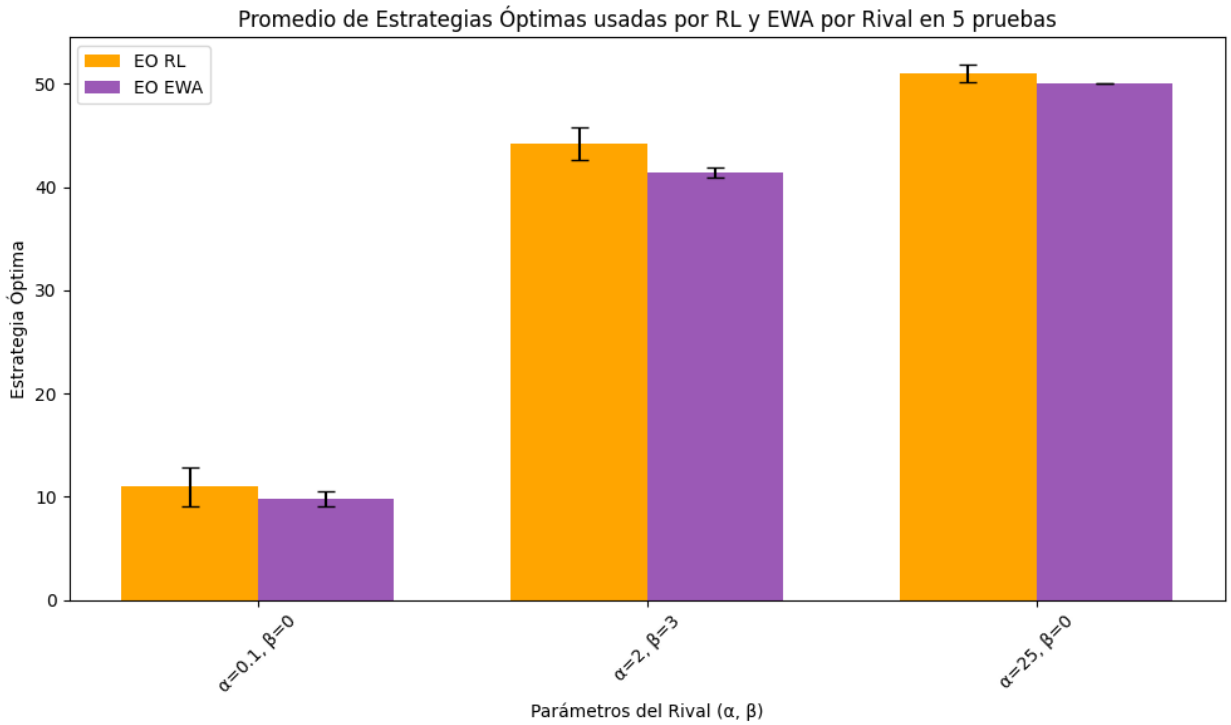

Figura 168: Proporción de estrategias usadas por los agentes RL y FEWA cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA donde todos tienen una proporción uniforme.

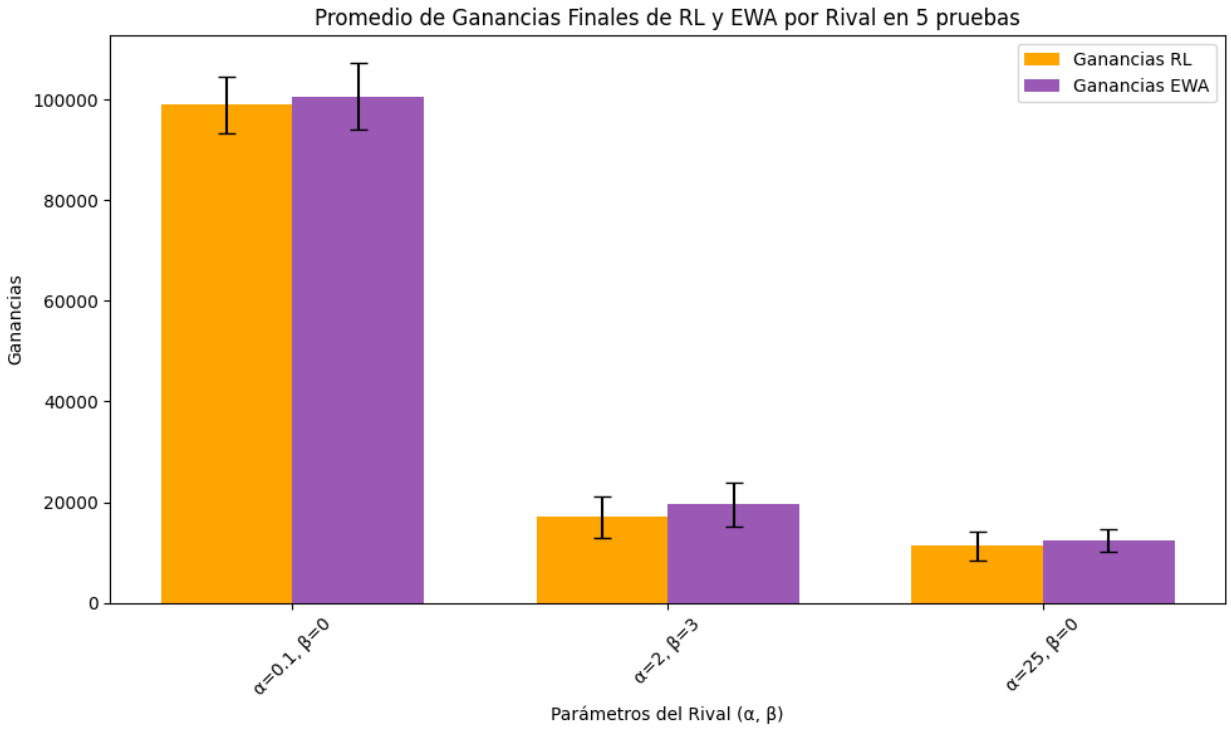

Figura 169: Ganancias que obtivieron RL y FEWA por cada grupo cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA donde todos tienen una proporción uniforme.

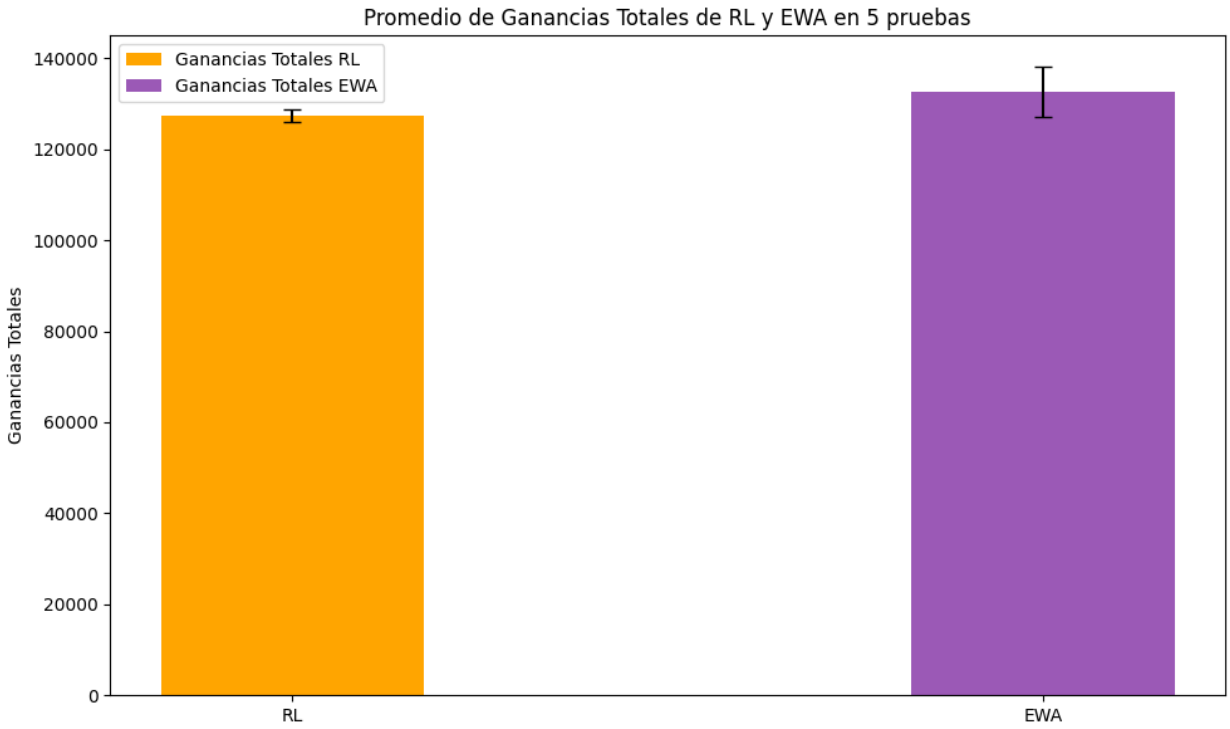

Figura 170: Ganancias totales que obtivieron RL y FEWA cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA donde todos tienen una proporción uniforme.

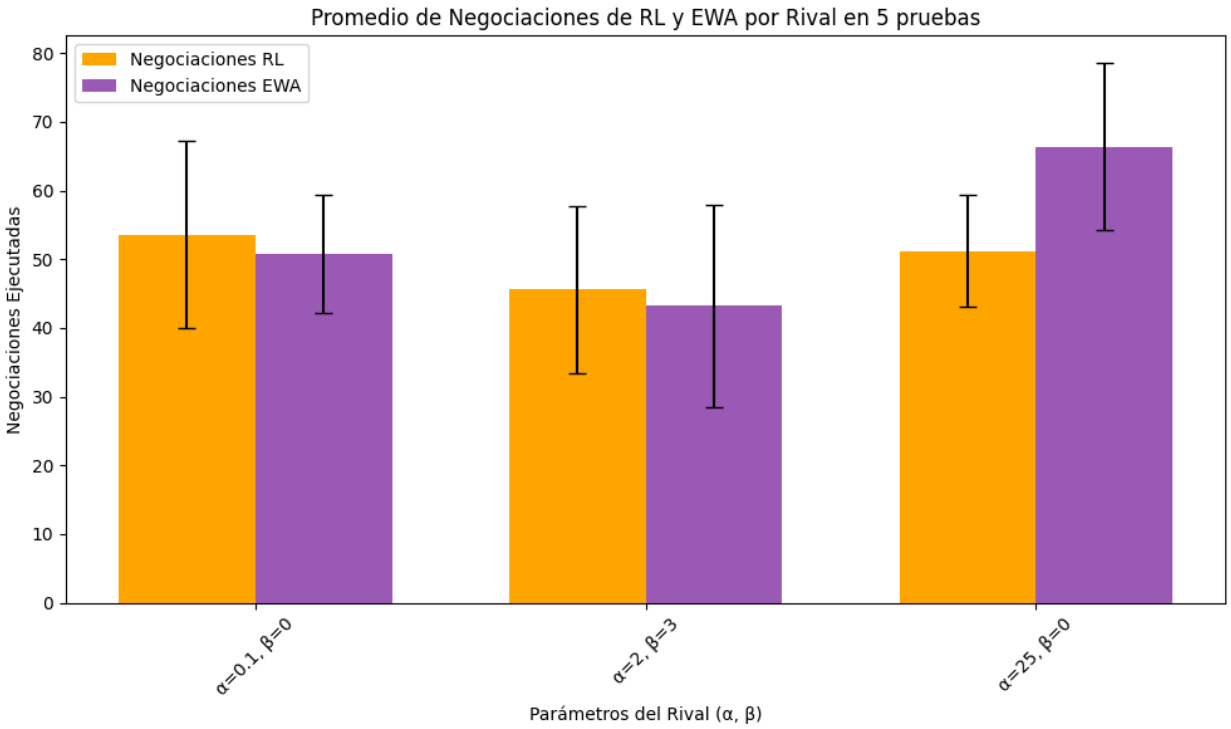

Figura 171: Negociaciones que llevaron a cabo RL y FEWA por cada grupo cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA donde todos tienen una proporción uniforme.

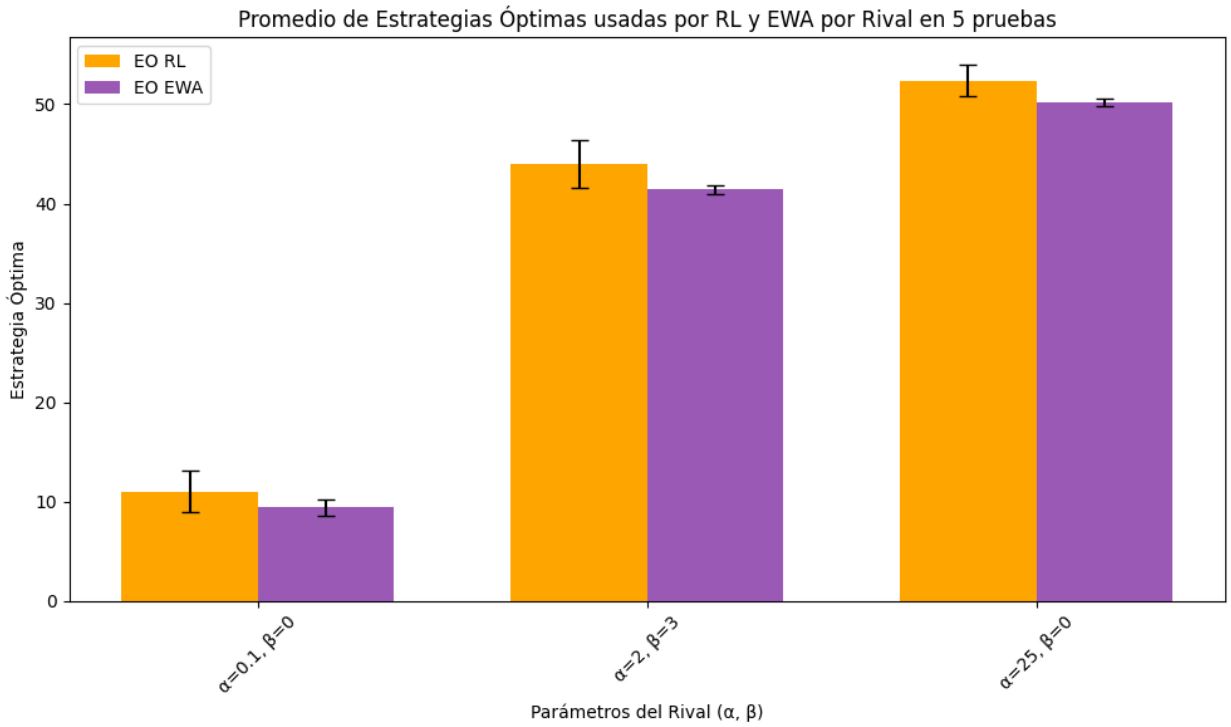

Figura 172: Estrategias usadas por los agentes RL y FEWA cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia alta a resultados ventajosos tienen mayoría.

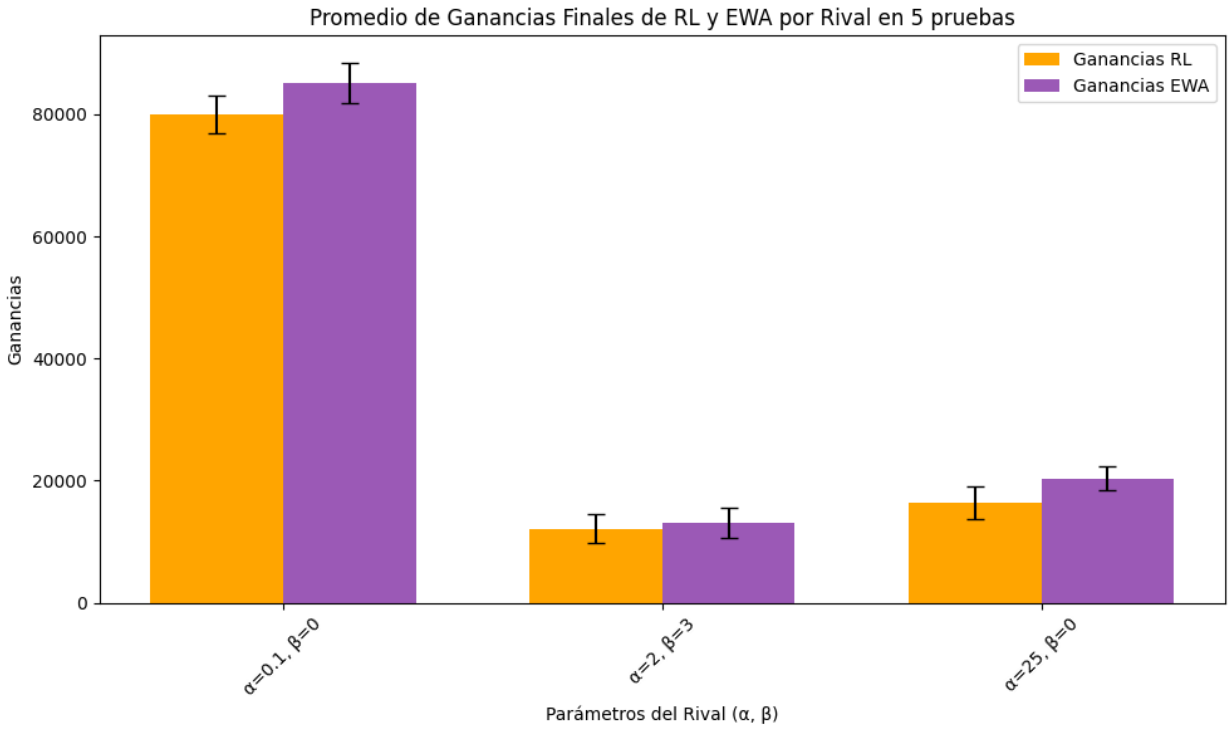

Figura 173: Ganancias que obtivieron RL y FEWA por cada grupo cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia alta a resultados ventajosos tienen mayoría.

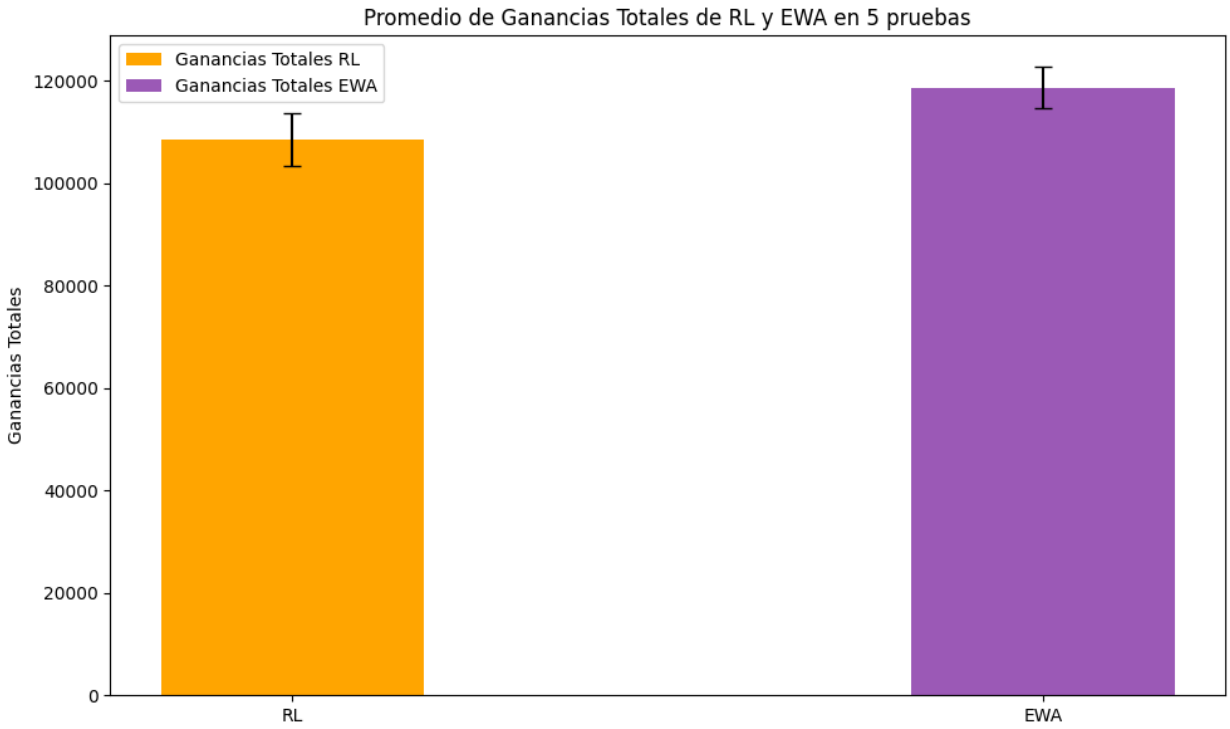

Figura 174: Ganancias totales que obtivieron RL y FEWA cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia alta a resultados ventajosos tienen mayoría.

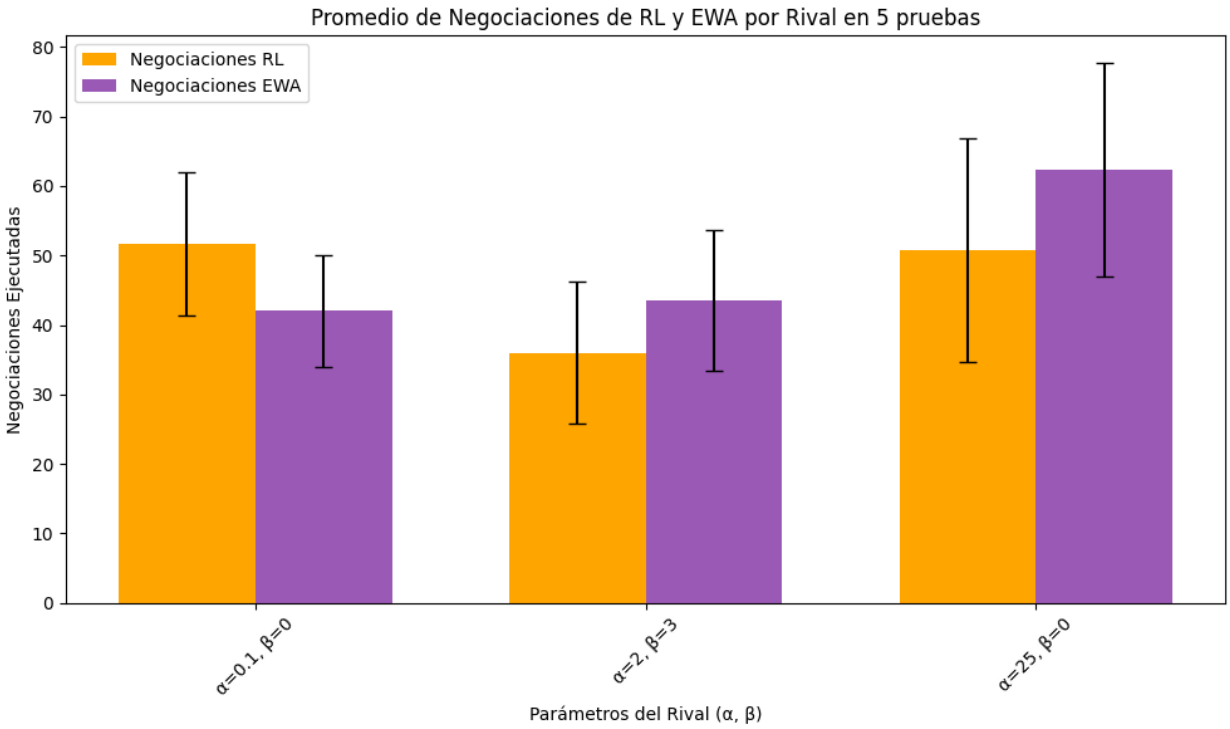

Figura 175: Negociaciones que llevaron a cabo RL y FEWA por cada grupo cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia alta a resultados ventajosos tienen mayoría.

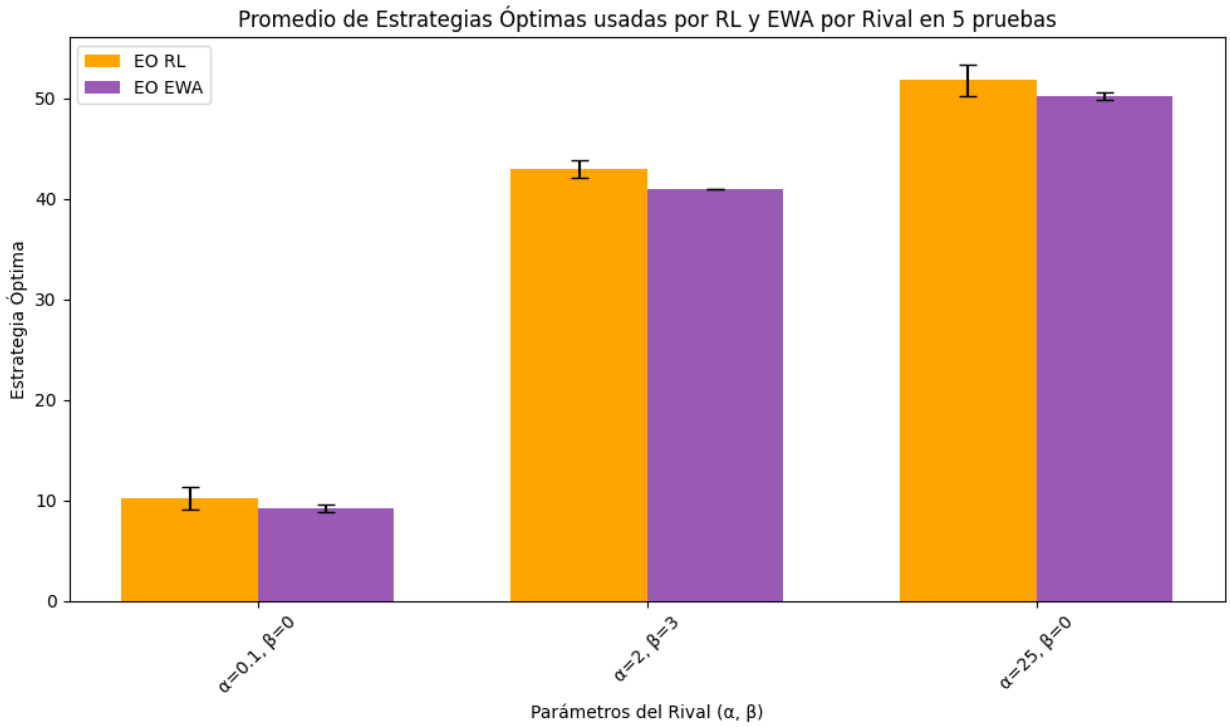

Figura 176: Estrategias usadas por los agentes RL y FEWA cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia baja a resultados ventajosos tienen mayoría.

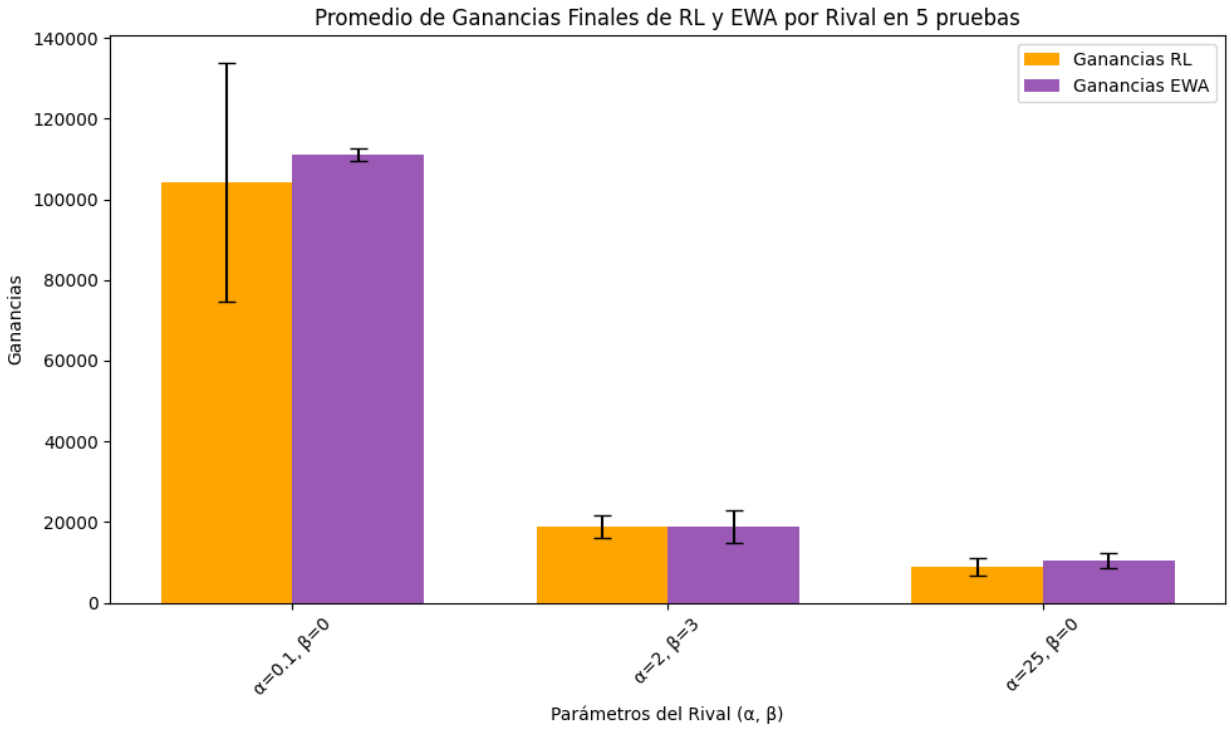

Figura 177: Ganancias que obtivieron RL y FEWA por cada grupo cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia baja a resultados ventajosos tienen mayoría.

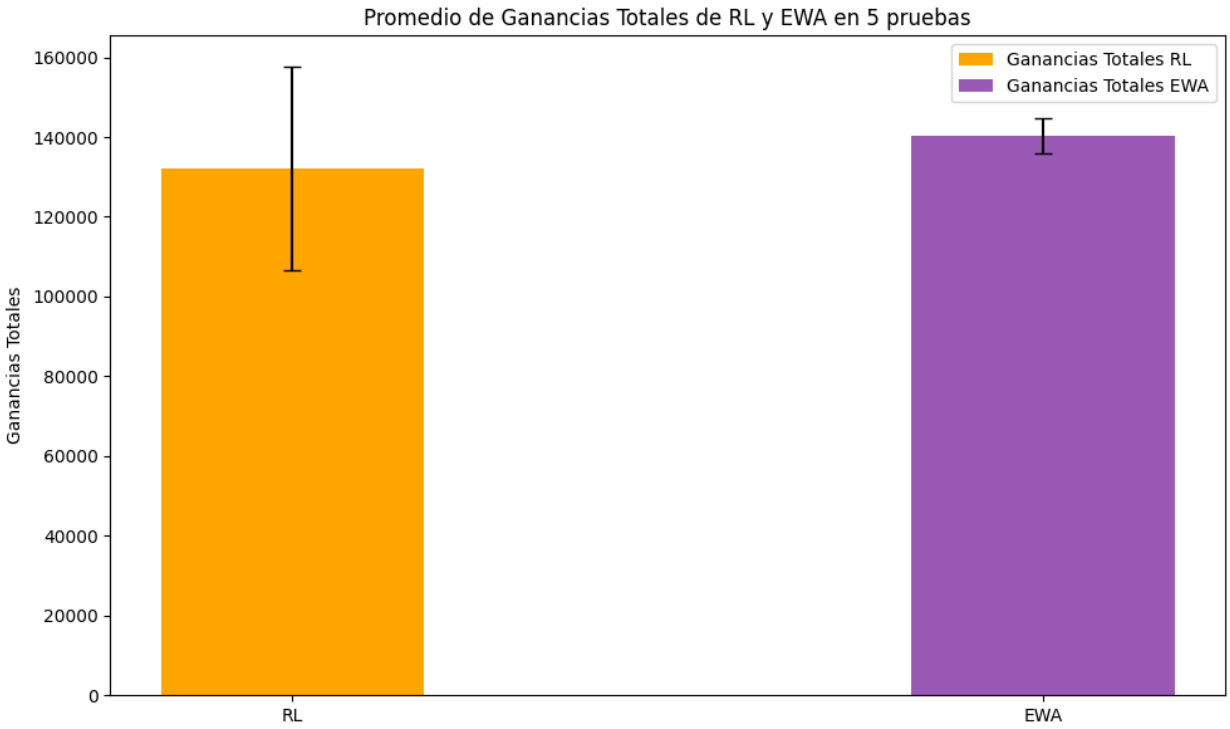

Figura 178: Ganancias totales que obtivieron RL y FEWA por cada grupo cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia baja a resultados ventajosos tienen mayoría.

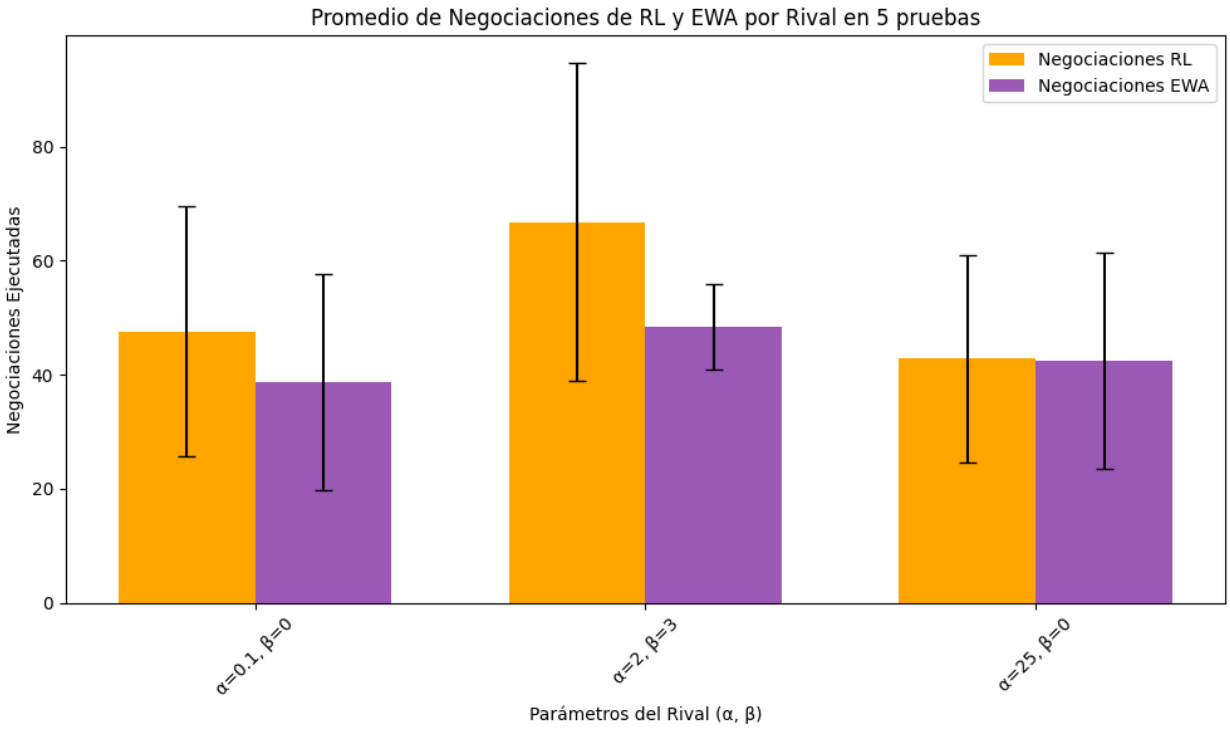

Figura 179: Negociaciones que llevaron a cabo RL y FEWA por cada grupo cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia baja a resultados ventajosos tienen mayoría.

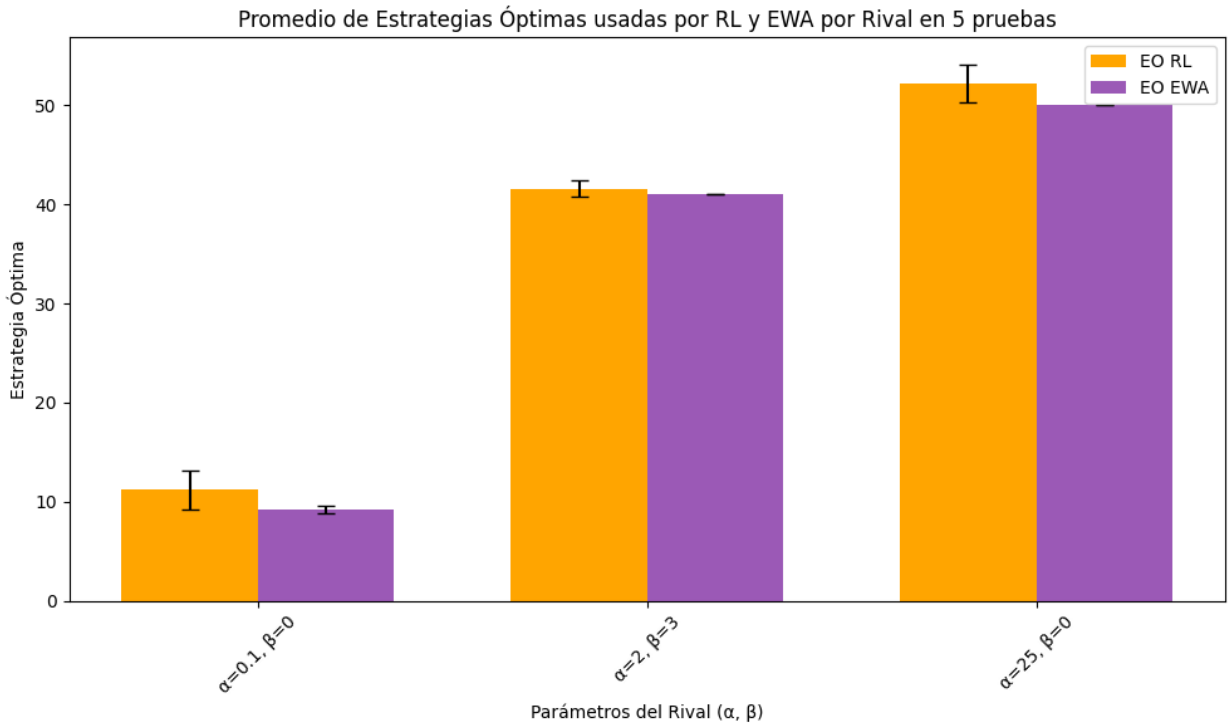

Figura 180: Estrategias usadas por los agentes RL y FEWA cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia media a resultados justos tienen mayoría.

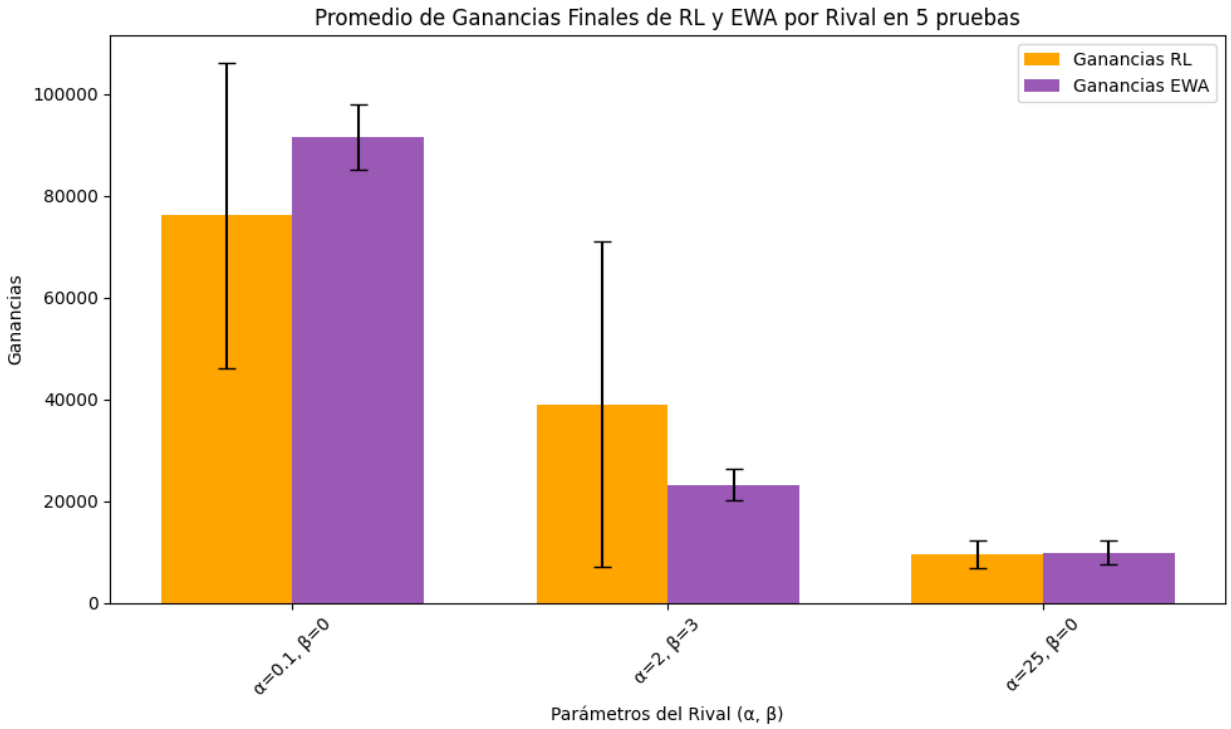

Figura 181: Ganancias que obtivieron RL y FEWA por cada grupo cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia media a resultados justos tienen mayoría.

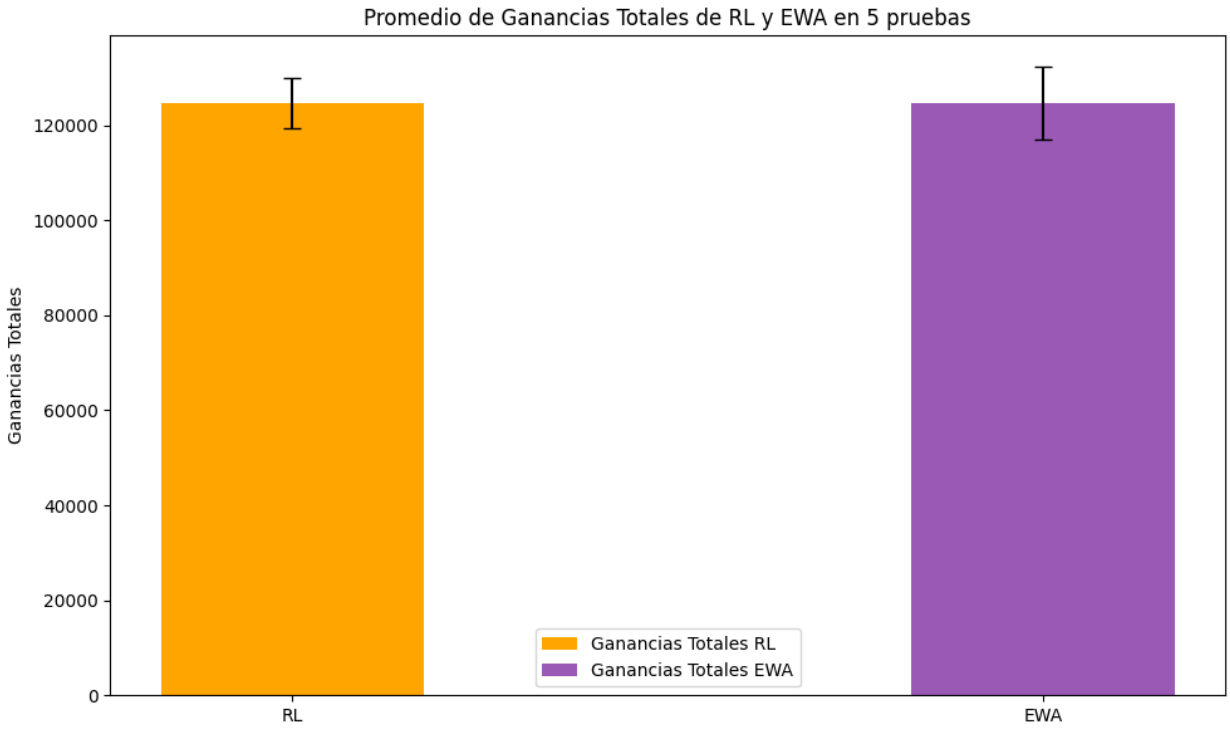

Figura 182: Ganancias totales que obtivieron RL y FEWA cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia media a resultados justos tienen mayoría.

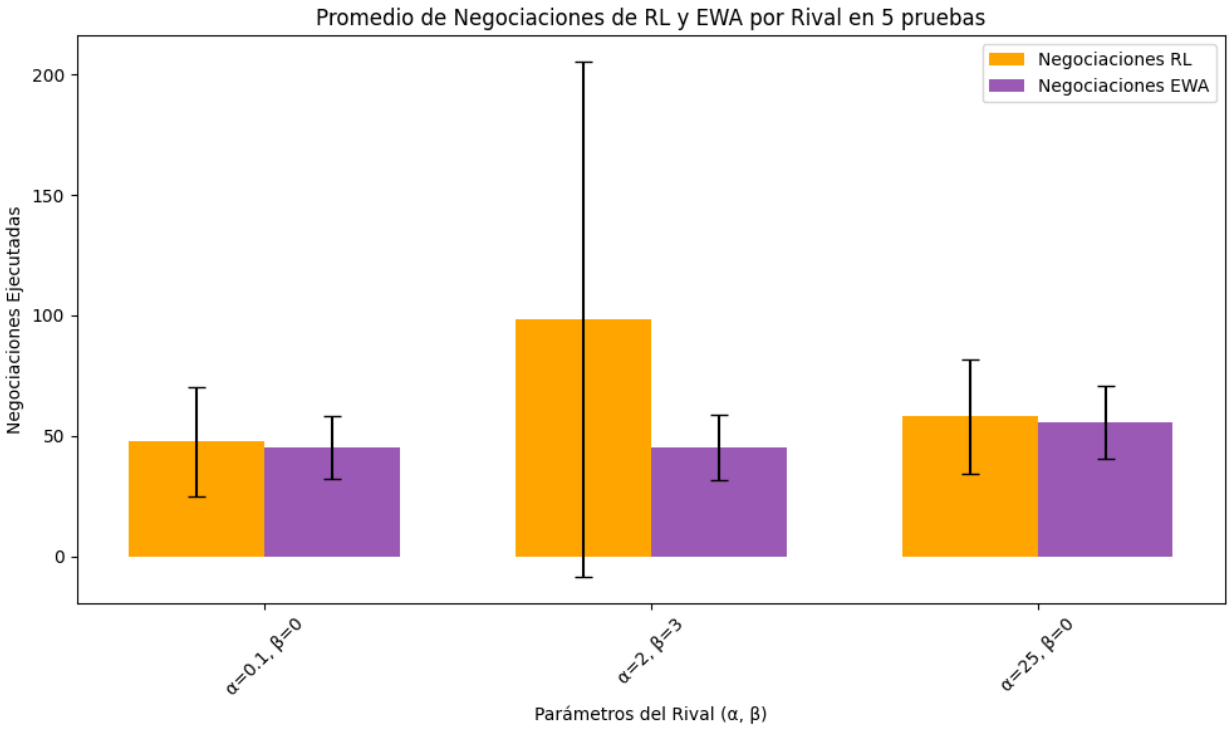

Figura 183: Negociaciones que llevaron a cabo RL y FEWA por cada grupo cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA donde los rivales con preferencia media a resultados justos tienen mayoría.

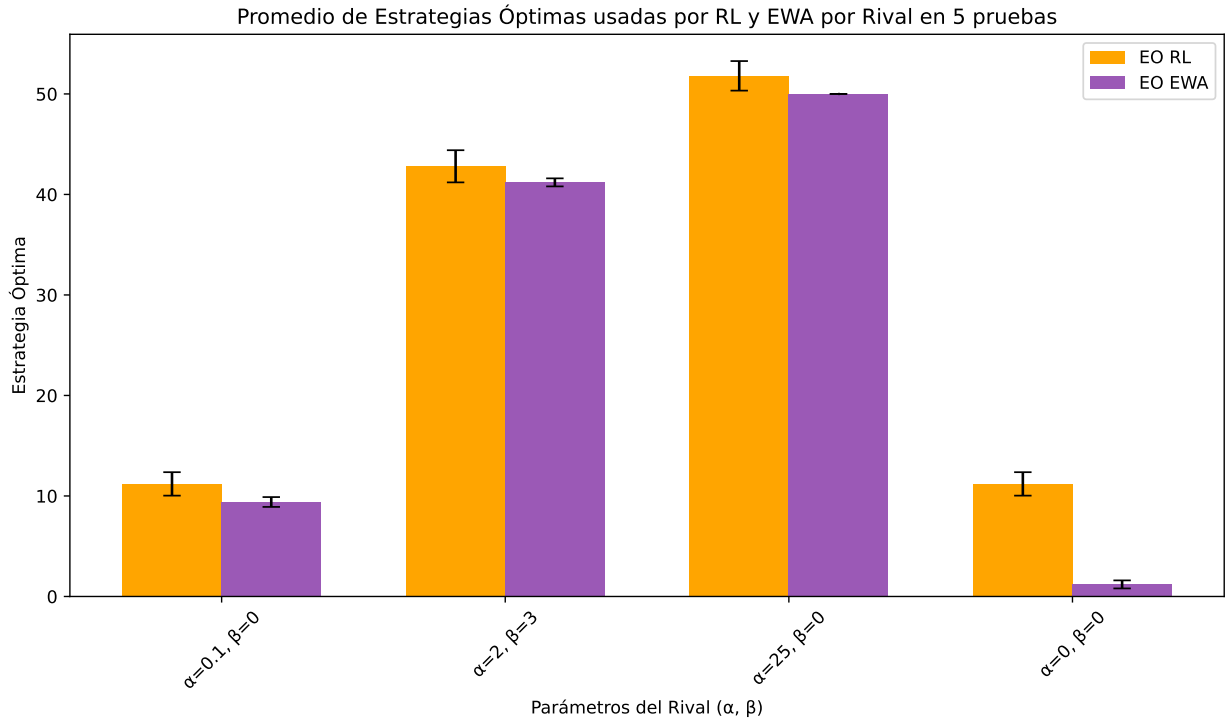

Figura 184: Estrategias usadas por los agentes RL y FEWA cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde todos tienen una proporción uniforme.

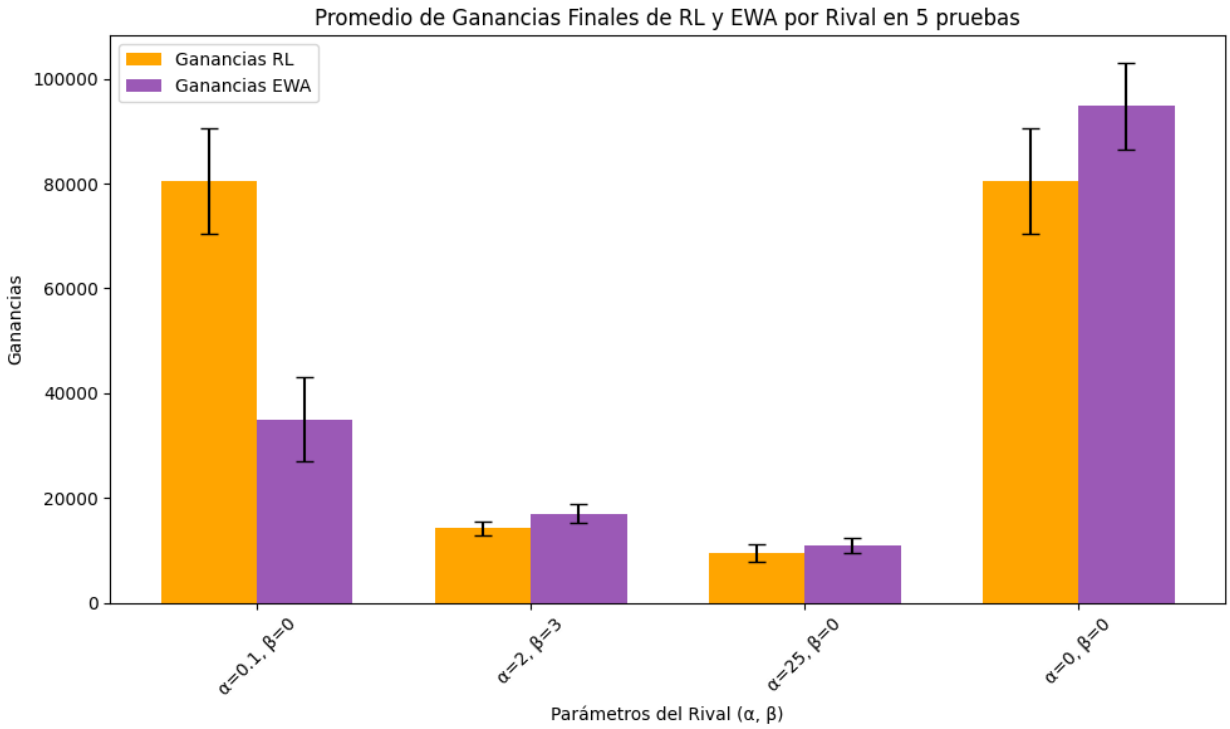

Figura 185: Ganancias que obtivieron RL y FEWA por cada grupo cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde todos tienen una proporción uniforme.

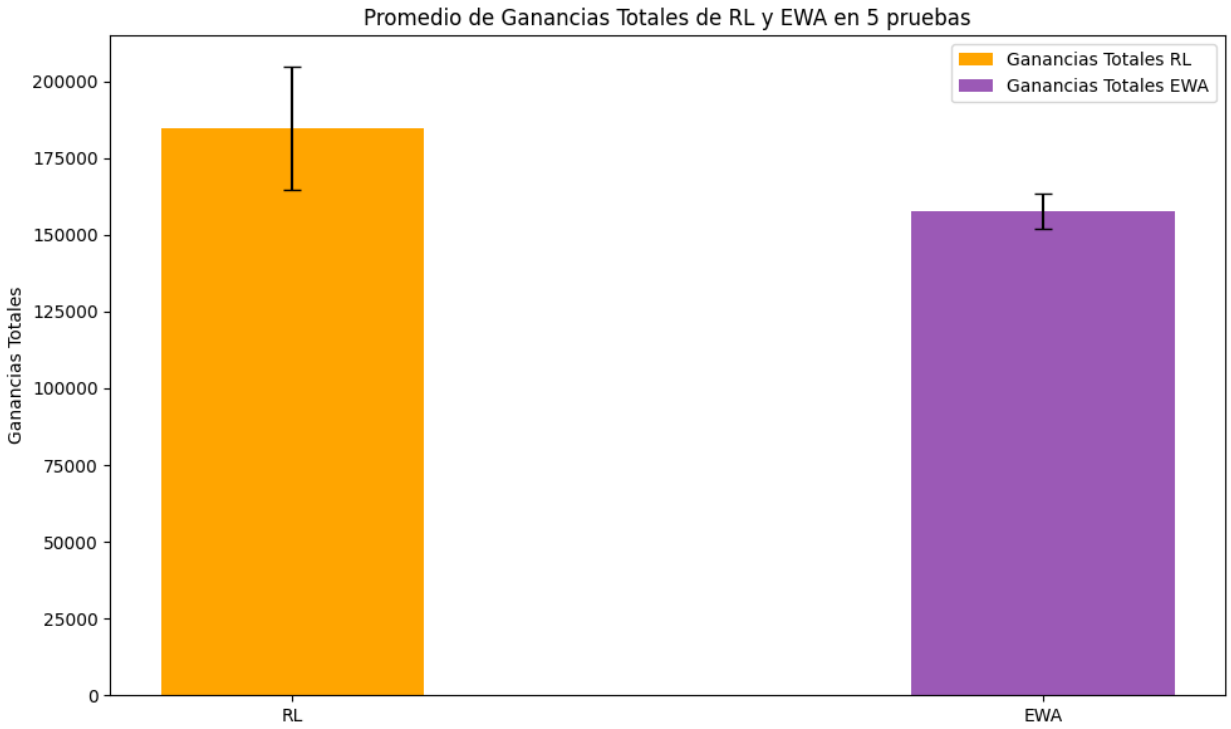

Figura 186: Ganancias totales que obtivieron RL y FEWA cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde todos tienen una proporción uniforme.

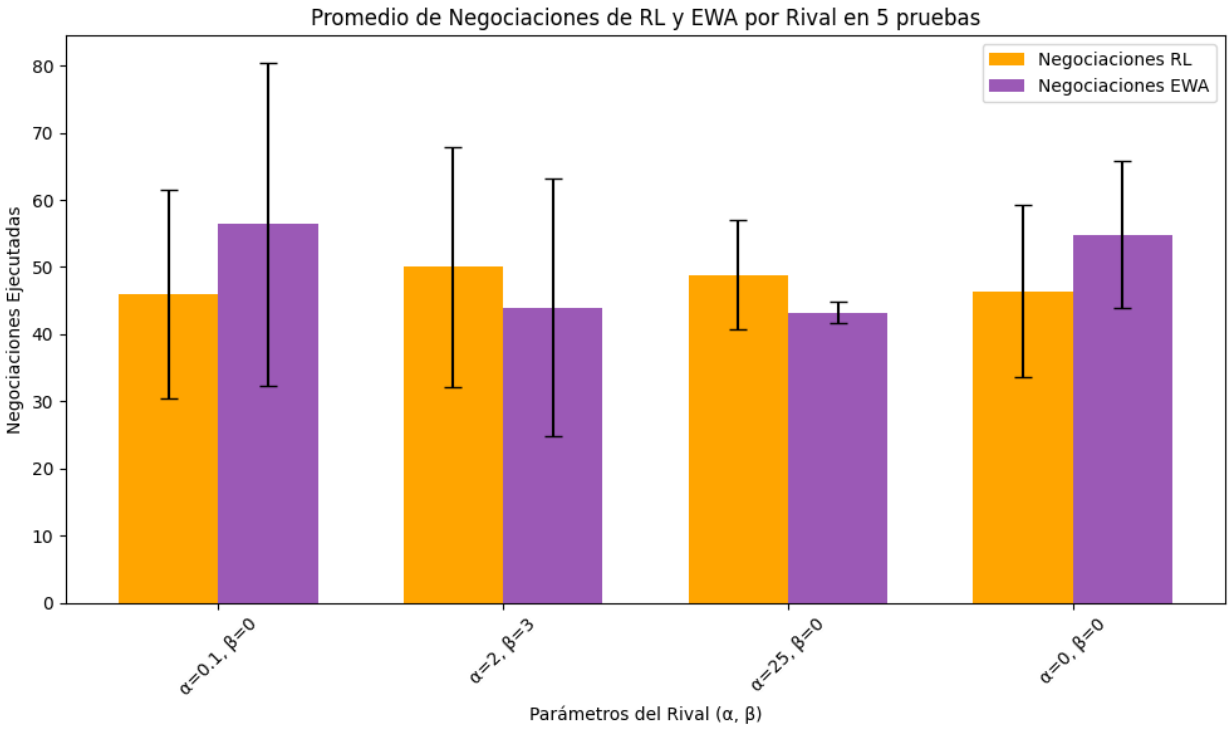

Figura 187: Negociaciones que llevaron a cabo RL y FEWA por cada grupo cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde todos tienen una proporción uniforme.

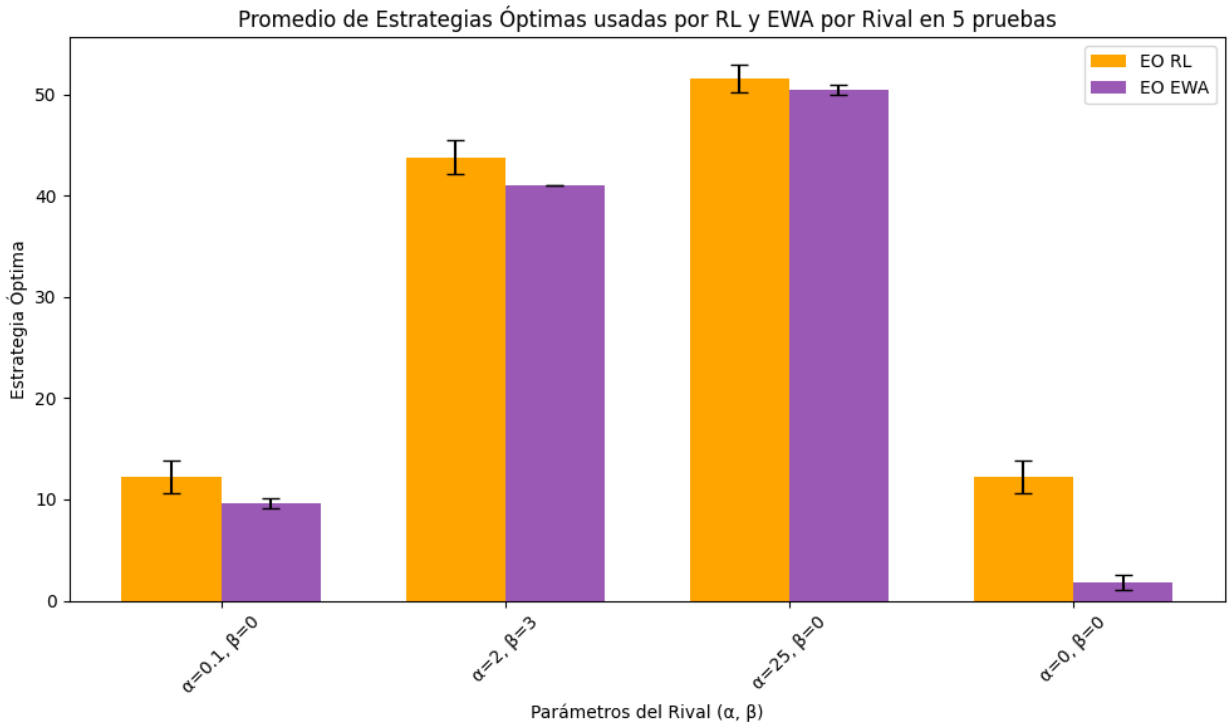

Figura 188: Estrategias usadas por los agentes RL y FEWA cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene minoría.

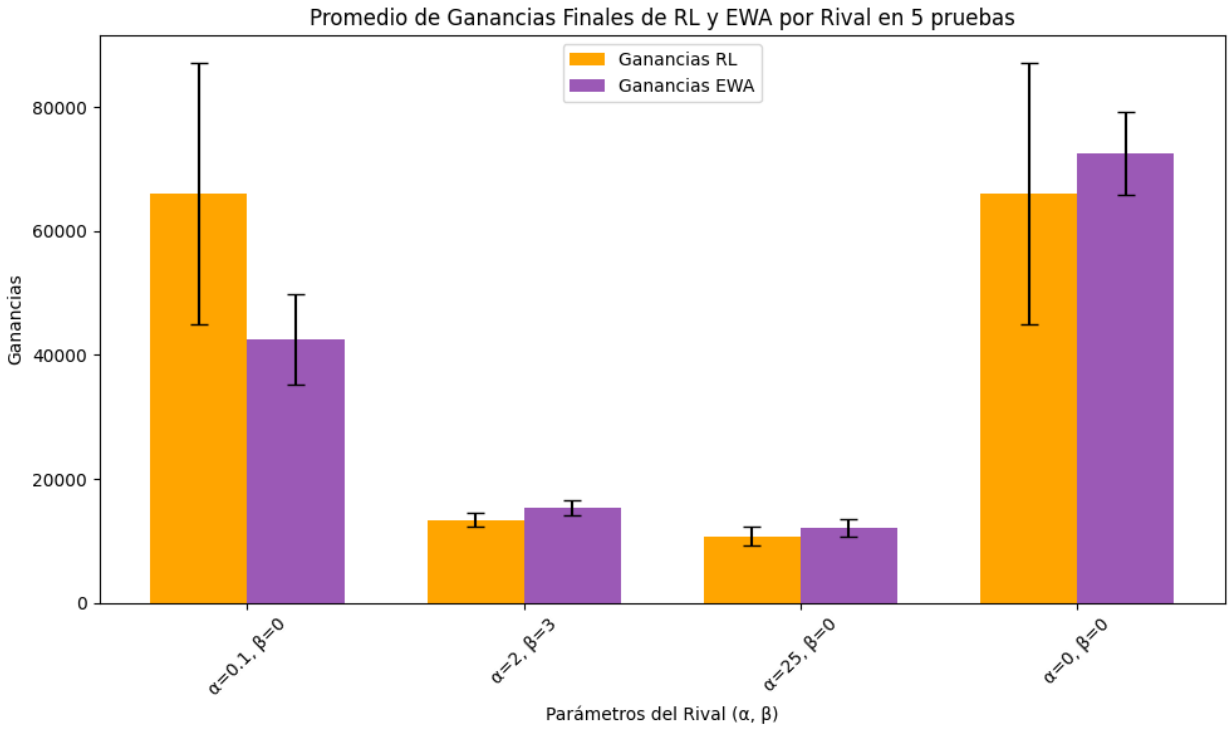

Figura 189: Ganancias que obtivieron RL y FEWA por cada grupo cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene minoría.

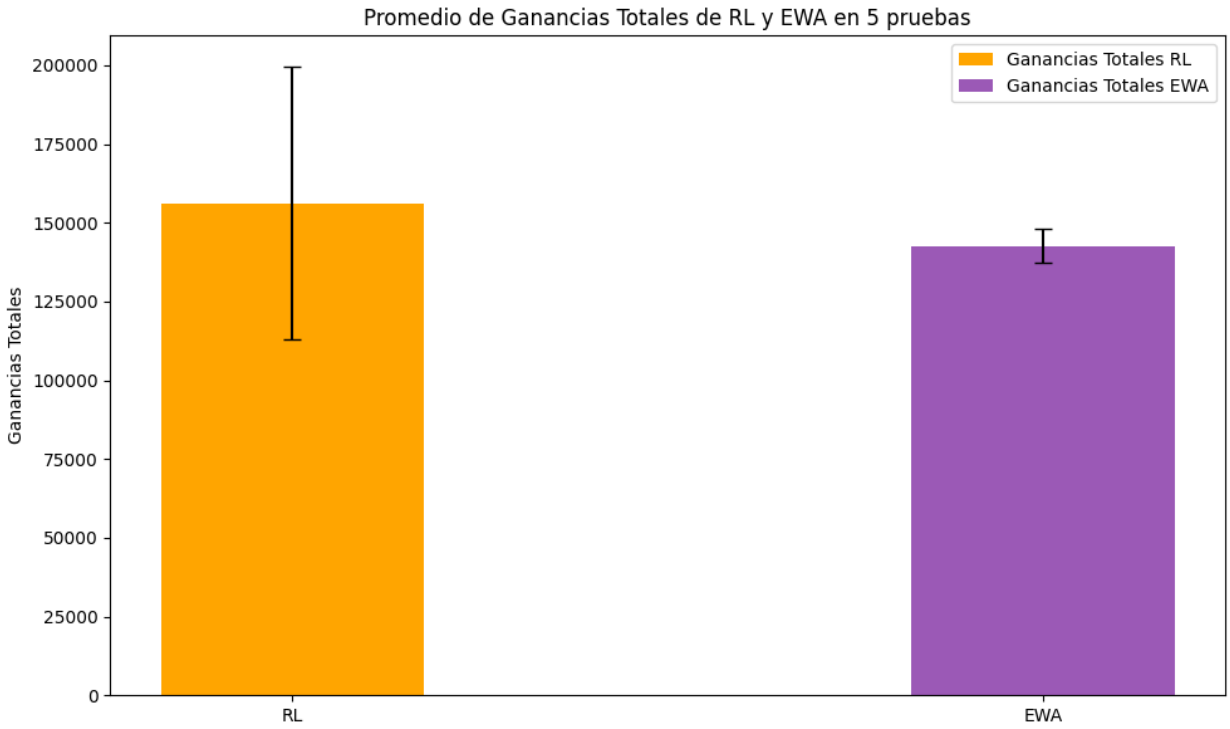

Figura 190: Ganancias totales que obtivieron RL y FEWA cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene minoría.

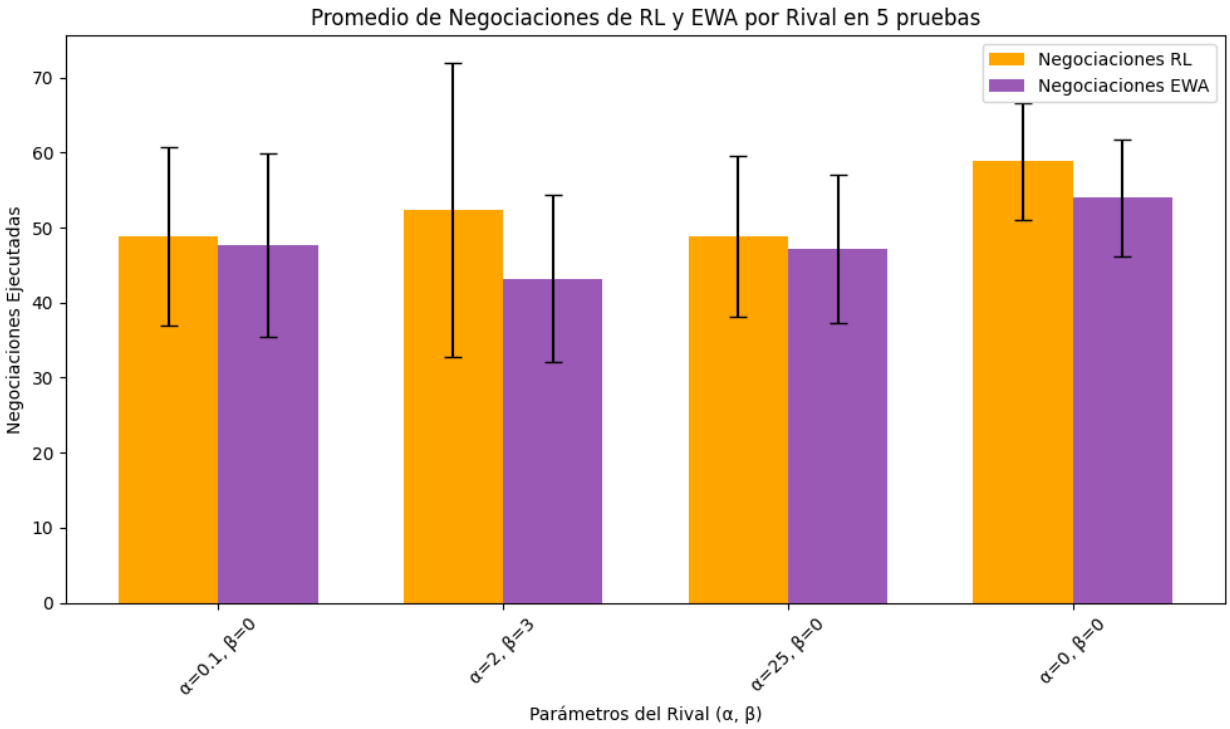

Figura 191: Negociaciones que llevaron a cabo RL y FEWA por cada grupo cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene minoría.

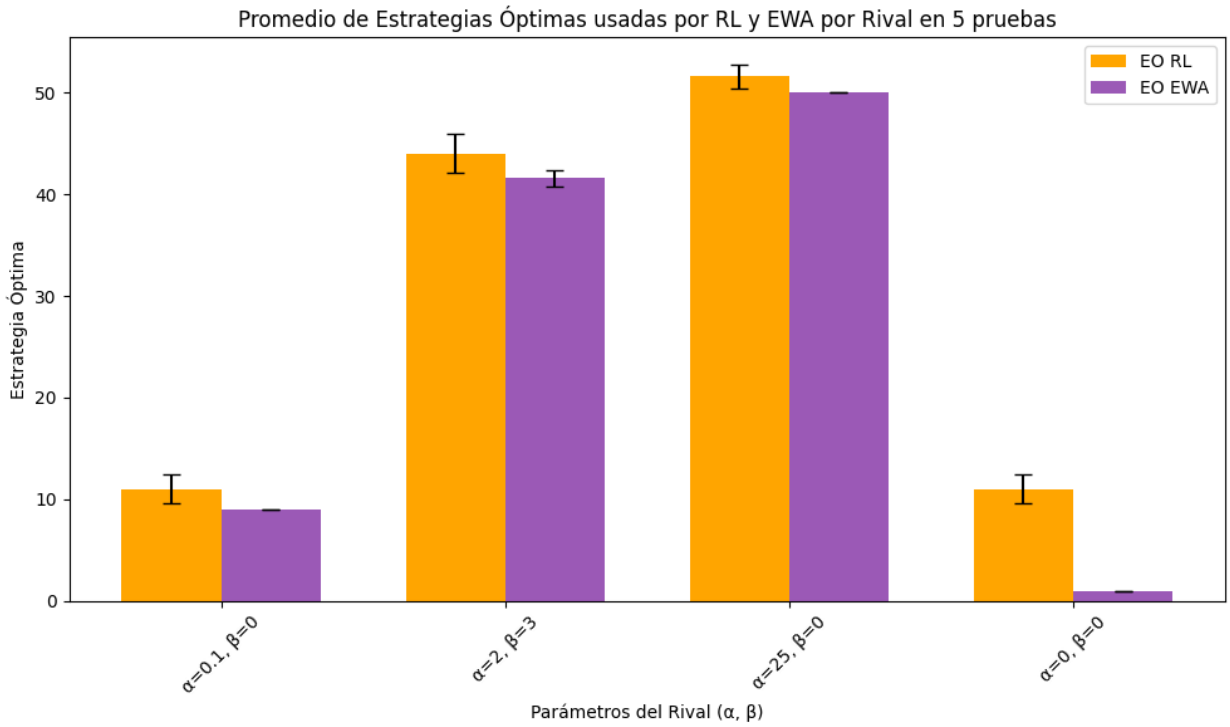

Figura 192: Estrategias usadas por los agentes RL y FEWA cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene mayoría.

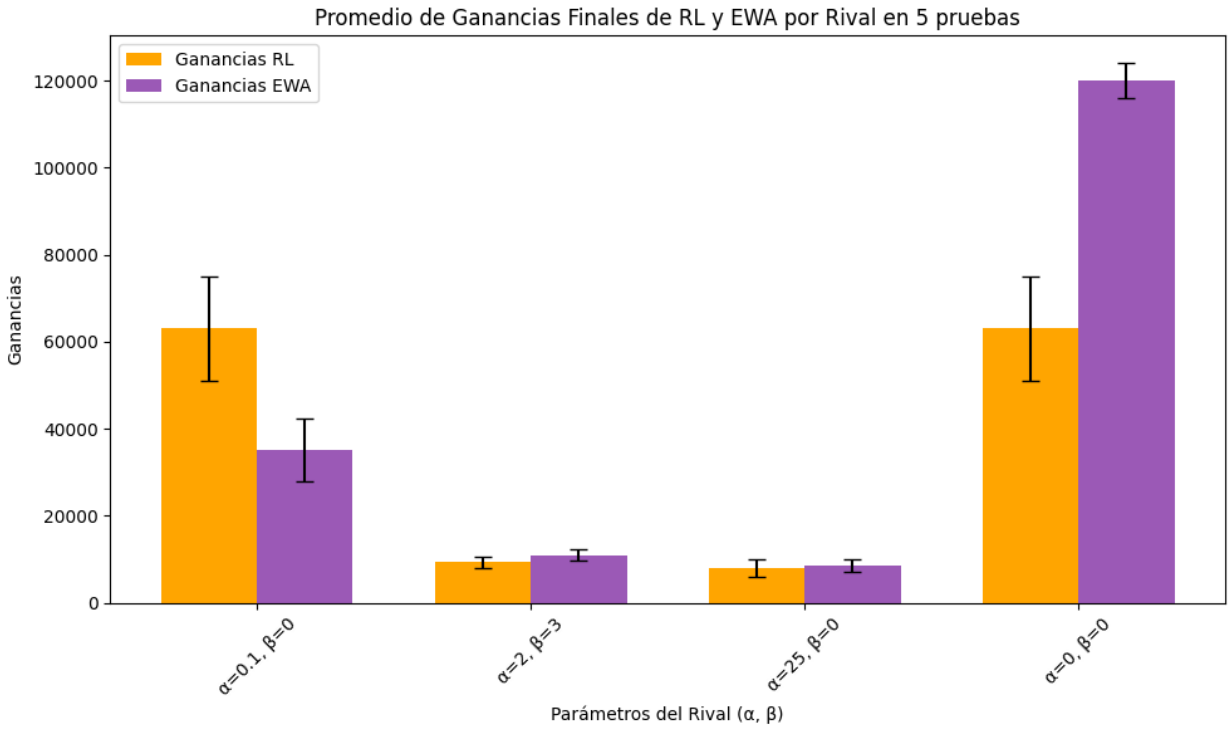

Figura 193: Ganancias que obtivieron RL y FEWA por cada grupo cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene mayoría.

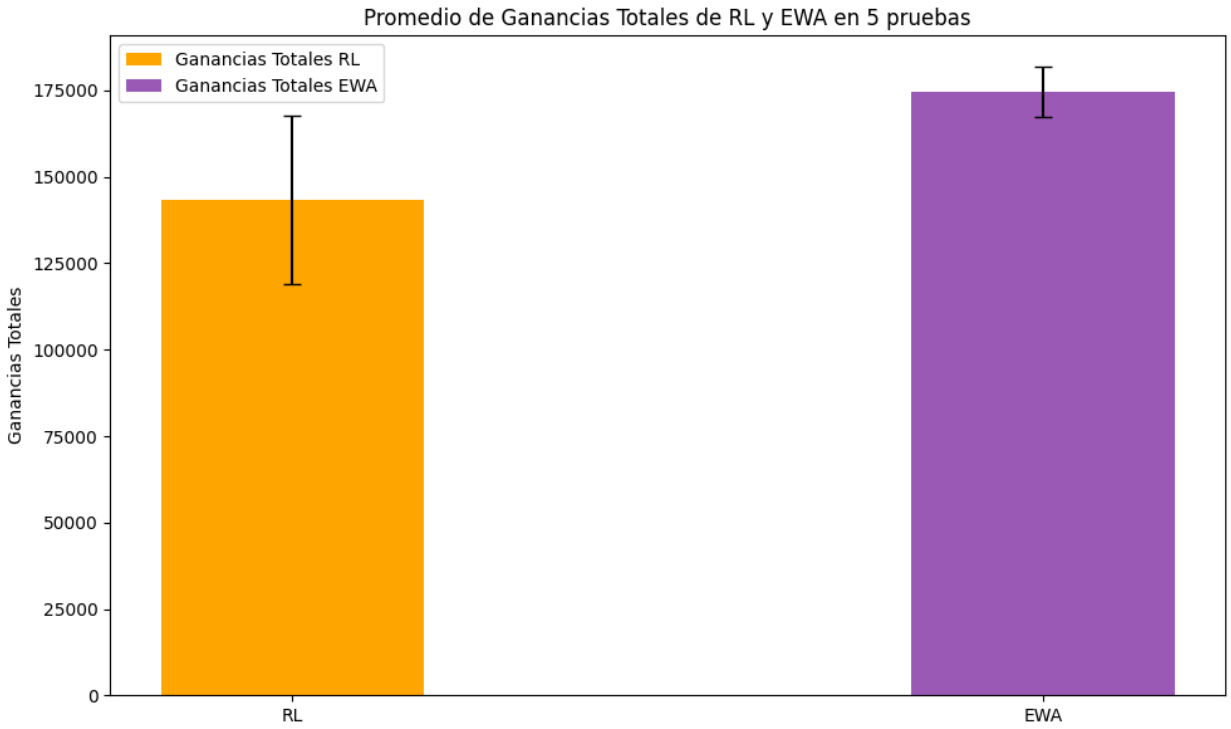

Figura 194: Ganancias totales que obtivieron RL y FEWA cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene mayoría.

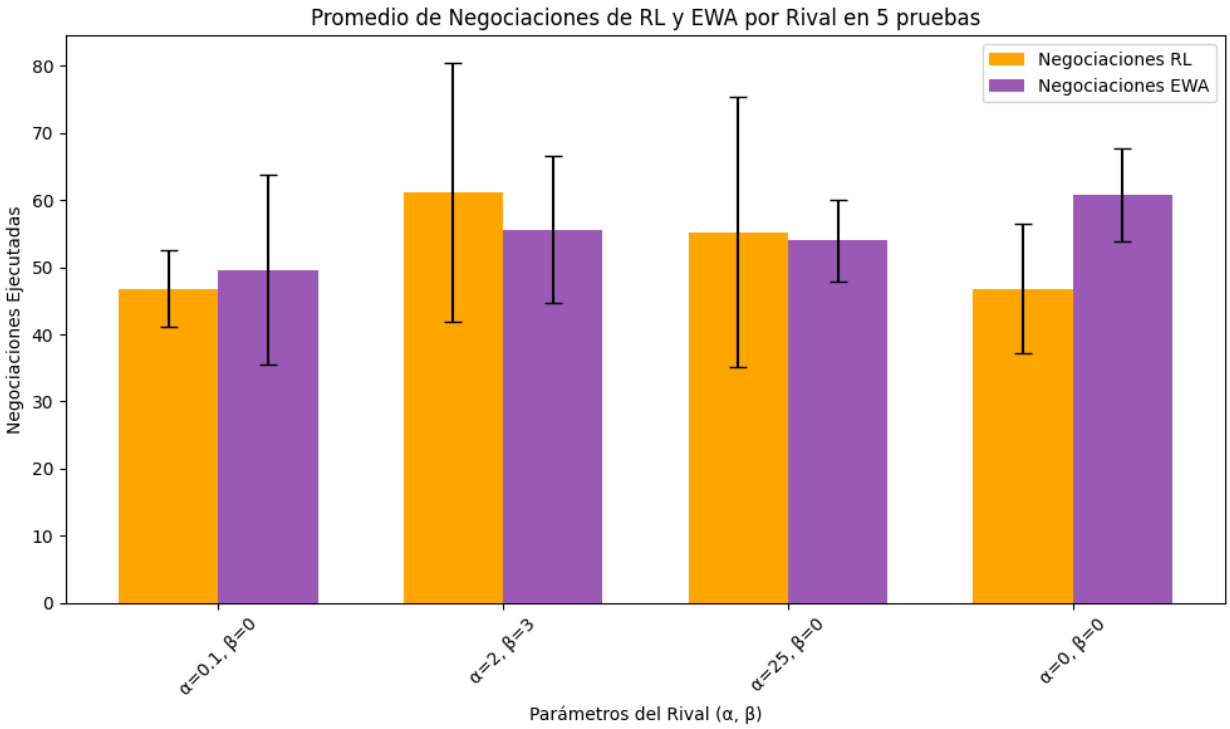

Figura 195: Negociaciones que llevaron a cabo RL y FEWA por cada grupo cuando cuando la población es de 50 agentes y la distribución consta de rivales conocidos para RL y FEWA y un rival desconocido donde el rival desconocido tiene mayoría.

9.7. Gráficos de Segregación

9.7.1. Distribuciones para tamaño de muestra de 150 agentes

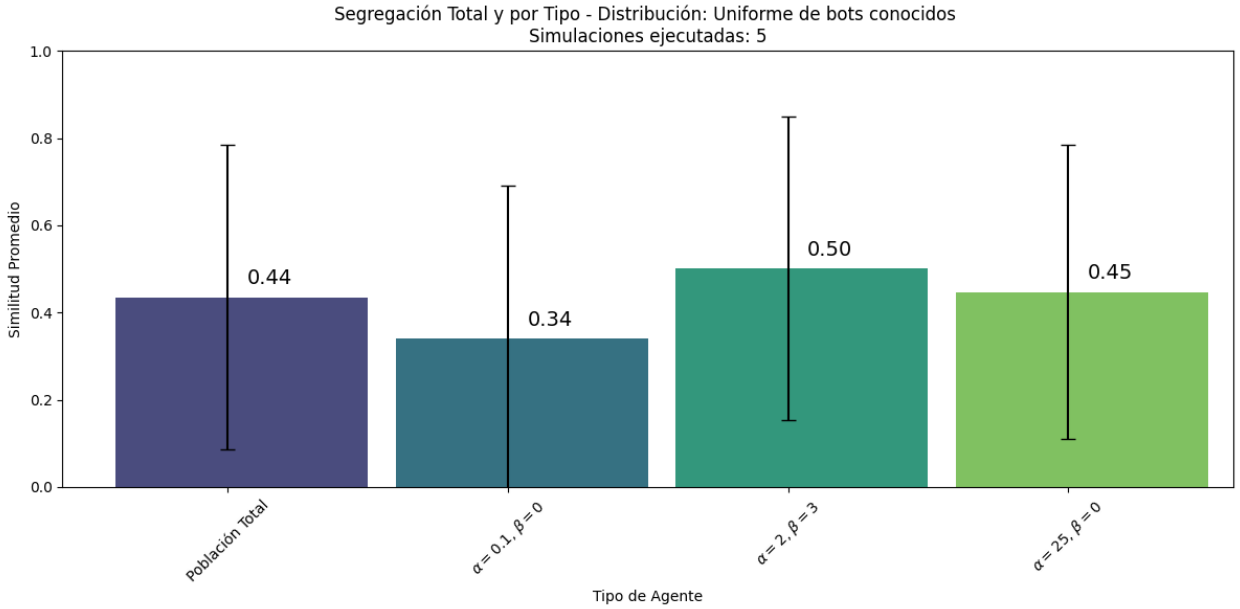

Figura 196: Segregación para una población de 150 rivales con una distribución uniforme de los tres rivales conocidos para los agentes RL y FEWA.

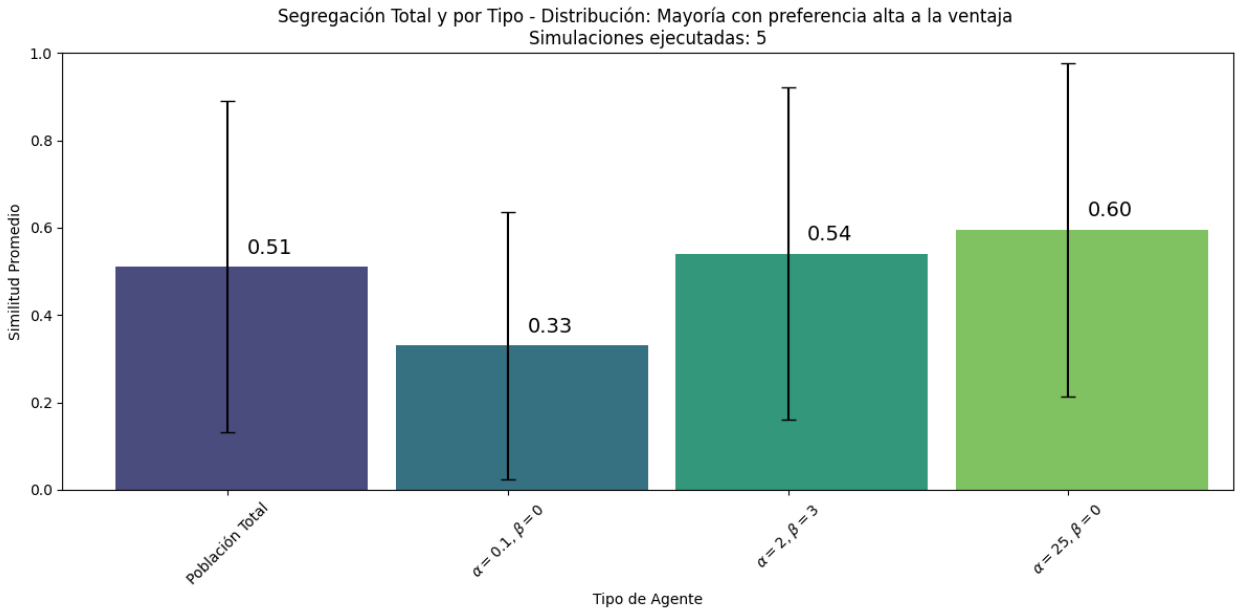

Figura 197: Segregación para una población de 150 rivales con una distribución de los tres rivales conocidos para los agentes RL y FEWA donde el rival con preferencia alta a resultados ventajosos tiene una proporción más alta de bots en la población y los agentes restantes tiene la misma proporción.

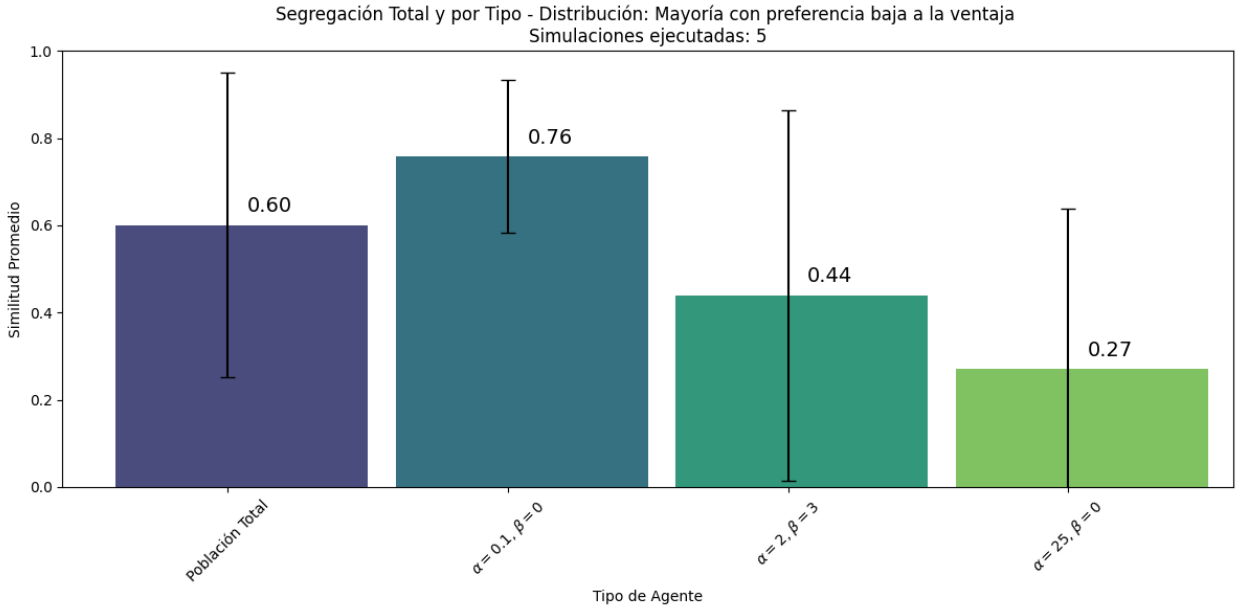

Figura 198: Segregación para una población de 150 rivales con una distribución de los tres rivales conocidos para los agentes RL y FEWA donde el rival con preferencia baja a resultados ventajosos tiene una proporción más alta de bots en la población y los agentes restantes tiene la misma proporción.

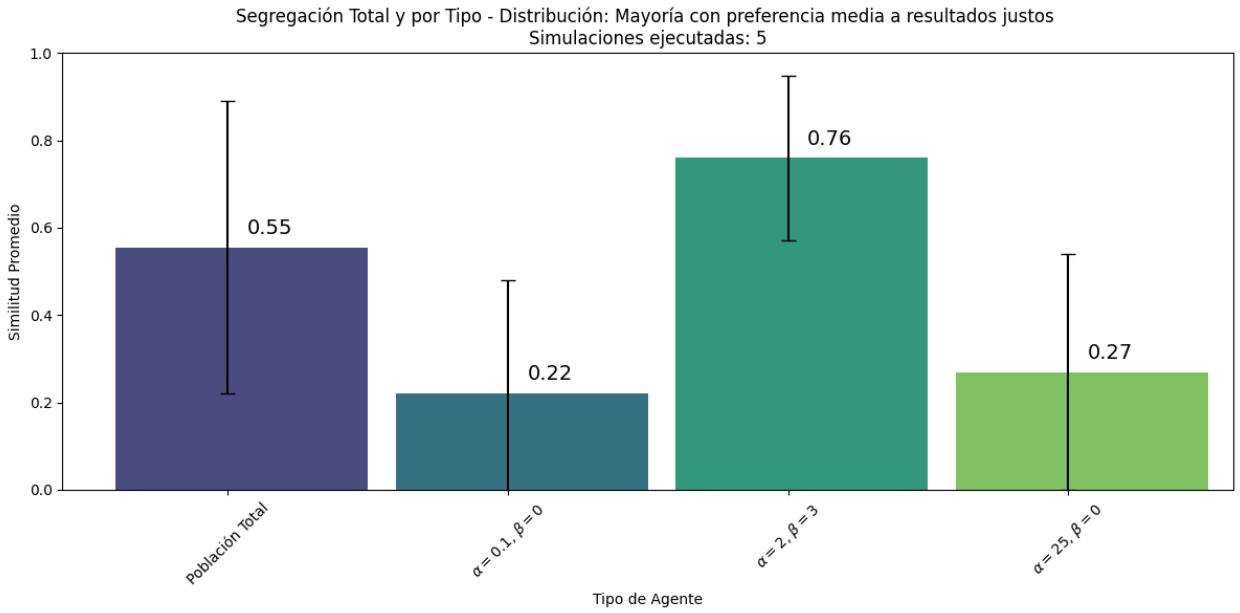

Figura 199: Segregación para una población de 150 rivales con una distribución de los tres rivales conocidos para los agentes RL y FEWA donde el rival con preferencia media a resultados justos tiene una proporción más alta de bots en la población y los agentes restantes tiene la misma proporción..

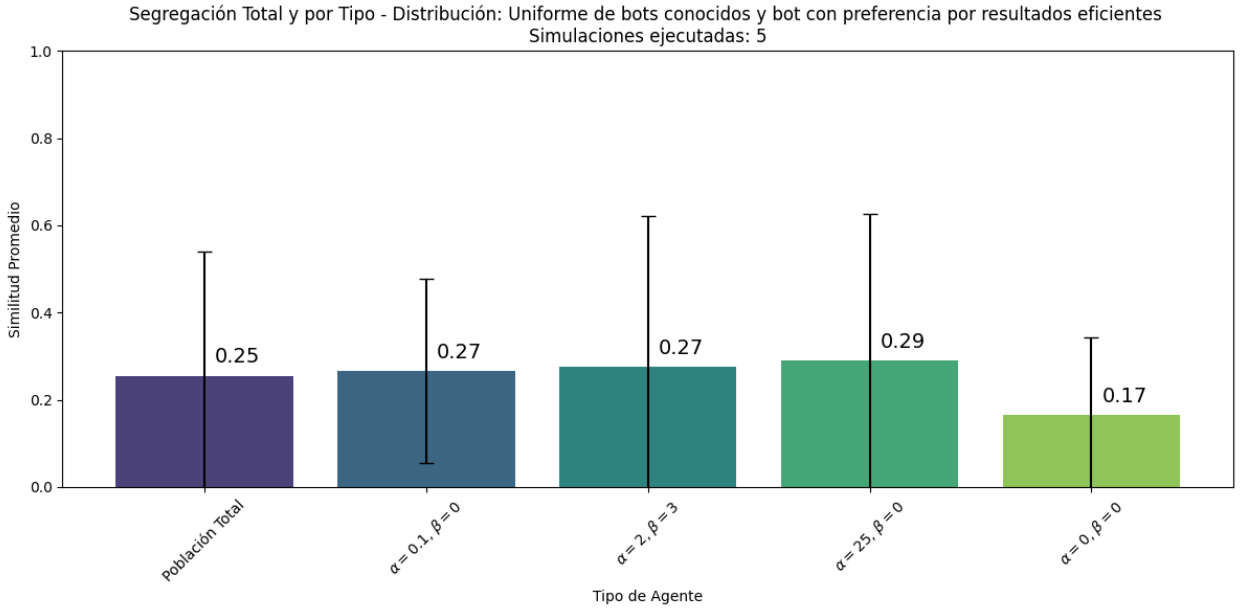

Figura 200: Segregación para una población de 150 rivales con una distribución uniforme de los tres rivales conocidos para los agentes RL y FEWA y de un rival desconocido, el rival con preferencia por resultados eficientes.

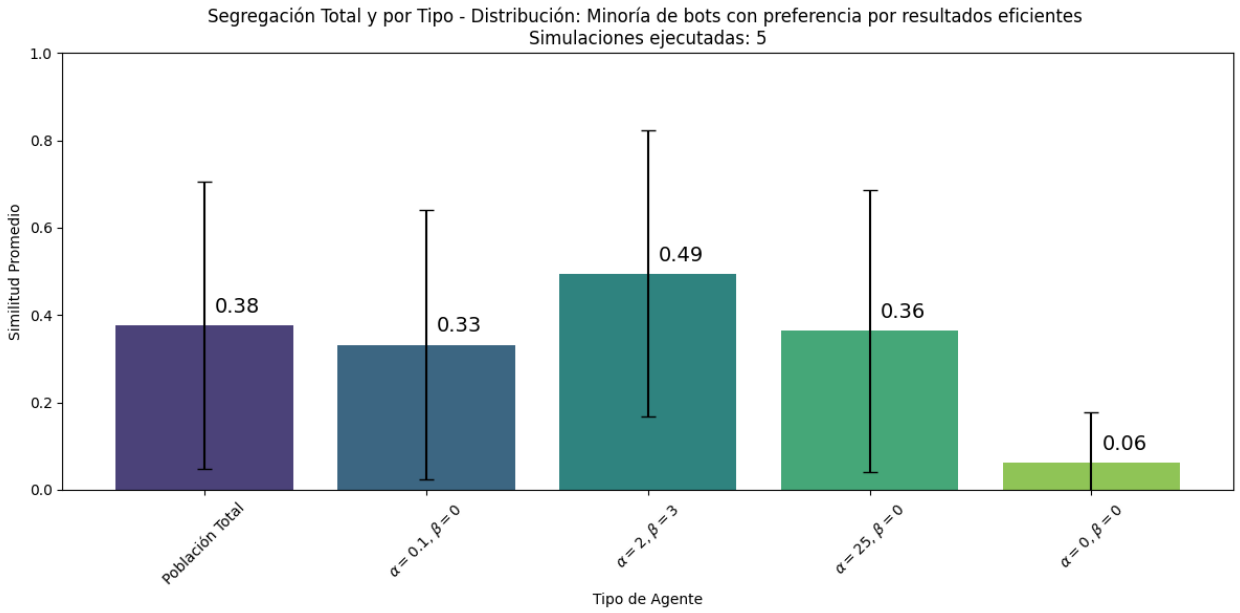

Figura 201: Segregación para una población de 150 rivales con una distribución de los tres rivales conocidos para los agentes RL y FEWA y el rival desconocido, donde el rival desconocido con preferencia por resultados eficientes tiene la proporción más baja de bots en la población, y los agentes restantes tiene la misma proporción.

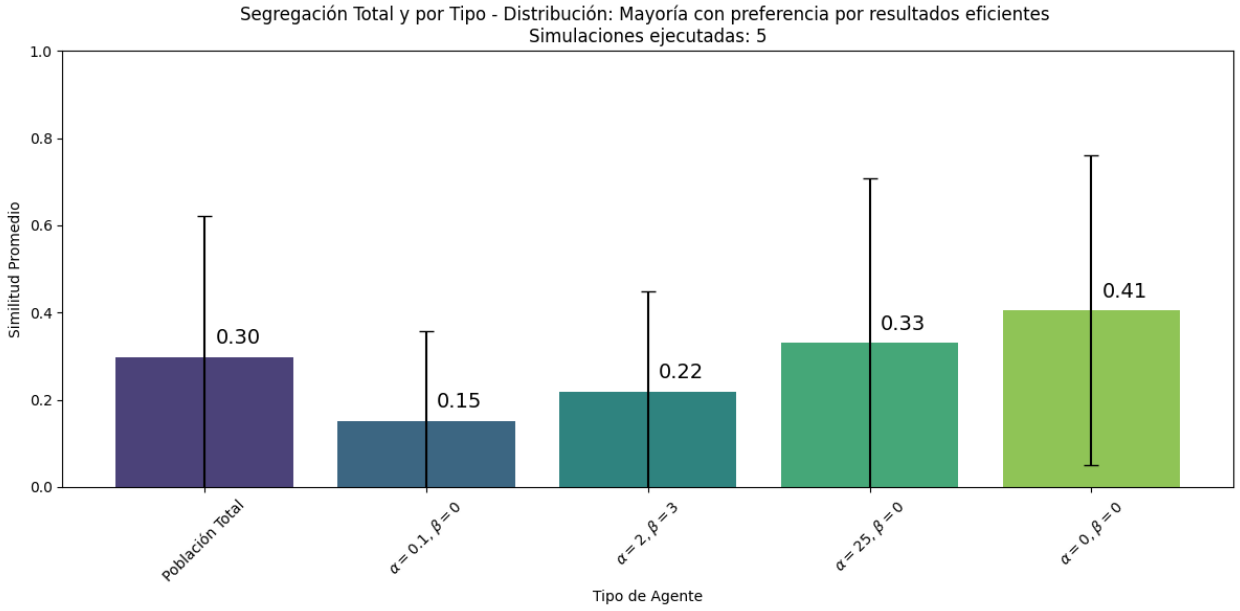

Figura 202: Segregación para una población de 150 rivales con una distribución de los tres rivales conocidos para los agentes RL y FEWA y el rival desconocido, donde el rival desconocido con preferencia por resultados eficientes tiene la proporción más baja de bots en la población, y los agentes restantes tiene la misma proporción.

9.7.2. Distribuciones para tamaño de muestra de 300 agentes

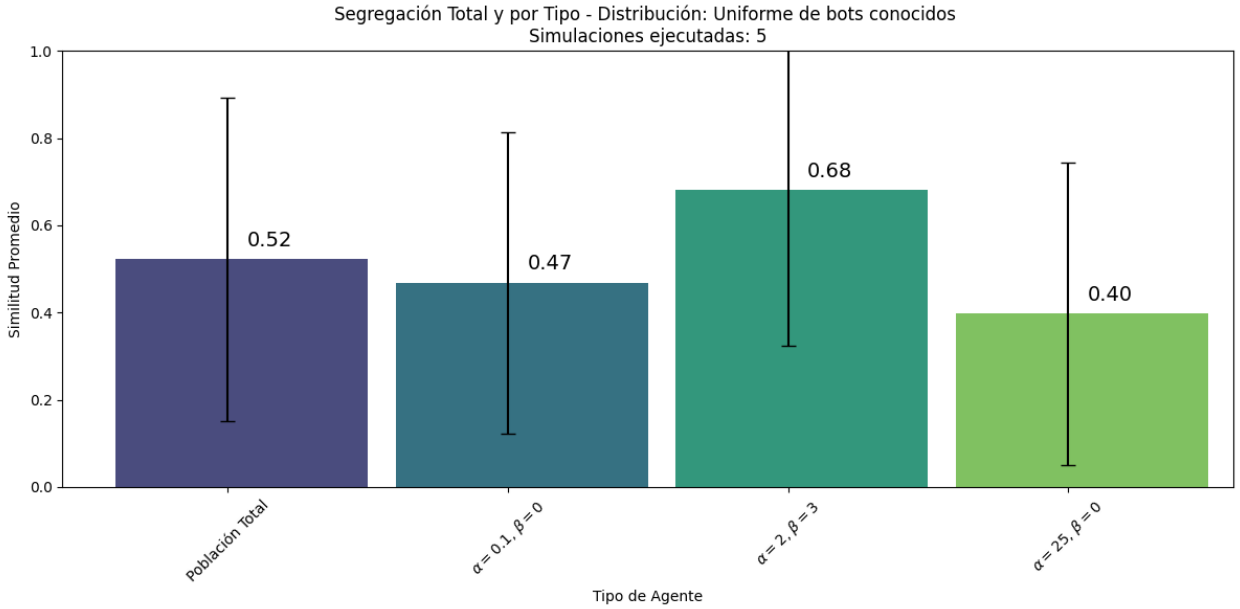

Figura 203: Segregación para una población de 300 rivales con una distribución uniforme de los tres rivales conocidos para los agentes RL y FEWA.

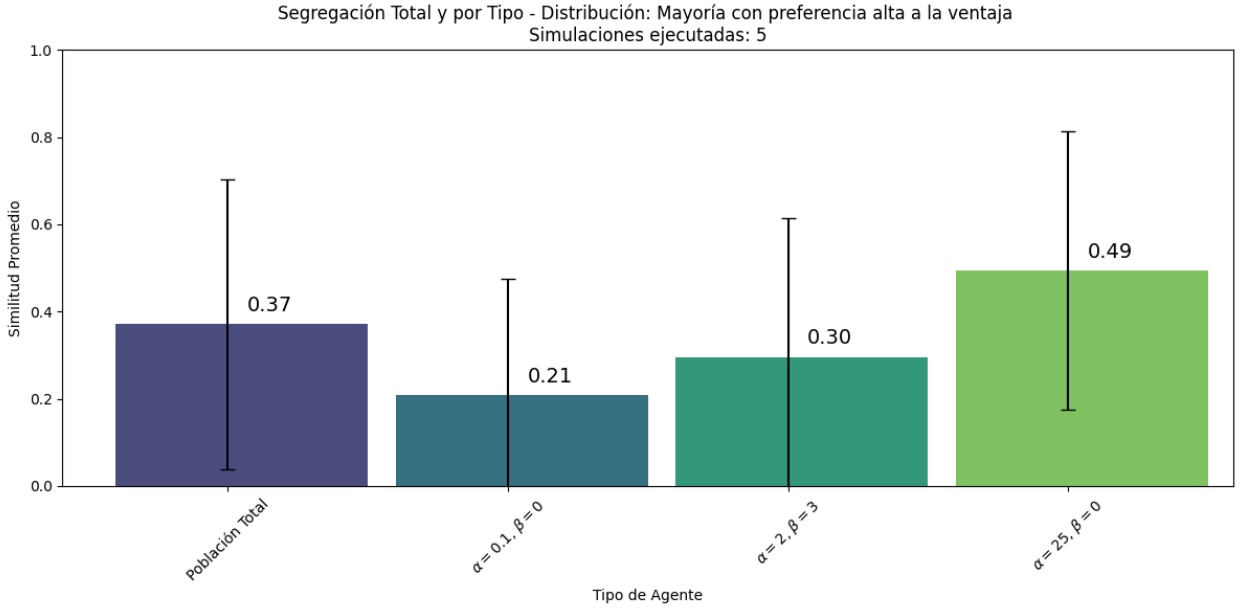

Figura 204: Segregación para una población de 300 rivales con una distribución de los tres rivales conocidos para los agentes RL y FEWA donde el rival con preferencia alta a resultados ventajosos tiene una proporción más alta de bots en la población y los agentes restantes tiene la misma proporción.

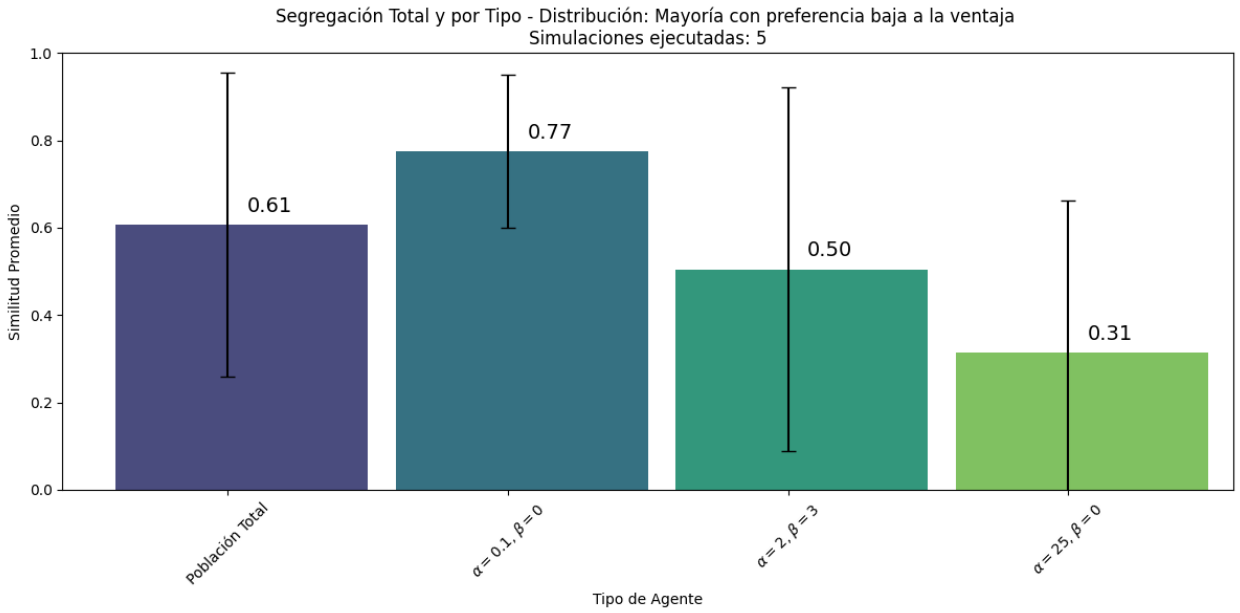

Figura 205: Segregación para una población de 300 rivales con una distribución de los tres rivales conocidos para los agentes RL y FEWA donde el rival con preferencia baja a resultados ventajosos tiene una proporción más alta de bots en la población y los agentes restantes tiene la misma proporción.

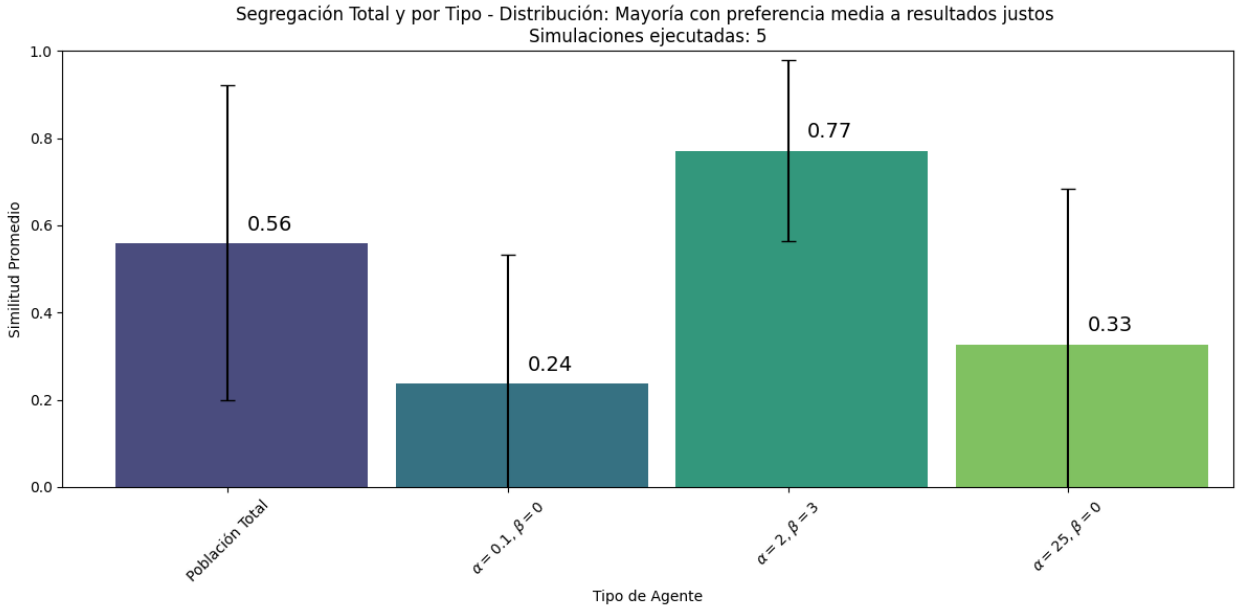

Figura 206: Segregación para una población de 300 rivales con una distribución de los tres rivales conocidos para los agentes RL y FEWA donde el rival con preferencia media a resultados justos tiene una proporción más alta de bots en la población y los agentes restantes tiene la misma proporción..

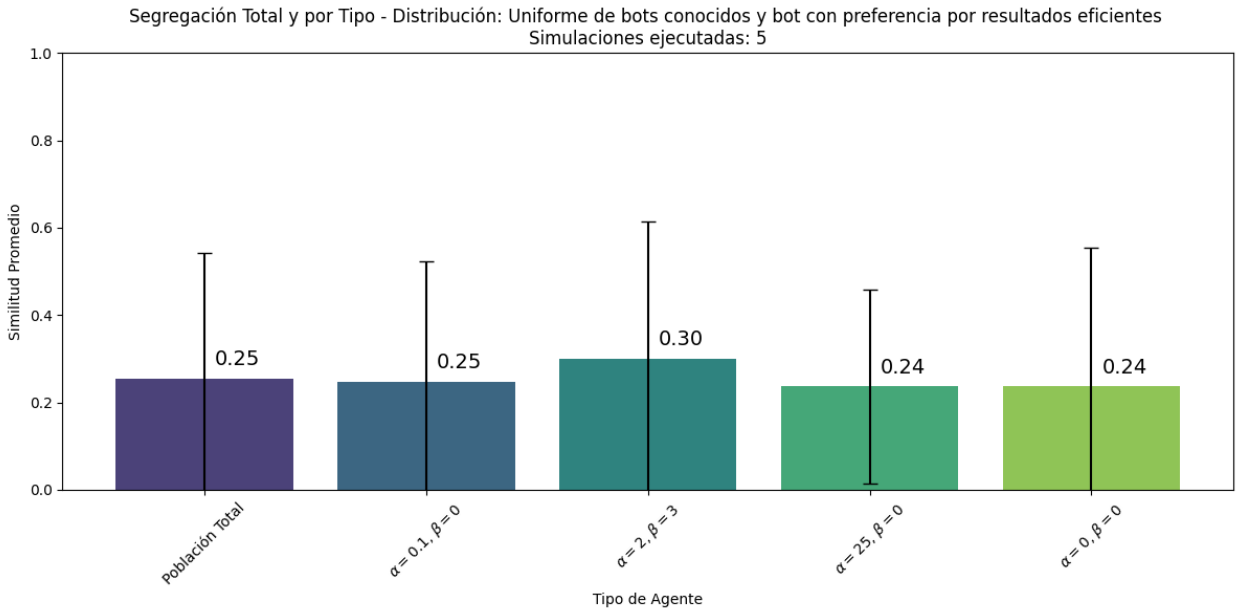

Figura 207: Segregación para una población de 300 rivales con una distribución uniforme de los tres rivales conocidos para los agentes RL y FEWA y de un rival desconocido, el rival con preferencia por resultados eficientes.

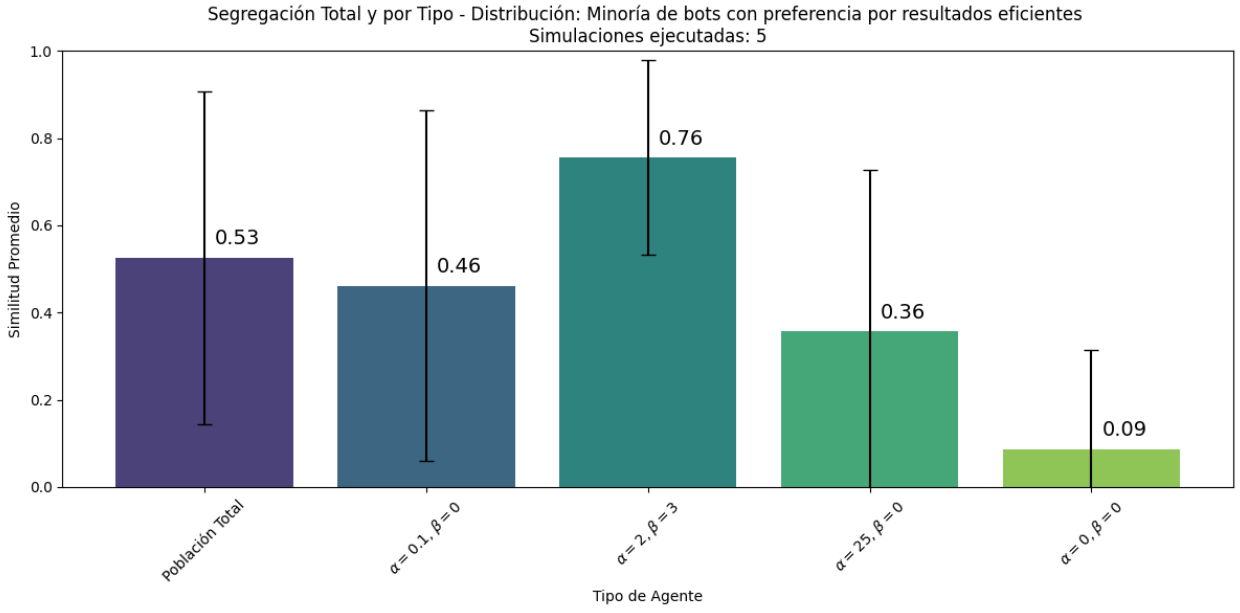

Figura 208: Segregación para una población de 300 rivales con una distribución de los tres rivales conocidos para los agentes RL y FEWA y el rival desconocido, donde el rival desconocido con preferencia por resultados eficientes tiene la proporción más baja de bots en la población, y los agentes restantes tiene la misma proporción.

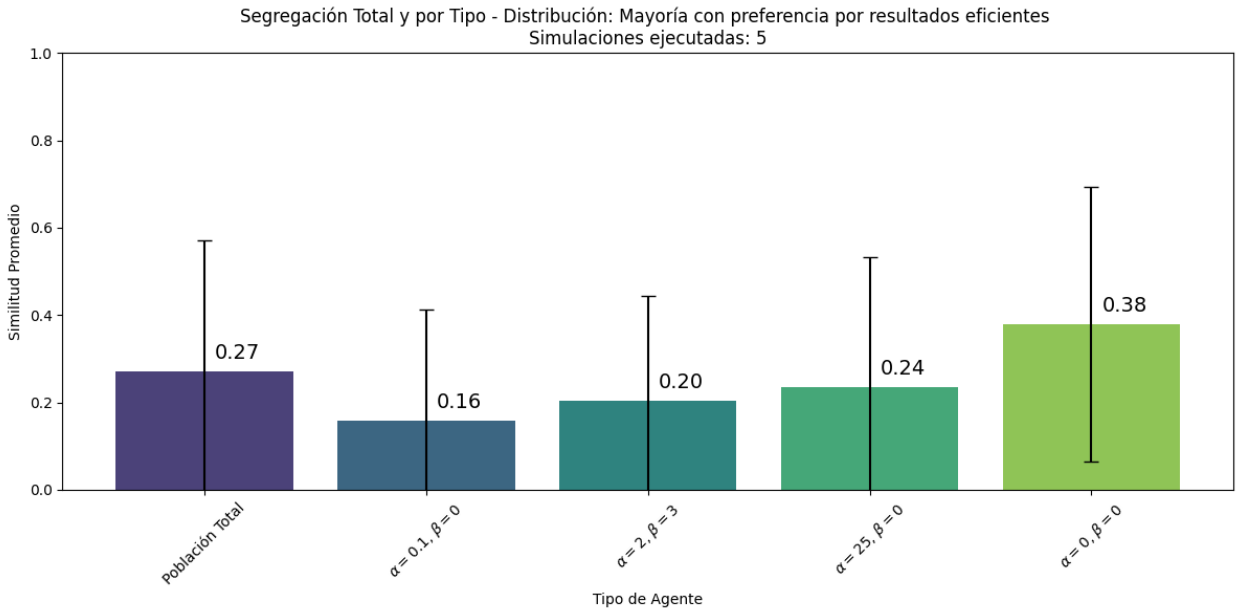

Figura 209: Segregación para una población de 300 rivales con una distribución de los tres rivales conocidos para los agentes RL y FEWA y el rival desconocido, donde el rival desconocido con preferencia por resultados eficientes tiene la proporción más baja de bots en la población, y los agentes restantes tiene la misma proporción.

9.7.3. Distribuciones para tamaño de muestra de 50 agentes

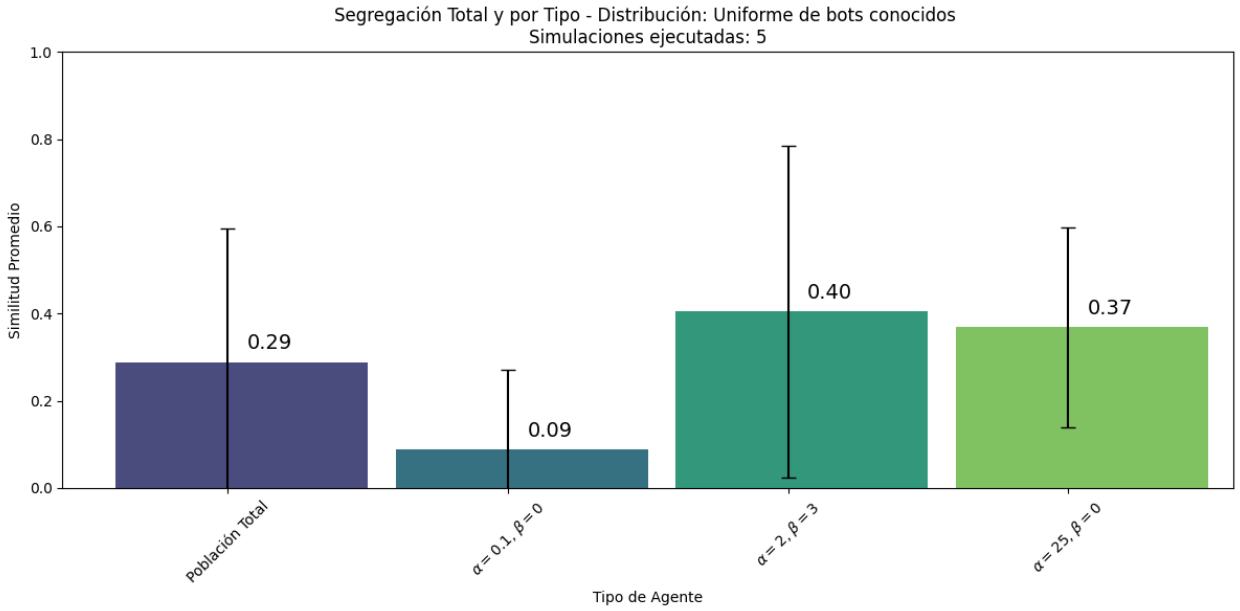

Figura 210: Segregación para una población de 50 rivales con una distribución uniforme de los tres rivales conocidos para los agentes RL y FEWA.

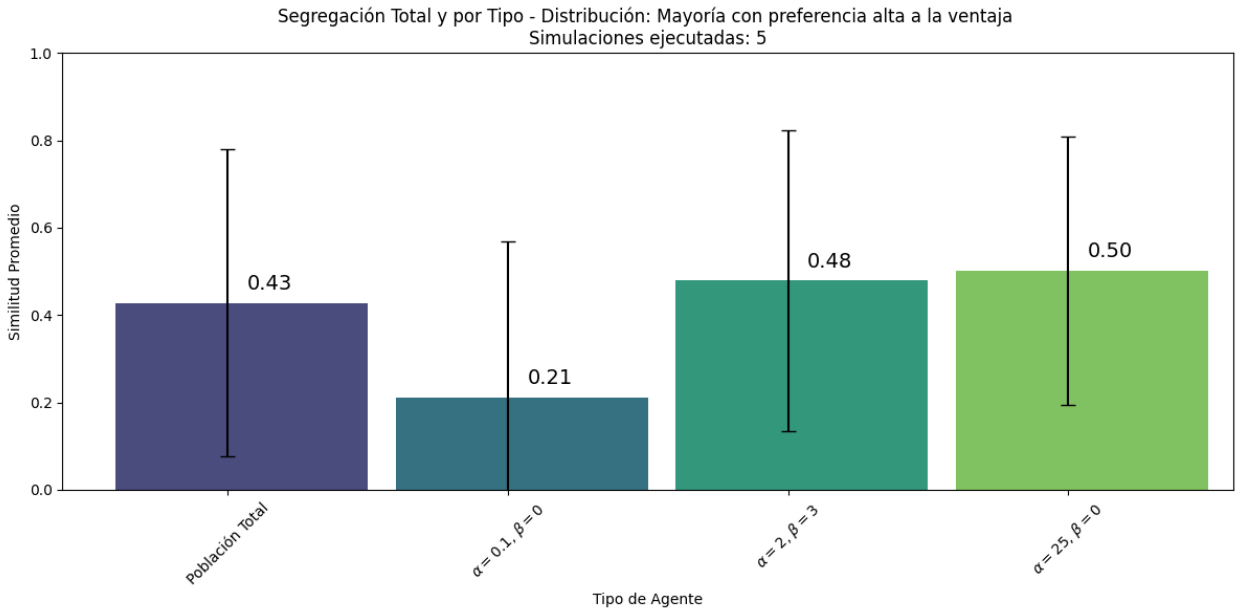

Figura 211: Segregación para una población de 50 rivales con una distribución de los tres rivales conocidos para los agentes RL y FEWA donde el rival con preferencia alta a resultados ventajosos tiene una proporción más alta de bots en la población y los agentes restantes tiene la misma proporción.

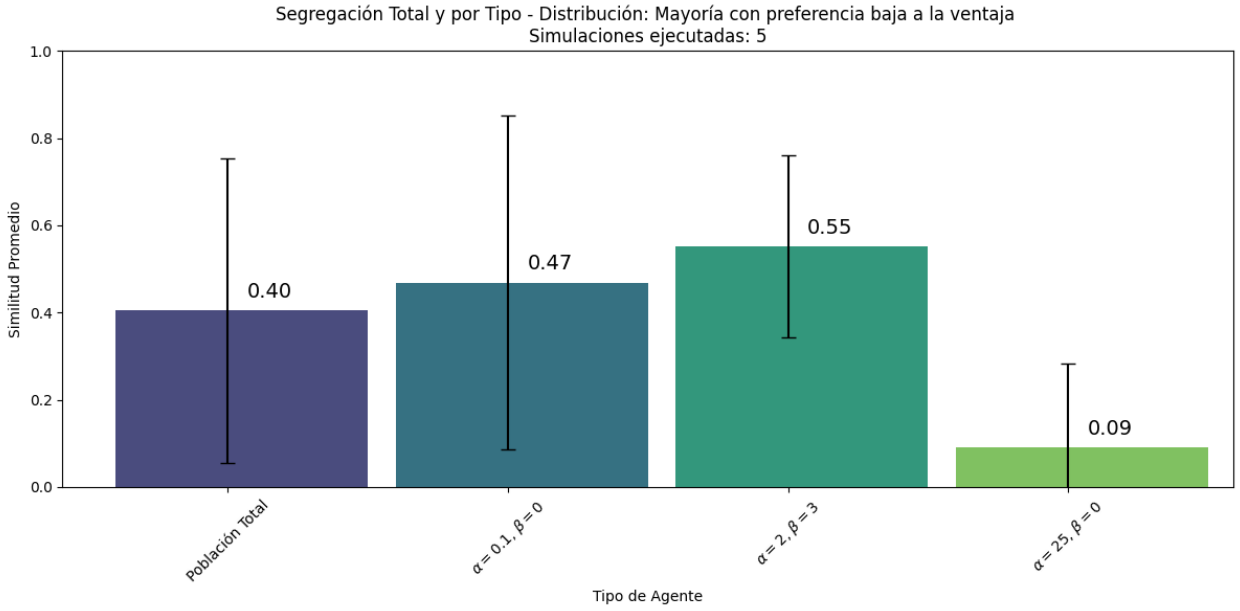

Figura 212: Segregación para una población de 50 rivales con una distribución de los tres rivales conocidos para los agentes RL y FEWA donde el rival con preferencia baja a resultados ventajosos tiene una proporción más alta de bots en la población y los agentes restantes tiene la misma proporción.

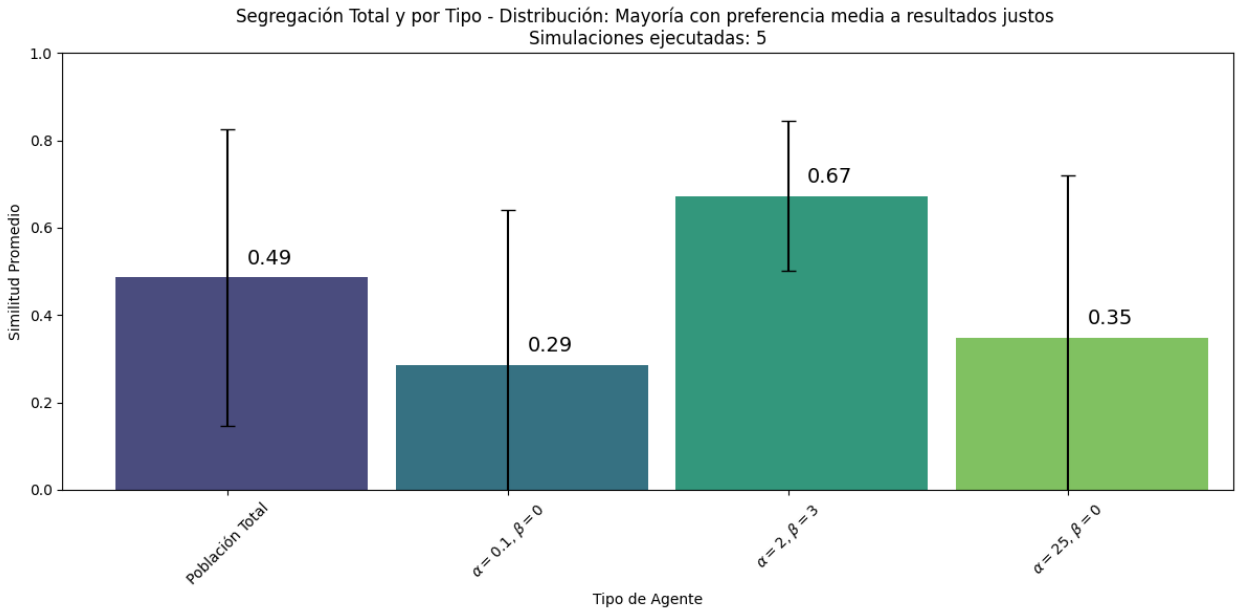

Figura 213: Segregación para una población de 50 rivales con una distribución de los tres rivales conocidos para los agentes RL y FEWA donde el rival con preferencia media a resultados justos tiene una proporción más alta de bots en la población y los agentes restantes tiene la misma proporción..

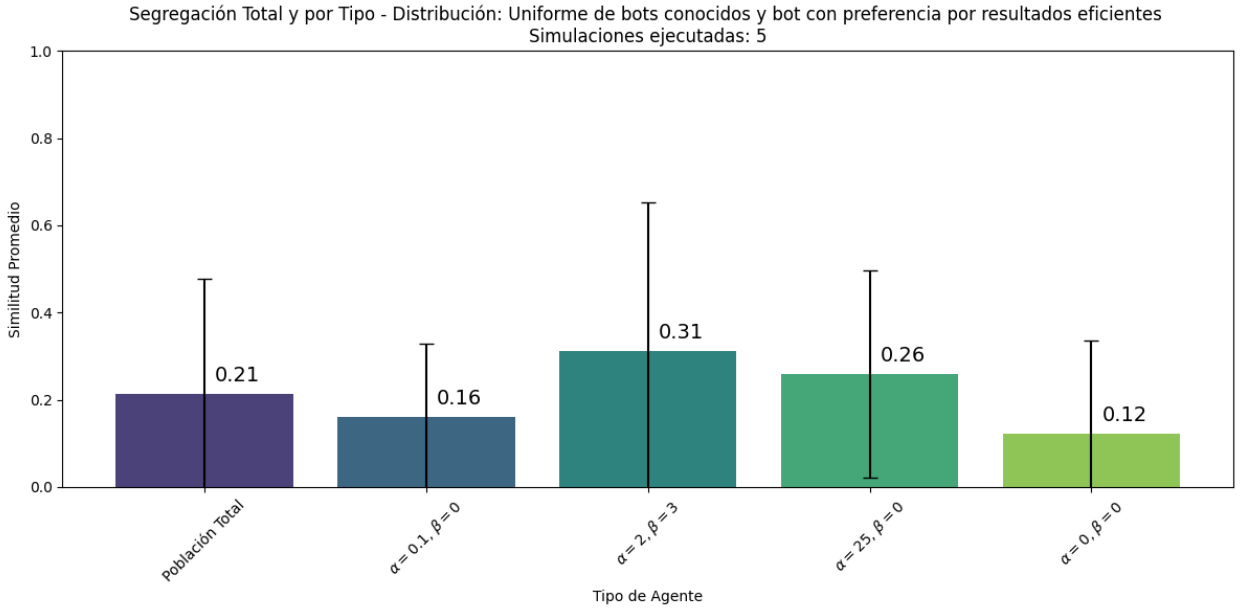

Figura 214: Segregación para una población de 50 rivales con una distribución uniforme de los tres rivales conocidos para los agentes RL y FEWA y de un rival desconocido, el rival con preferencia por resultados eficientes.

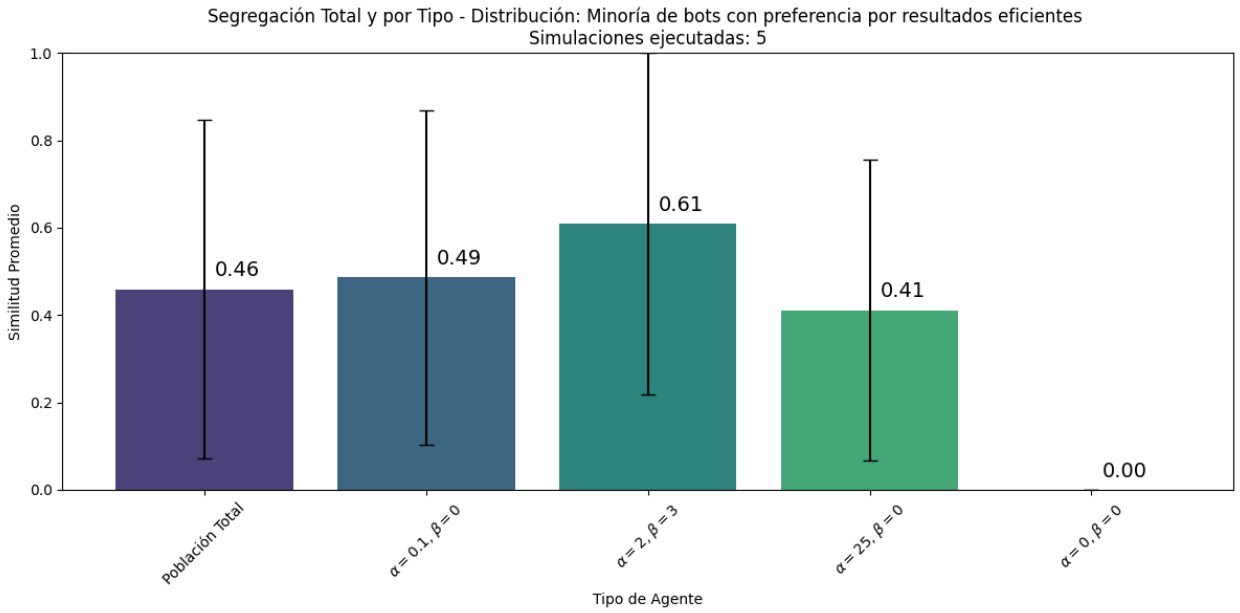

Figura 215: Segregación para una población de 50 rivales con una distribución de los tres rivales conocidos para los agentes RL y FEWA y el rival desconocido, donde el rival desconocido con preferencia por resultados eficientes tiene la proporción más baja de bots en la población, y los agentes restantes tiene la misma proporción.

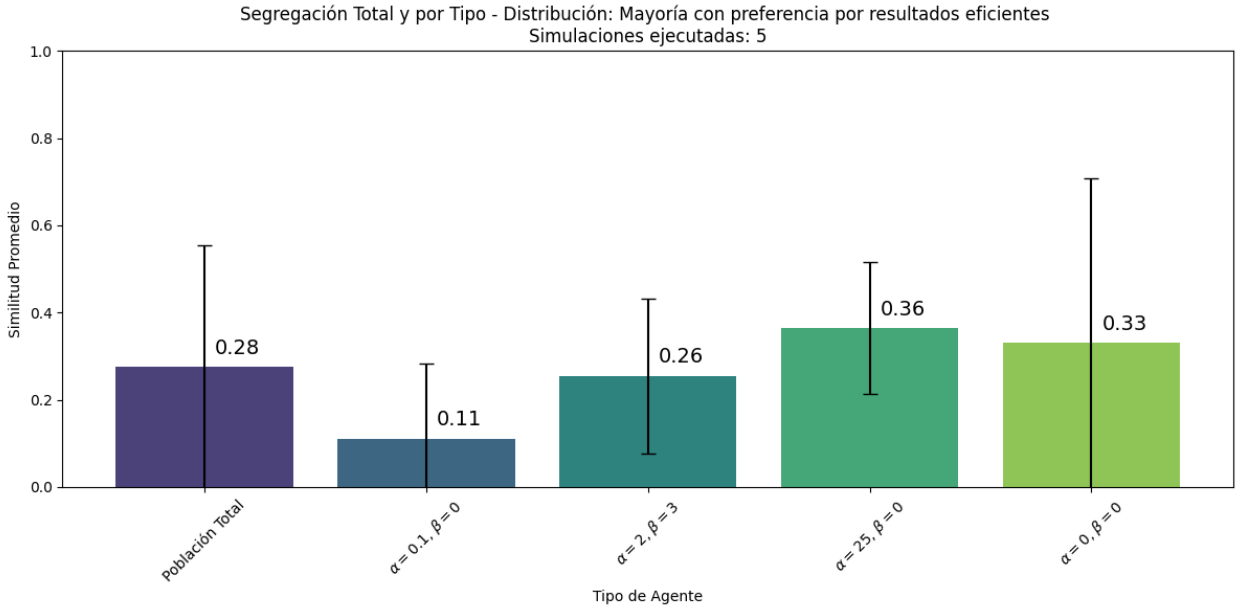

Figura 216: Segregación para una población de 50 rivales con una distribución de los tres rivales conocidos para los agentes RL y FEWA y el rival desconocido, donde el rival desconocido con preferencia por resultados eficientes tiene la proporción más baja de bots en la población, y los agentes restantes tiene la misma proporción.

9.8. Gráficos de Equilibrio

9.8.1. Distribuciones para tamaño de muestra de 300 agentes

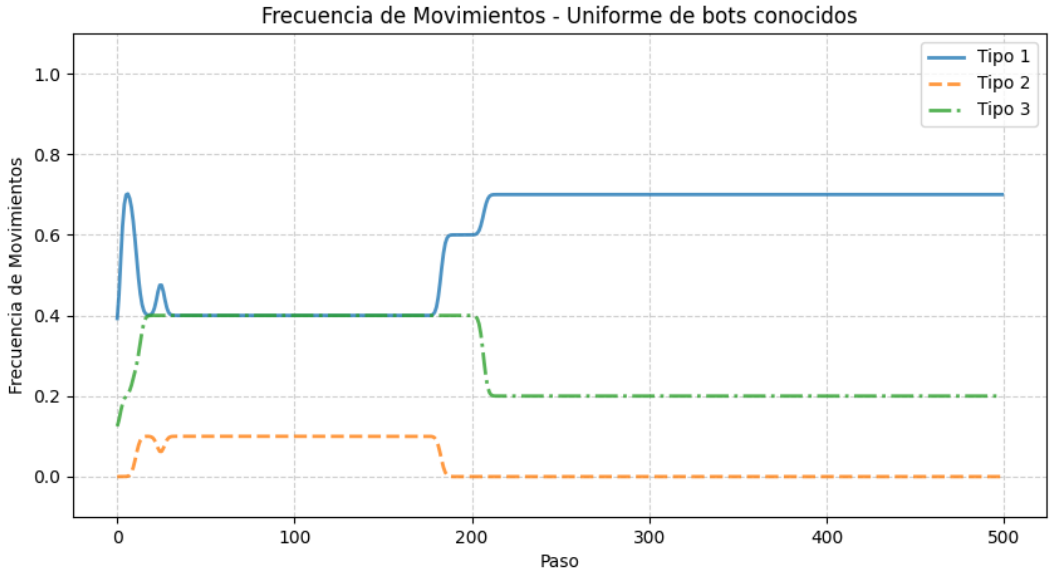

Figura 217: Movimientos de cada grupo cuando la población es de 300 agentes y la distribución es uniforme de bots conocidos.

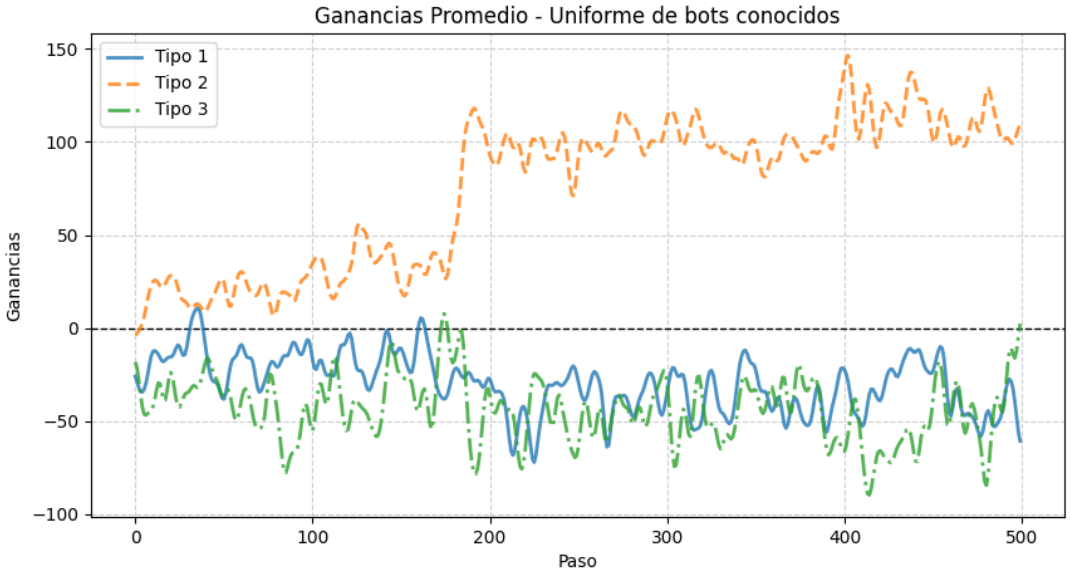

Figura 218: Ganancias de cada grupo cuando la población es de 300 agentes y la distribución es uniforme de rivales conocidos.

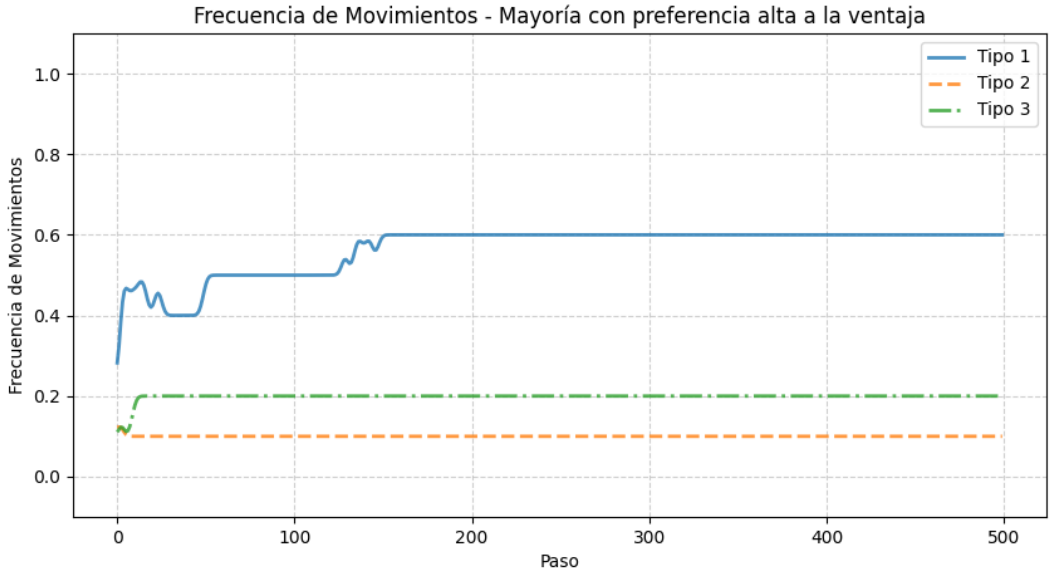

Figura 219: Movimientos de cada grupo cuando la población es de 300 agentes y la distribución es de rivales conocidos pero el rival con preferencia alta a la ventaja tiene mayoría.

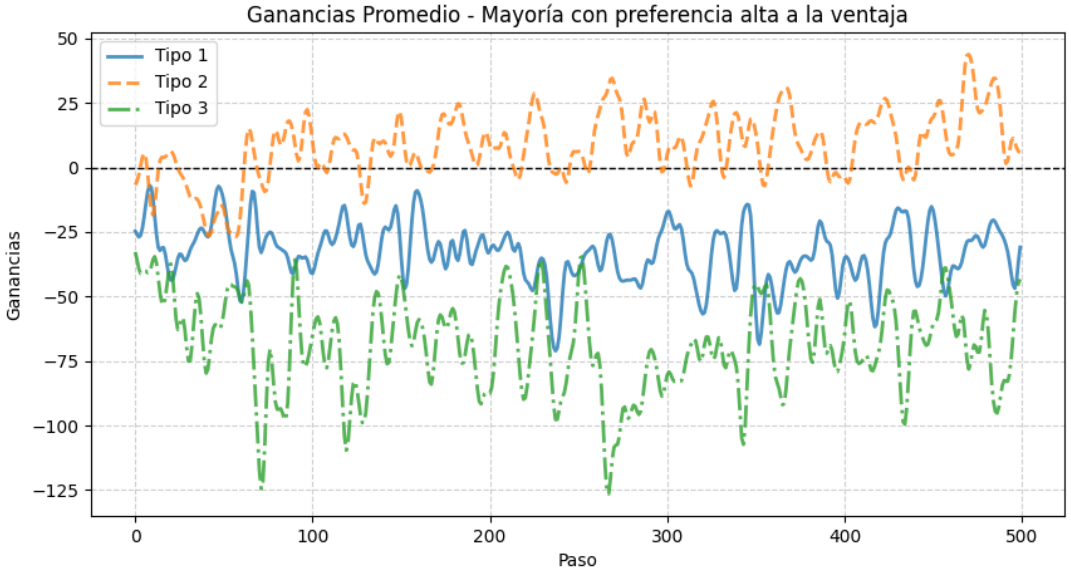

Figura 220: Ganancias de cada grupo cuando la población es de 300 agentes y la distribución es de rivales conocidos pero el rival con preferencia alta a la ventaja tiene mayoría.

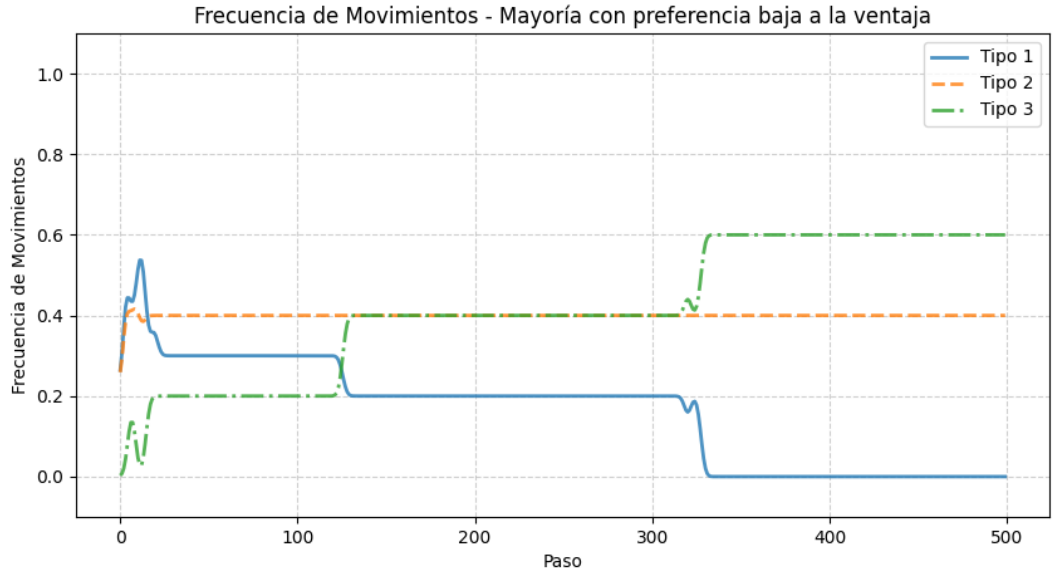

Figura 221: Movimientos de cada grupo cuando la población es de 300 agentes y la distribución es de rivales conocidos pero el rival con preferencia baja a la ventaja tiene mayoría.

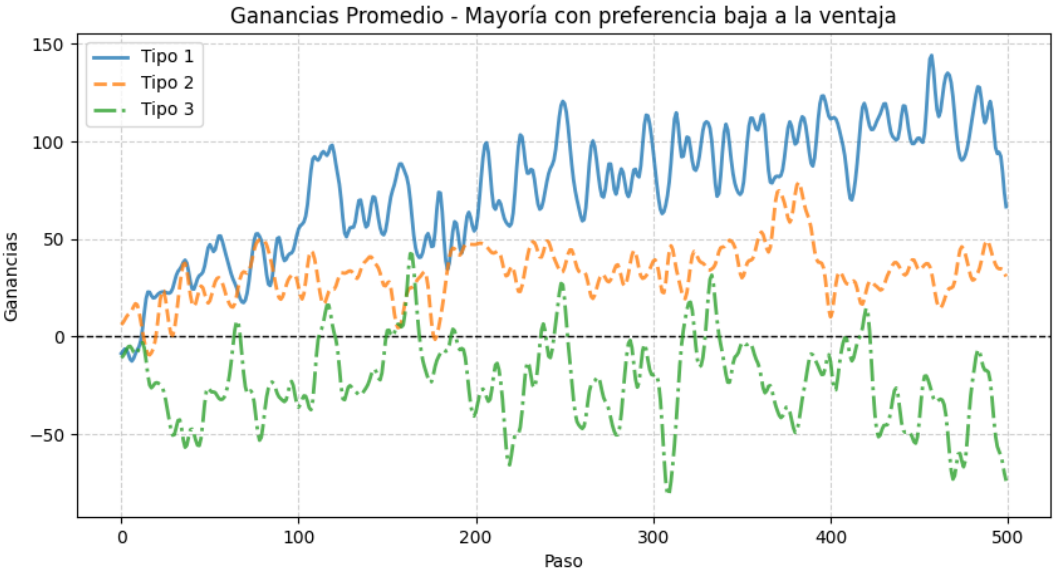

Figura 222: Ganancias de cada grupo cuando la población es de 300 agentes y la distribución es de rivales conocidos pero el rival con preferencia baja a la ventaja tiene mayoría.

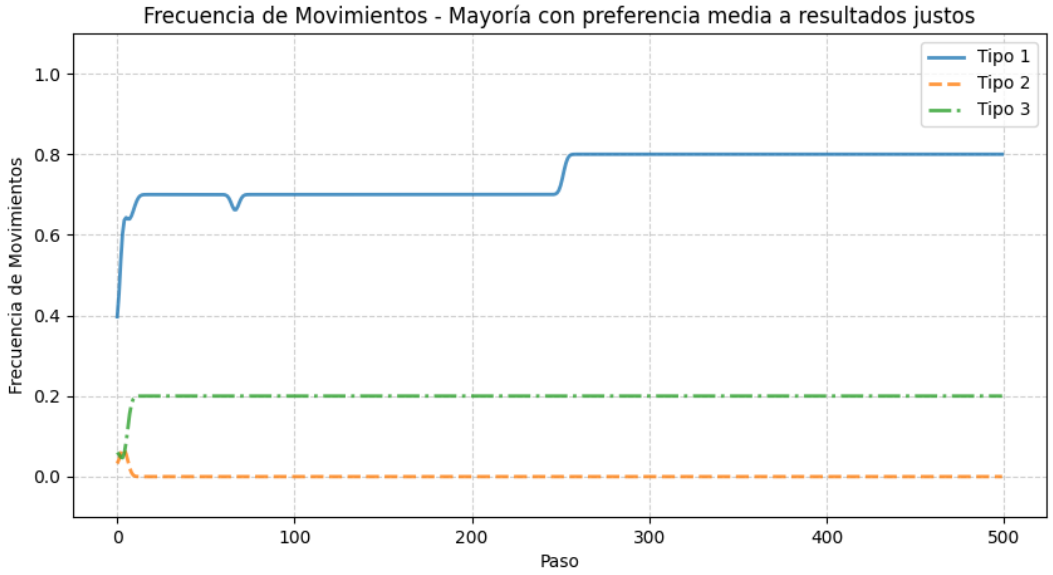

Figura 223: Movimientos de cada grupo cuando la población es de 300 agentes y la distribución es de rivales conocidos pero el rival con preferencia media a resultados justos tiene mayoría.

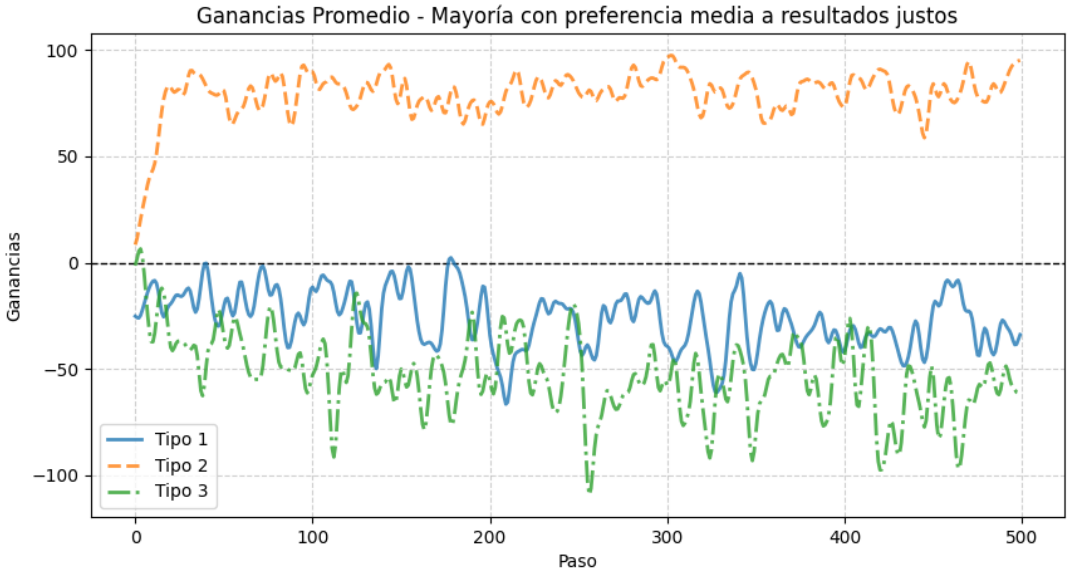

Figura 224: Ganancias de cada grupo cuando la población es de 300 agentes y la distribución es de rivales conocidos pero el rival con preferencia media a resultados justos tiene mayoría.

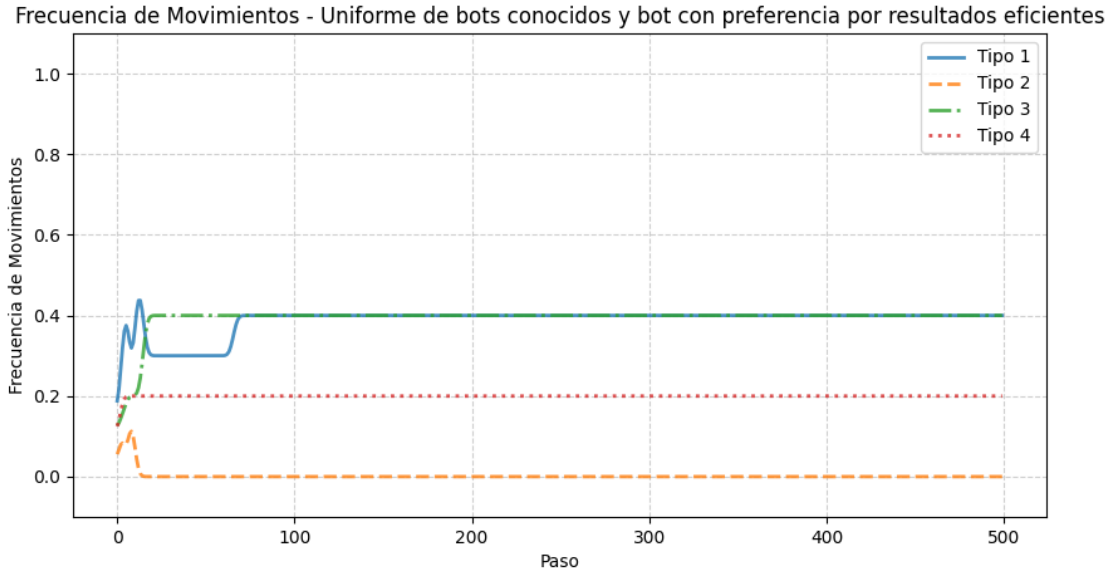

Figura 225: Movimientos de cada grupo cuando la población es de 300 agentes y la distribución es uniforme de rivales conocidos y un rival desconocido con preferencia por resultados eficientes.

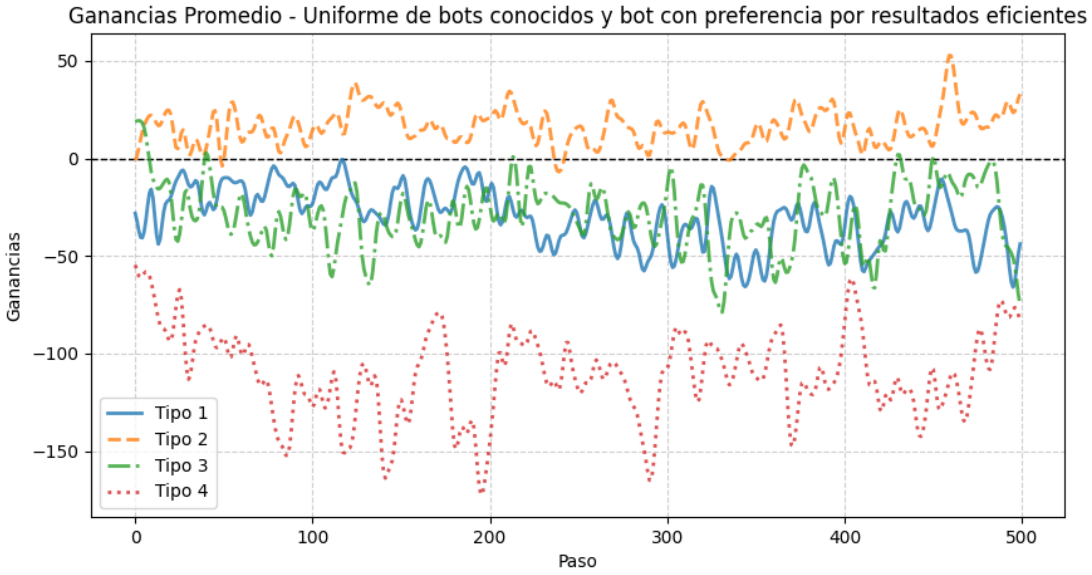

Figura 226: Ganancias de cada grupo cuando la población es de 300 agentes y la distribución es uniforme de rivales conocidos y un rival desconocido con preferencia por resultados eficientes.

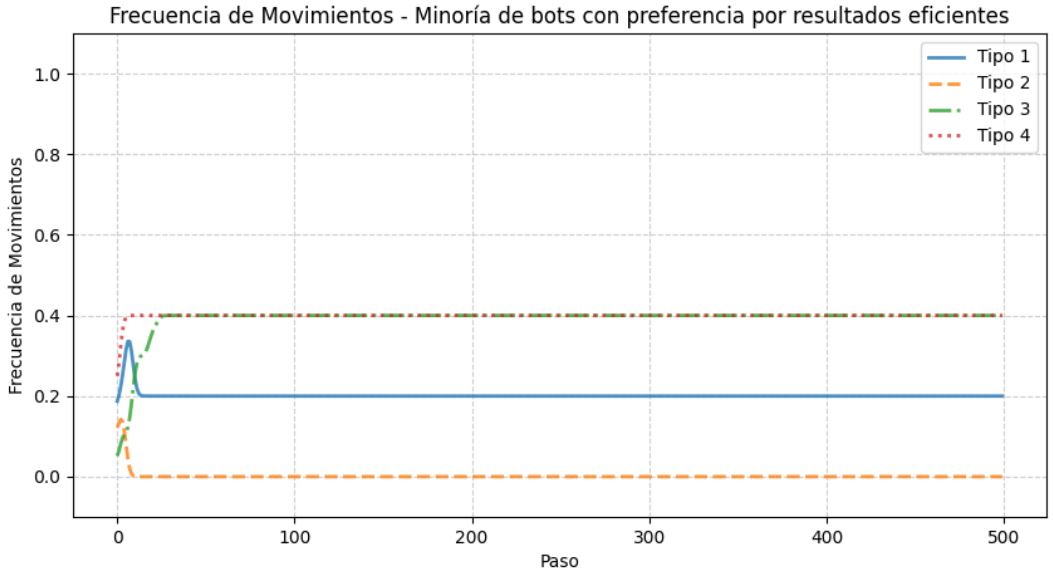

Figura 227: Movimientos de cada grupo cuando la población es de 300 agentes y la distribución de rivales conocidos y una minoría de rivales desconocido con preferencia por resultados eficientes.

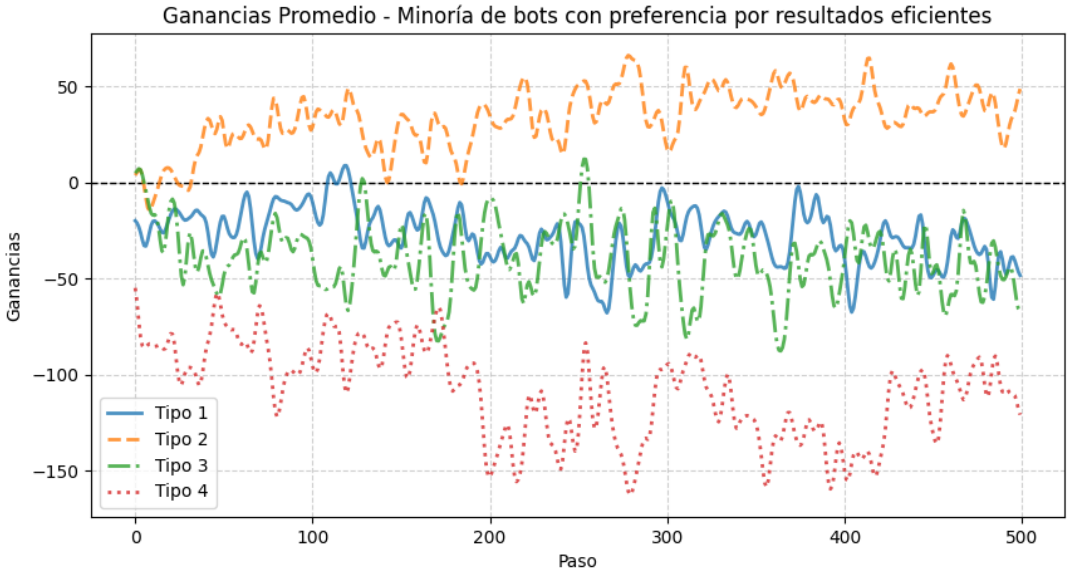

Figura 228: Ganancias de cada grupo cuando la población es de 300 agentes y la distribución de rivales conocidos y una minoría de rivales desconocido con preferencia por resultados eficientes.

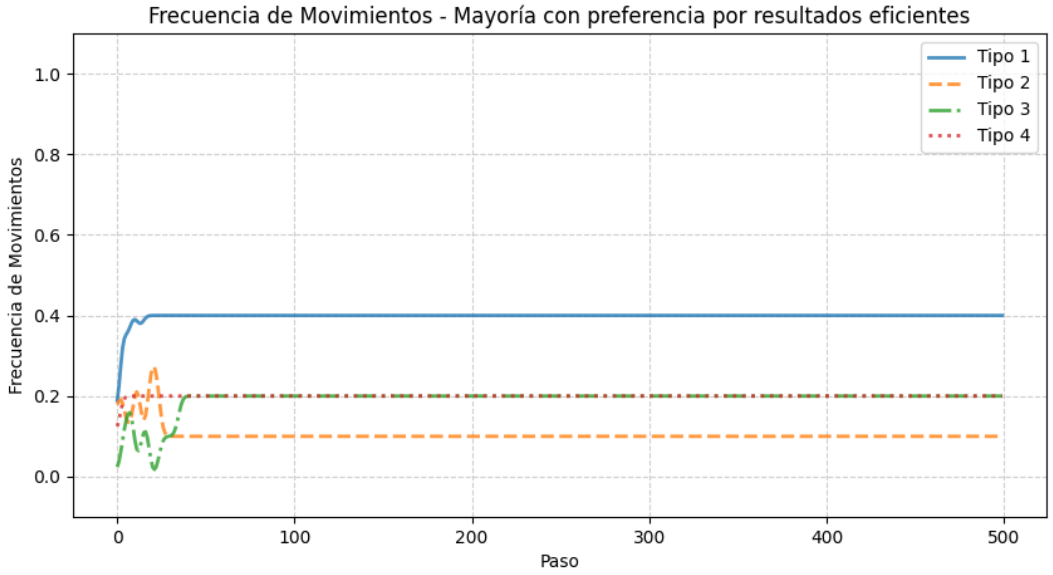

Figura 229: Movimientos de cada grupo cuando la población es de 300 agentes y la distribución de rivales conocidos y una mayoría de rivales desconocido con preferencia por resultados eficientes.

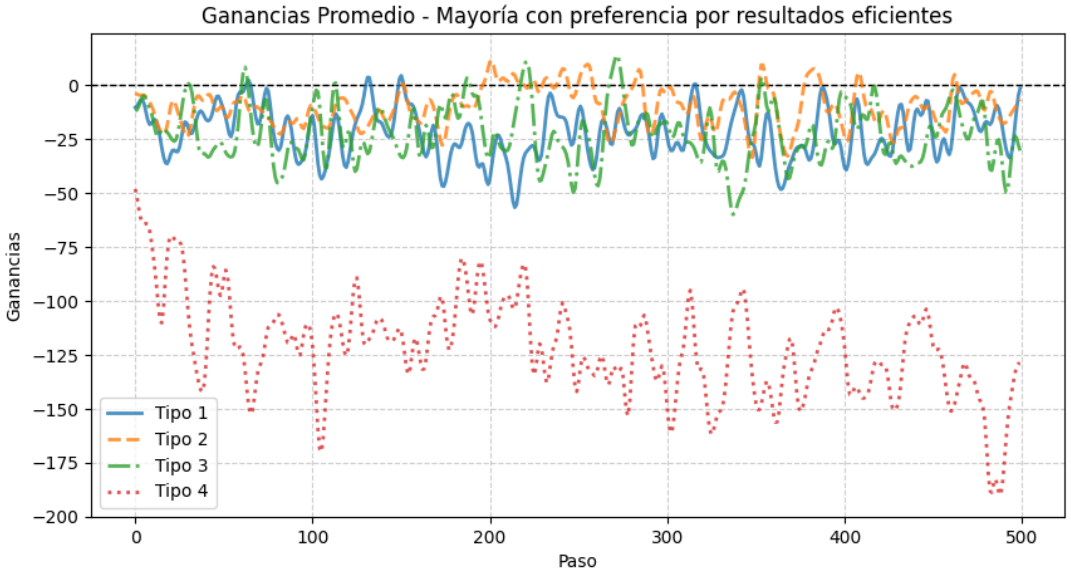

Figura 230: Ganancias de cada grupo cuando la población es de 300 agentes y la distribución de rivales conocidos y una mayoría de rivales desconocido con preferencia por resultados eficientes.

9.8.2. Distribuciones para tamaño de muestra de 50 agentes

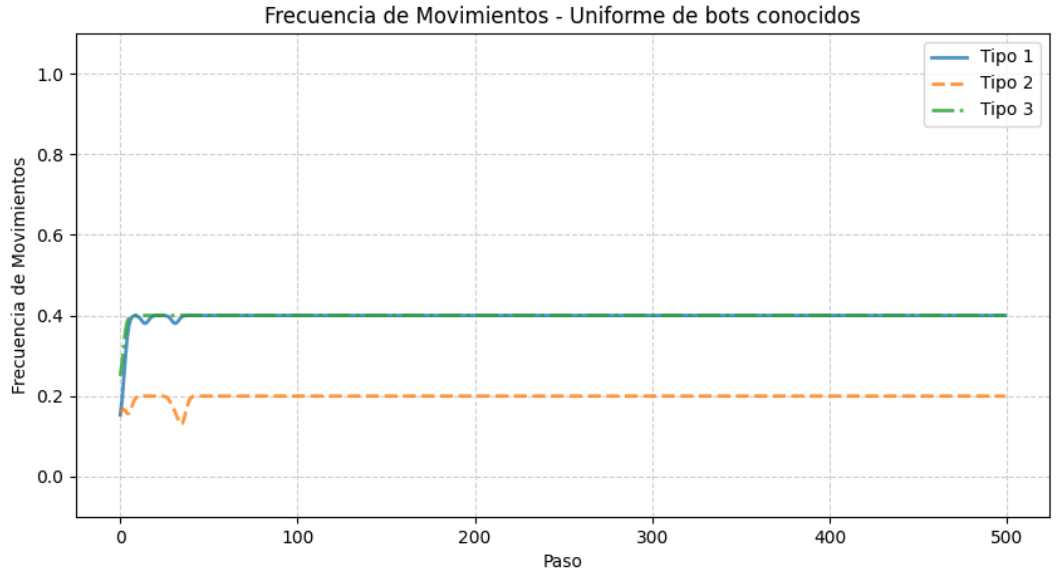

Figura 231: Movimientos de cada grupo cuando la población es de 50 agentes y la distribución es uniforme de bots conocidos.

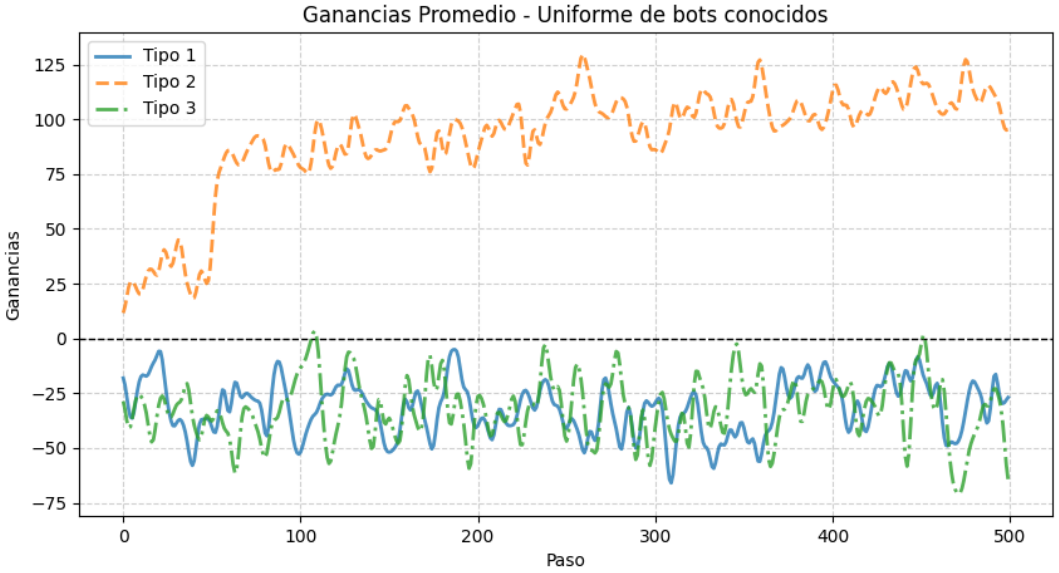

Figura 232: Ganancias de cada grupo cuando la población es de 50 agentes y la distribución es uniforme de rivales conocidos.

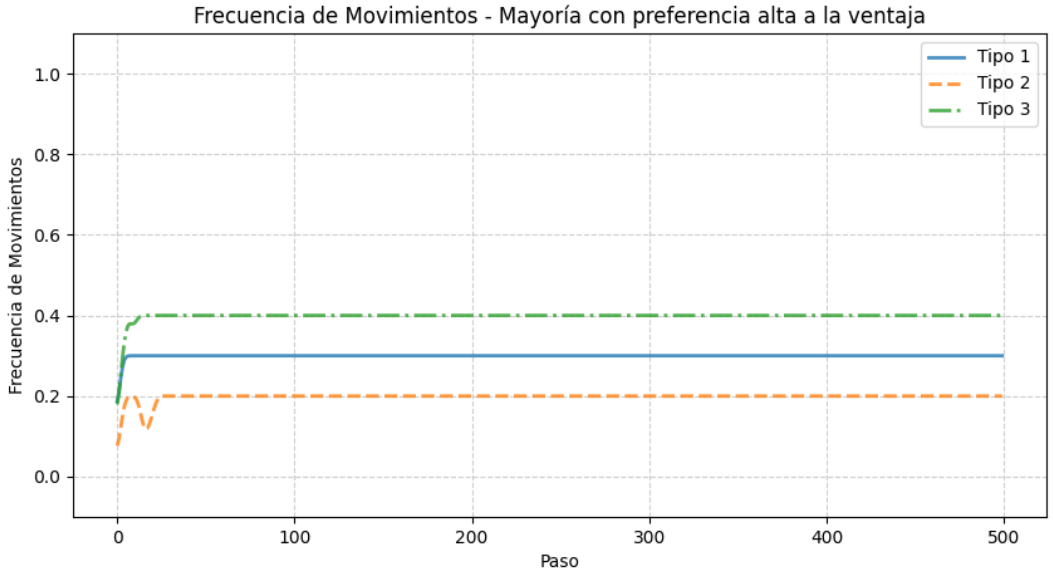

Figura 233: Movimientos de cada grupo cuando la población es de 50 agentes y la distribución es de rivales conocidos pero el rival con preferencia alta a la ventaja tiene mayoría.

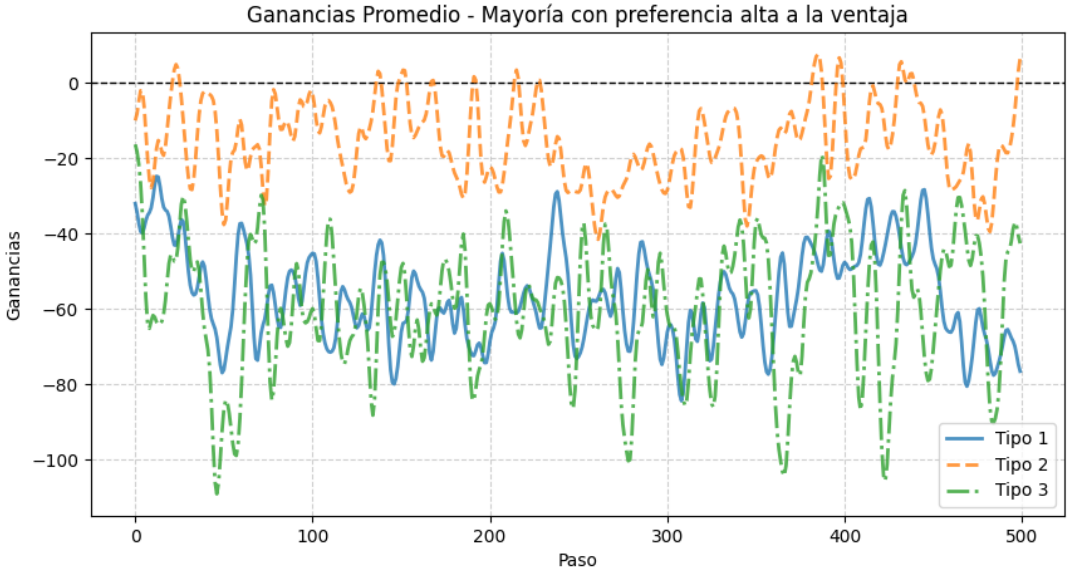

Figura 234: Ganancias de cada grupo cuando la población es de 50 agentes y la distribución es de rivales conocidos pero el rival con preferencia alta a la ventaja tiene mayoría.

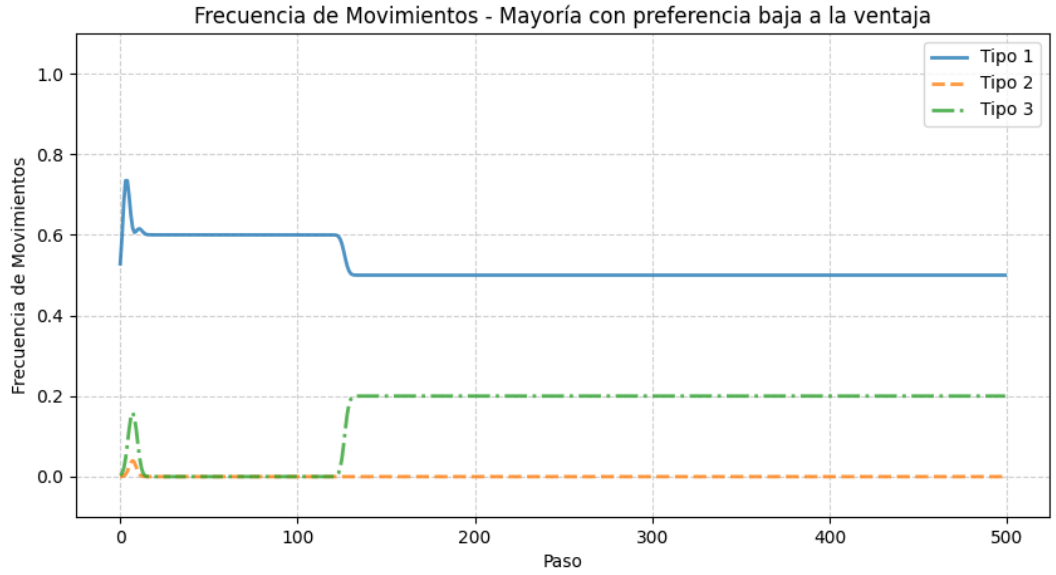

Figura 235: Movimientos de cada grupo cuando la población es de 50 agentes y la distribución es de rivales conocidos pero el rival con preferencia baja a la ventaja tiene mayoría.

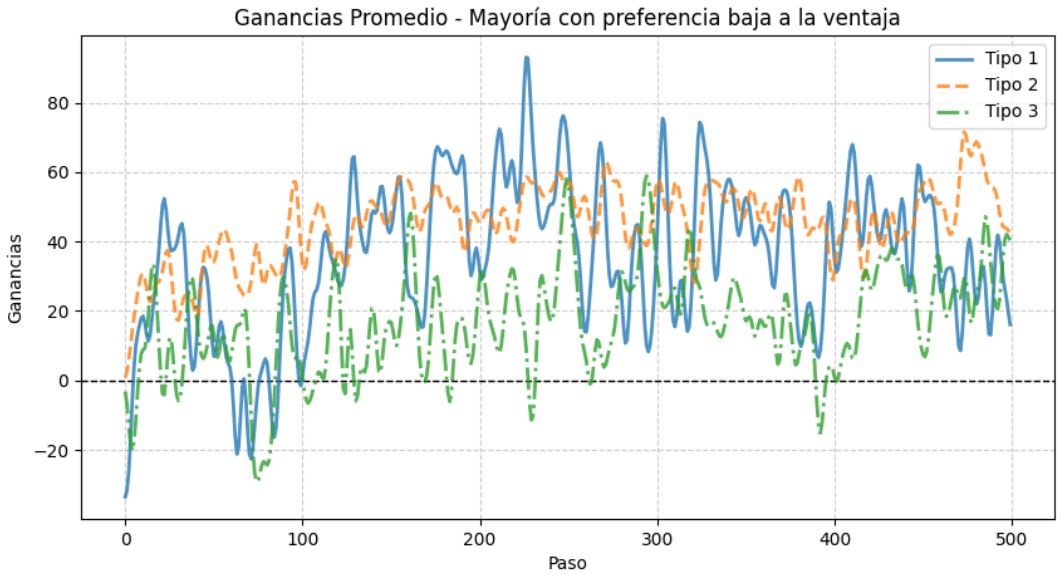

Figura 236: Ganancias de cada grupo cuando la población es de 50 agentes y la distribución es de rivales conocidos pero el rival con preferencia baja a la ventaja tiene mayoría.

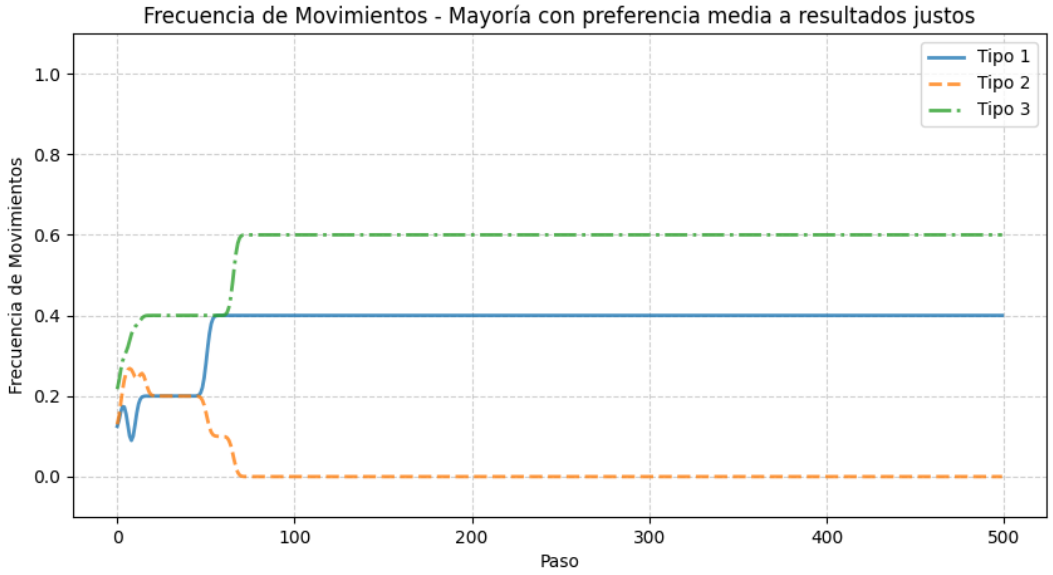

Figura 237: Movimientos de cada grupo cuando la población es de 50 agentes y la distribución es de rivales conocidos pero el rival con preferencia media a resultados justos tiene mayoría.

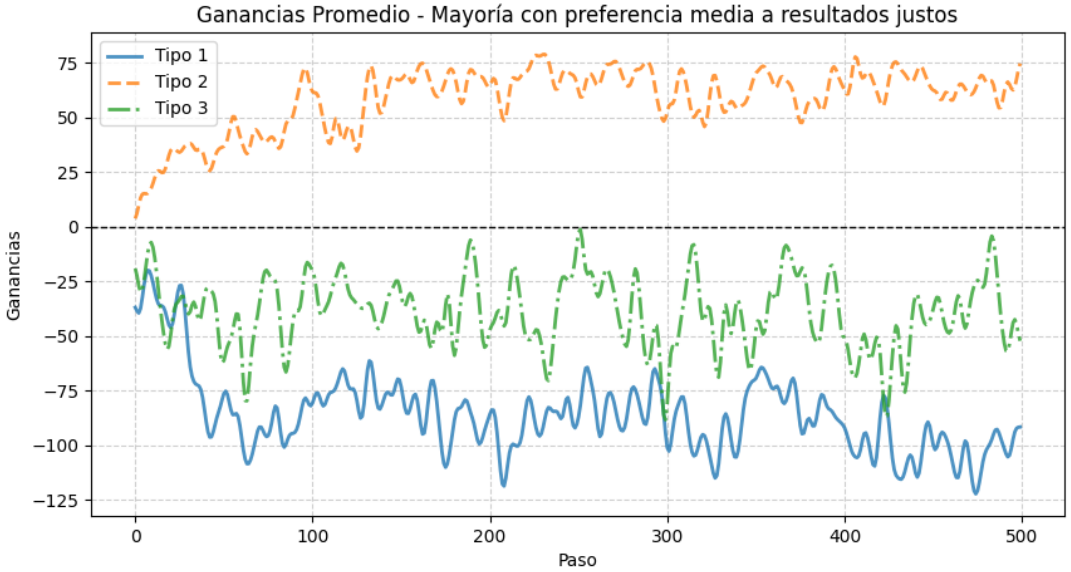

Figura 238: Ganancias de cada grupo cuando la población es de 50 agentes y la distribución es de rivales conocidos pero el rival con preferencia media a resultados justos tiene mayoría.

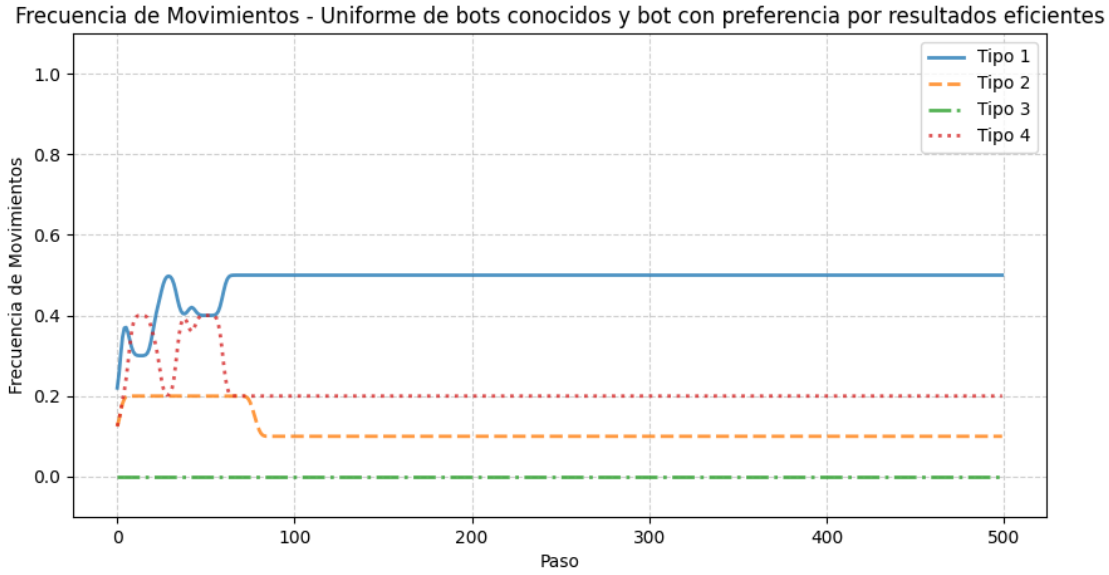

Figura 239: Movimientos de cada grupo cuando la población es de 50 agentes y la distribución es uniforme de rivales conocidos y un rival desconocido con preferencia por resultados eficientes.

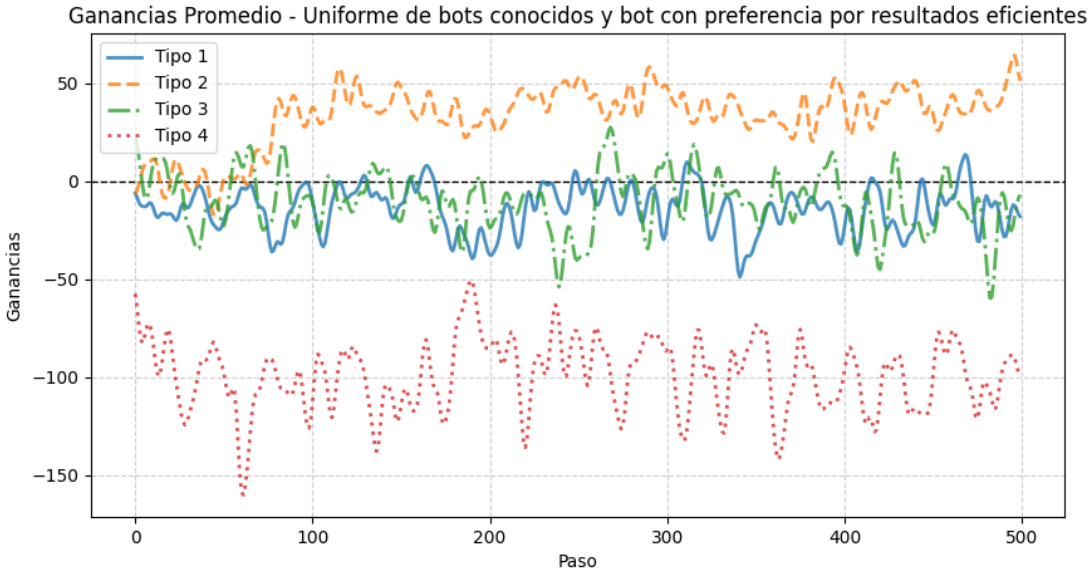

Figura 240: Ganancias de cada grupo cuando la población es de 50 agentes y la distribución es uniforme de rivales conocidos y un rival desconocido con preferencia por resultados eficientes.

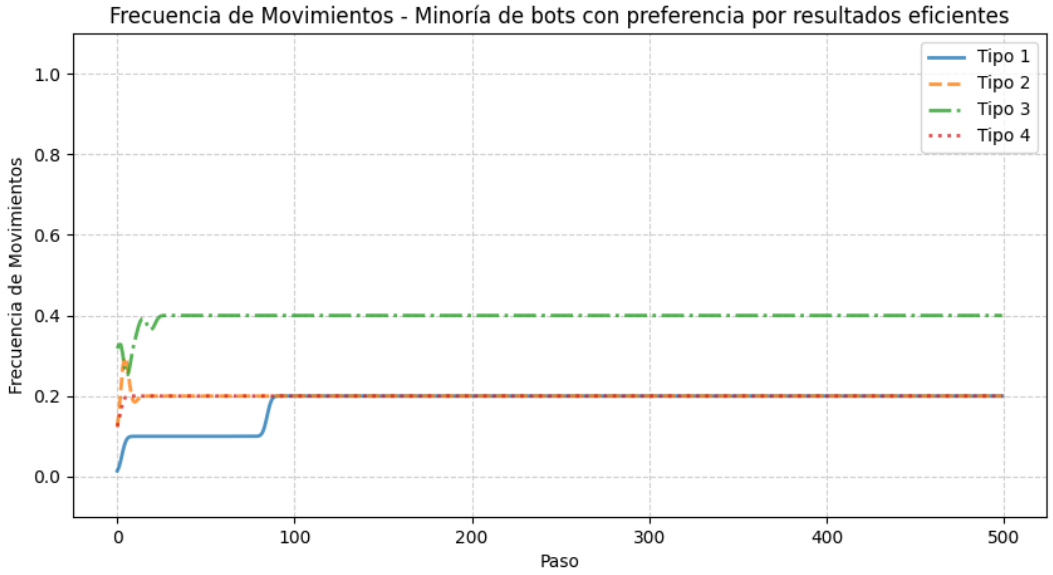

Figura 241: Movimientos de cada grupo cuando la población es de 50 agentes y la distribución de rivales conocidos y una minoría de rivales desconocido con preferencia por resultados eficientes.

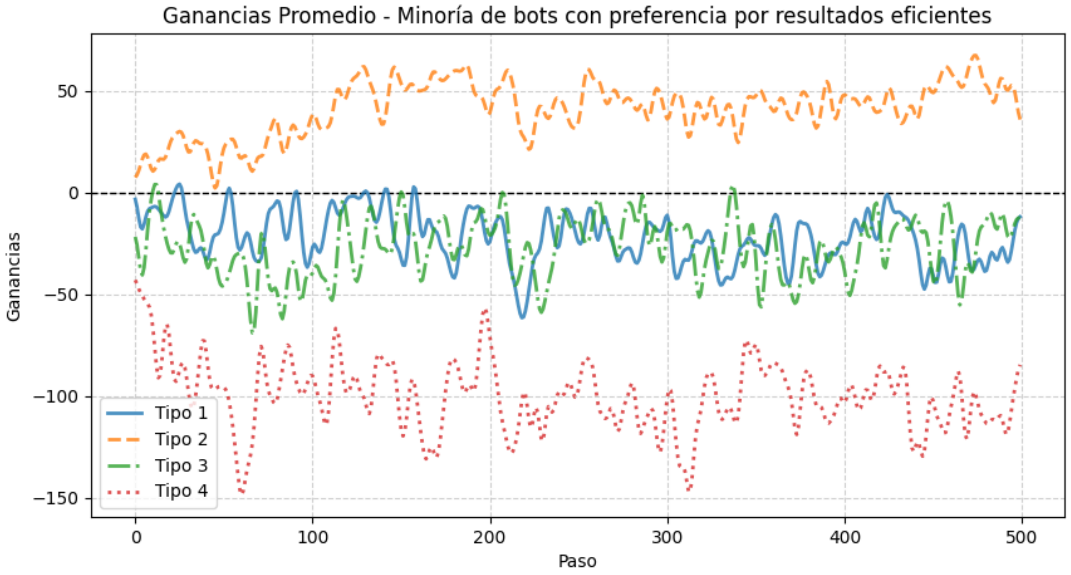

Figura 242: Ganancias de cada grupo cuando la población es de 50 agentes y la distribución de rivales conocidos y una minoría de rivales desconocido con preferencia por resultados eficientes.

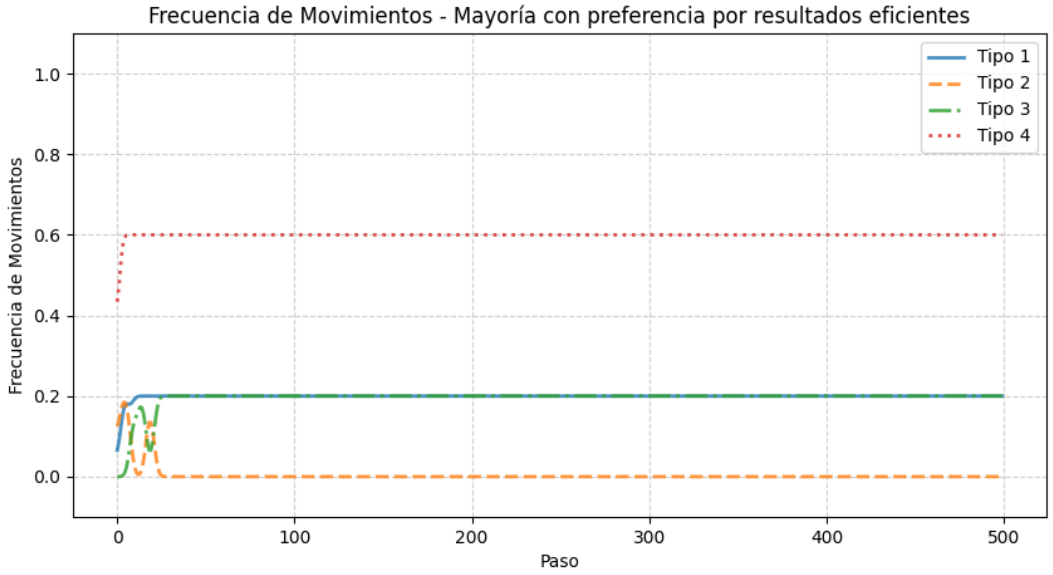

Figura 243: Movimientos de cada grupo cuando la población es de 50 agentes y la distribución de rivales conocidos y una mayoría de rivales desconocido con preferencia por resultados eficientes.

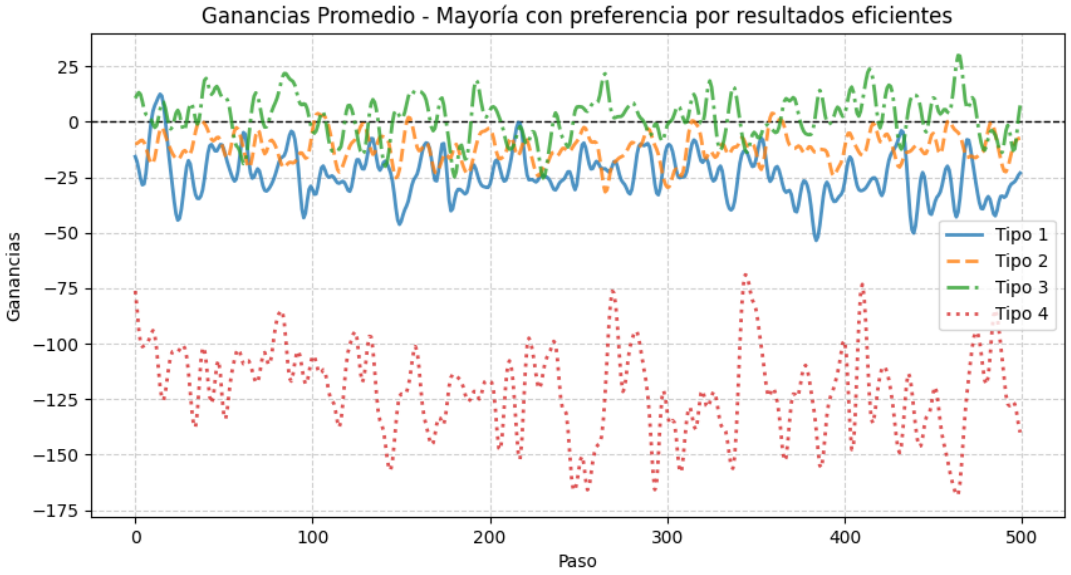

Figura 244: Ganancias de cada grupo cuando la población es de 50 agentes y la distribución de rivales conocidos y una mayoría de rivales desconocido con preferencia por resultados eficientes.