Optimización de la Gestión de Inventarios en cadenas de suministros de alimentos mediante Aprendizaje por Refuerzo

Camilo Aguilar León

Universidad de los Andes, c.aguilarl@uniandes.edu.co

Este trabajo se centra en mejorar las decisiones de las empresas de alimentos que compran productos agrícolas para satisfacer la demanda de los consumidores finales, un proceso conocido como la 'primera milla' de la cadena de suministro. Dado que estas decisiones se ven afectadas por factores dinámicos como la demanda, los precios, y la disponibilidad de productos, es crucial una toma de decisiones óptima para maximizar las ventas y minimizar los costos relacionados con el transporte, la pérdida de inventario y otros gastos. En este contexto volátil, propongo una serie de modelos basados en aprendizaje por refuerzo, especialmente utilizando Q Learning, para captar la naturaleza estocástica de la cadena de suministros. Estos modelos están diseñados para ayudar a las empresas a tomar decisiones informadas y eficientes. Se exploran diferentes enfoques, variando en complejidad, que integran variables clave como la demanda, los precios de compra, las cantidades disponibles, los costos de transporte y la gestión de inventarios en cada período. El objetivo es maximizar la recompensa a corto y largo plazo. La implementación de estos modelos requiere de estrategias creativas para adaptar la información disponible a los algoritmos de aprendizaje por refuerzo. Tras un entrenamiento adecuado, los modelos demuestran su capacidad para aprender patrones que maximizan la eficiencia operativa. La comparación con políticas de manejo de inventarios tradicionales muestra una mejora significativa, evidenciando la eficacia de estos modelos en la toma de decisiones estratégicas, considerando tanto el corto como el largo plazo.

Contenido

1 INTRODUCCIÓN

El sector de alimentos, crucial para la economía y el bienestar social, se enfrenta a desafíos únicos debido a su complejidad inherente y a la incertidumbre constante en la demanda. Esta incertidumbre afecta significativamente la vida útil de los productos, dada su alta perecibilidad (Gutiérrez, 2021) y en general a la toma decisiones de las partes de la cadena. A esto se suma la falta de un intercambio eficiente de información entre los diferentes agentes de la cadena de suministro, lo que impide un control óptimo y dinámico de las operaciones (Gutiérrez, 2021).

En este contexto, las cadenas de suministro de alimentos cortas están ganando atención por su capacidad para generar beneficios sociales, económicos y ambientales, marcando un contraste con los enfoques más tradicionales (EU Food Information Council, 2021). Sin embargo, las expectativas sobre el valor aportado por las cadenas de suministro son cada vez mayores, especialmente en un entorno marcado por la volatilidad del mercado, el aumento de los costos y la aceleración de la digitalización (Kaizen Institute, 2021).

La agilidad se ha convertido en un aspecto fundamental para las cadenas de suministro modernas. La capacidad de responder rápidamente a cambios en la demanda del cliente, la competencia o interrupciones en el suministro es esencial para mantener la competitividad en el mercado (IBM, 2021). No obstante, los métodos tradicionales de toma de decisiones, particularmente en la optimización y ubicación del inventario, han demostrado ser ineficientes frente a cómo funciona actualmente el comercio (Cozowicz, 2021).

Este trabajo propone un modelo innovador diseñado para manejar eficazmente la incertidumbre inherente en las cadenas de suministros de alimentos. El objetivo es facilitar la toma de decisiones óptimas respecto a la adquisición de productos, considerando una serie de variables críticas como la perecibilidad de los productos, la demanda fluctuante, precios y cantidades inciertas, costos de transporte, y la gestión eficiente del inventario.

Para simplificar el análisis y reducir la complejidad inherente a este tipo de cadena de suministro, el modelo se centra en la 'primera milla', es decir, desde los proveedores hasta el depósito. En cuanto a las decisiones de transporte, se asumen viajes directos de ida y vuelta sin escalas intermedias. Respecto a la demanda y las ventas, se consideran como realizadas directamente en el depósito. Así, el desafío se centra en la toma de decisiones de compra basada en la información recogida de la cadena de suministros, analizada periodo a periodo dentro de un horizonte temporal específico.

La complejidad del problema radica en la necesidad de integrar y balancear múltiples factores. Decisiones basadas exclusivamente en la demanda sin considerar los costos de transporte, o en los precios sin tener en cuenta la demanda, pueden resultar en utilidades subóptimas. Además, la incertidumbre constante en la cadena de suministros agrega un nivel adicional de dificultad,

obligando a los agentes a tomar decisiones sin una comprensión completa de lo que ocurrirá en periodos futuros. Esta incertidumbre subraya la necesidad de un enfoque que no solo sea ‘greedy’ sino que también anticipe y planifique para futuras eventualidades.

Para abordar los desafíos presentados en la cadena de suministros de alimentos, este trabajo propone la creación de un entorno virtual que simule todos los elementos de una cadena de suministro. Este entorno facilitará la interacción con distintos agentes basados en Aprendizaje por Refuerzo (RL, por sus siglas en inglés). Estos agentes, diseñados para procesar los datos de la cadena de suministro de manera única, tomarán decisiones diversas, permitiendo así el análisis de varias alternativas.

Los agentes se fundamentan en dos técnicas de RL: Q Learning y DQN (Deep Q Network). DQN es una evolución de Q Learning que integra el aprendizaje profundo para superar algunas limitaciones de su predecesor. Mediante un proceso iterativo, cada agente aprende interactuando con el ambiente diseñado, capturando los patrones inherentes a este.

La efectividad de estos agentes se analizó comparándolos con políticas de manejo de inventarios implementadas de manera 'greedy'. Los resultados demostraron una mejora significativa en la toma de decisiones por parte de los agentes de RL frente a estas políticas convencionales, aprendiendo a optimizar sus decisiones a lo largo de los episodios. Los hallazgos clave de esta investigación se resumen de la siguiente manera:

- **Capacidad de Captura de Dinámicas Complejas:** Las técnicas de Q Learning y DQN demuestran su eficacia para manejar la complejidad de sistemas como las cadenas de suministros. Sin embargo, hay que considerar que a mayor complejidad del sistema implica un aumento en el costo computacional para entrenar estos agentes. Es importante resaltar igualmente que después de entrenar, la toma de decisiones sucede casi de manera instantánea para ambas técnias.

- **Mejora sobre Políticas Tradicionales de Manejo de Inventarios:** Al comparar con políticas tradicionales de manejo de inventarios como R,Q y S,S, los agentes de RL más complejos muestran una mejora sustancial en los resultados.

- **Toma de Decisiones Multifactorial:** Los agentes no solo reaccionan a situaciones de bajo inventario o alta demanda, sino que también aprovechan oportunidades como precios bajos y altas disponibilidades de producto.

- **Visión a Largo Plazo en la Toma de Decisiones:** Los agentes consideran cómo sus decisiones afectan tanto el periodo actual como los futuros. Las decisiones están

interconectadas, de modo que el estado al que llega un agente y la decisión que debe tomar en esta situación afecta a la toma de decisión inicial que llevo a ese estado.

- **Eficacia de la Discretización de Datos:** La discretización de datos continuos no afecta negativamente el rendimiento de manera significativa, siempre y cuando se mantenga una granularidad adecuada que represente distintos estados de manera efectiva.

En conclusión, este trabajo aspira a contribuir significativamente a la tecnificación y a la implementación de la inteligencia artificial en las cadenas de suministros de alimentos, a través de la construcción de agentes de toma de decisiones. Si bien estas técnicas requieren grandes volúmenes de datos para el aprendizaje, su capacidad para facilitar decisiones rápidas y precisas las hace ideales en contextos donde la velocidad es prioritaria. A pesar de que el entrenamiento inicial puede ser intensivo en términos de recursos y datos, una vez implementados, estos modelos tienen la capacidad de seguir aprendiendo y adaptándose en operación real con un costo marginal mínimo. Este proceso, conocido como 'aprendizaje en línea' ('online learning'), es particularmente valioso en entornos cambiantes que requieren la integración continua de nuevos datos (Hugging Face, n.d.).

2 DESCRIPCIÓN GENERAL

2.1 Objetivos

El propósito principal de esta investigación es desarrollar un modelo de aprendizaje por refuerzo que integre con eficacia los componentes estocásticos de las cadenas de suministros, con el fin de facilitar la toma de decisiones eficientes e informadas. Para lograr este objetivo general, el trabajo se enfoca en varios objetivos específicos. Primero, se identificarán y analizarán los elementos clave y la dinámica temporal de una cadena de suministro para comprender su evolución y complejidad. A continuación, se diseñarán y desarrollarán modelos de aprendizaje por refuerzo capaces de procesar de manera efectiva los datos de la cadena de suministros, proporcionando un soporte crucial para la toma de decisiones estratégicas. Además, se implementarán los modelos propuestos de manera que el trabajo sea replicable y verificable en diferentes contextos. Otro objetivo clave es validar el aprendizaje y el rendimiento de estos modelos en diversos escenarios, comparándolos con un modelo de optimización determinístico, que dispone de todos los datos de todos los periodos, y con las políticas tradicionales de manejo de inventarios. Finalmente, se identificarán las áreas en las que estas técnicas de aprendizaje por refuerzo sobresalen al ser aplicadas en las cadenas de suministros, destacando sus ventajas y áreas de mejora.

2.2 Antecedentes

En el ámbito de las cadenas de suministro, la literatura abarca una amplia gama de modelos decisionales que abordan problemáticas de diversa complejidad. Un ejemplo notable es la investigación de Gomez et al. (2023), que ha sido una fuente de inspiración significativa para este trabajo. Este estudio destaca por su enfoque innovador, que integra la gestión dinámica de inventarios de múltiples productos perecederos a lo largo de varios períodos, con el enrutamiento de adquisiciones caracterizado por demandas fluctuantes, precios de compra variables y suministro incierto. Este trabajo y sus formulaciones son bastantes útiles para el diseño e implementación de las técnicas de RL, utilizando definiciones y notaciones alineadas con este enfoque. Se realizaron ajustes específicos en este entorno para facilitar la implementación efectiva de los agentes de RL.

Ahora bien, la aplicación de la inteligencia artificial, especialmente el aprendizaje por refuerzo, en la optimización de cadenas de suministros es un campo relativamente nuevo con escasos estudios formales. Sin embargo, los avances recientes en hardware y tecnología han propiciado un incremento en la implementación de modelos complejos de RL.

En esta línea, Hubbs et al. (2019) introdujeron 'or-gym', un entorno de simulación especializado para abordar problemas de investigación de operaciones para casos como el de cadena de suministro y manejo de inventarios. Este entorno simula escenarios realistas, lo que permite el entrenamiento y validación de modelos de RL en contextos que reflejan los retos actuales. Además, este trabajo propone formulaciones detalladas para varios problemas de esta área, como el de la cadena de suministros.

Por su parte, Hutse (2019) contribuye al ámbito incrementando la complejidad de los problemas al incorporar variables como tiempos de entrega y espacios de acción continuos. Este estudio explora técnicas avanzadas de RL, como DQN y DDPG, ofreciendo un enfoque más preciso en la actualización de recompensas y en la estimación de estados y acciones. La utilización de la librería GYM de Python en este trabajo facilita notablemente su replicabilidad.

En un enfoque similar, van Helsdingen (2022) desarrolla y evalúa agentes de Q Learning y DQN para las cadenas de suministros, enfocándose en aspectos críticos como el ajuste de hiperparámetros. Este estudio también incluye una comparación de estos agentes con métodos de simulación comerciales, utilizando datos reales para validar sus hallazgos.

Finalmente, van Hasselt (2010) aportó significativamente al campo con la introducción de Double Q Learning, una variante de Q Learning diseñada para entornos de RL en espacios continuos. Esta metodología, que utiliza dos redes de valor para minimizar la sobreestimación en las acciones, ha demostrado ser más robusta y eficiente en comparación con los métodos tradicionales de Q Learning, especialmente en contextos estocásticos como las cadenas de suministro.

A pesar de estos avances, la mayoría de los estudios existentes se han centrado en la formulación inicial y básica de agentes de RL para problemas de cadenas de suministros, dejando un amplio margen para explorar y maximizar el potencial de estas técnicas avanzadas en investigaciones futuras.

3 DISEÑO Y ESPECIFICACIONES

3.1 Definición del problema

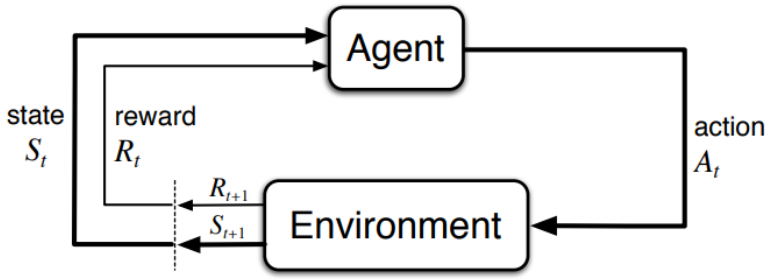
El trabajo que se desarrolla en esta tesis aborda una problemática intrínsecamente vinculada con la operatividad de las cadenas de suministro alimenticias, enfocándose en la etapa inicial o "primera milla". Específicamente, se examina el rol de un intermediario situado en un punto medio de la cadena, cuya función es adquirir alimentos de agricultores y otros productores para luego venderlos y obtener beneficios. En este contexto, se plantea un escenario donde múltiples proveedores ofrecen variados productos alimenticios, y el objetivo es maximizar la rentabilidad derivada de la compra y venta de estos productos.

Cada proveedor posee un catálogo específico de productos, y se recopila periódicamente información referente al precio y disponibilidad de cada ítem. Por otro lado, el intermediario, quien es el foco principal de este estudio, administra un almacén donde guarda el inventario adquirido. Este inventario debe satisfacer la demanda de productos, que se conoce al inicio de cada ciclo y que varía a lo largo del tiempo, considerando además la tasa de perecimiento fija de los alimentos. El precio de venta se mantiene constante durante todos los periodos y siempre es superior al de compra, que puede variar. Por otro lado, existen costos asociados a no satisfacer la demanda o tener inventarios sobrantes. De esta manera, el objetivo general es maximizar las ventas y minimizar el inventario sin afectar decisiones futuras.

Adicionalmente, la ubicación de los proveedores respecto al almacén es fija, pero se incurre en costos de transporte directamente relacionados con la distancia y la cantidad de vehículos necesarios, limitados por su capacidad de carga. Este problema se enfoca en un horizonte temporal discreto y fijo, buscando maximizar las utilidades considerando todos los factores mencionados. El intermediario, enfrentando la incertidumbre de la información futura, debe tomar decisiones de compra cada periodo, determinando cuánto y qué productos adquirir de cada proveedor.

3.2 Formulación del problema

En el contexto de este trabajo, la formulación del problema de la cadena de suministro alimentaria como un Proceso de Decisión de Markov (MDP) ofrece un marco robusto para aplicar técnicas de aprendizaje por refuerzo, como se explica en "Reinforcement Learning: An Introduction" de Sutton and Barto [http://www.incompleteideas.net/book/RLbook2020.pdf]. Este enfoque es particularmente pertinente dada la naturaleza secuencial y la interdependencia entre decisiones y recompensas en la gestión de la cadena de suministro.



(cita aquí http://www.incompleteideas.net/book/RLbook2020.pdf

En este MDP, el agente (el intermediario en la cadena de suministro) interactúa con un ambiente que cambia en respuesta a sus decisiones. Cada acción tomada por el agente se basa en el estado actual del ambiente, y como resultado de estas acciones, el ambiente proporciona una recompensa que refleja la efectividad de la decisión. El MDP se define por una tupla (S, A, P, R), donde S representa el conjunto de estados posibles, A el conjunto de acciones disponibles, P la función de transición de estados, y R la función de recompensa.

Antes de detallar los elementos específicos de la tupla que constituye el MDP, resulta esencial contextualizar y describir los atributos adicionales del ambiente en el que opera el sistema de cadena de suministro. Primordialmente, este entorno se caracteriza por funcionar en un horizonte de tiempo finito y discreto, representado por un conjunto de periodos, denotado como T, que comprende una serie de periodos temporales fijos y secuenciales.

Además, el sistema incluye un conjunto de proveedores, señalado como S, y una diversidad de productos, indicados como P. Cada proveedor s en el conjunto S está localizado a una distancia específica del depósito, identificada como l_s. Esta distancia es crítica, ya que influye directamente en los costos y la logística de transporte. También, se denota al conjunto de productos de un proveedor como P_s y al conjunto de proveedores que ofrecen un producto como S_p. Por otra parte, cada producto p en el conjunto P se caracteriza por una tasa de perecimiento, denotada como ϕ, y un precio de venta, x_p. De este precio de venta surgen los costos relacionados con no cumplir la demanda k_p o por mantener el inventario h_p. Estos valores son un porcentaje del precio de venta y pueden tener varios significados en el contexto de las cadenas de suministros por lo que fueron

incluidos para darle más flexibilidad al modelo y también para apoyar al diseño de los modelos. Estos factores son vitales para la gestión eficiente del inventario y la estrategia de precios.

En lo que respecta a la logística de transporte, cada vehículo utilizado para el traslado de alimentos posee una capacidad máxima, Q, y opera a una velocidad v. También se considera un tiempo de carga fijo, τ, y un costo de transporte, c, que se calcula por unidad de tiempo. Estos parámetros del vehículo son fundamentales para planificar las rutas de entrega y optimizar los costos operativos. Por lo que es necesario modelar también la cantidad de vehículos que se usan para recoger los productos de cada proveedor como n_{st} Todas estas características –distancias, tasas de perecimiento, precios de venta, y especificaciones de los vehículos– son constantes a lo largo de los periodos y establecen las condiciones iniciales del ambiente.

Ahora bien, con estos elementos intrínsecos al ambiente definidos, se puede detallar cada uno de los elementos del MDP en el contexto específico de nuestro problema:

- **Estados (S):** El estado es una tupla que incluye el inventario actual I_{pt}, la demanda d_{pt}, el precio de compra p_{spt}, y la cantidad disponible q_{spt} para cada proveedor $s \in S$, para cada producto $p \in P$ y para cada periodo $t \in T$. El inventario es directamente influenciado por las decisiones del agente, mientras que los otros componentes son determinados aleatoriamente por el ambiente.
- **Acciones (A):** Las acciones disponibles para el agente son definidas por la cantidad a comprar z_{spt} de cada producto $p \in P$, de cada proveedor $s \in S$ durante cada periodo $t \in T$. Estas acciones son cruciales para la actualización del estado y la gestión del inventario.
- **Función de Transición (P):** La función de transición modela la probabilidad de pasar de un estado a otro dada una acción específica. En este caso, la demanda, el precio de compra y las cantidades disponibles aportan elementos aleatorios a esta función de transición que no dependen ni del estado actual ni de la acción. Por otro lado, el inventario depende únicamente del estado actual y de la decisión por lo que esta parte es determinista.
- **Función de Recompensa (R):** La función de recompensa en este MDP es un poco distinta a la generalización que se suele hacer en el sentido de que depende únicamente del estado actual y la acción tomada, y no del estado resultante. Esto refleja la naturaleza inmediata de la recompensa obtenida de la decisión de compra, como las ganancias o pérdidas realizadas en un periodo dado.

Además, es crucial resaltar la naturaleza continua y potencialmente, pero poco probable, infinita de las variables involucradas (inventario, precios, demanda, disponibilidad). Aunque en la práctica, estas variables operarán dentro de un rango específico, su naturaleza continua y no finita agrega una capa de complejidad al problema. Además, el modelo incluye características fijas como la distancia de los proveedores al depósito, la tasa de perecimiento de los productos, los precios de

venta, y las especificaciones de los vehículos de transporte, que son constantes a lo largo del tiempo, pero influyen en las decisiones y recompensas.

De toda la formulación descrita anteriormente pueden surgir casos en los que los estados son inconsistentes por lo que se establecen ciertas restricciones en el ambiente que permiten modelas estas reglas que se deben cumplir para validar la factibilidad de las acciones tomadas y los estados del ambiente.

$$I_{pt} = I_{pt-1}(1-\phi) + \sum_{s\in S_p} z_{spt} - \min\left(d_{pt}, I_{pt-1}(1-\phi) + \sum_{s\in S_p} z_{spt}\right), \forall p \in P, t \in T | t > 0 \quad (1)$$

$$\sum_{p\in P} z_{spt} \leq Q n_{st}, \qquad \forall s \in S, t \in T \quad (2)$$

$$z_{spt} \leq q_{spt}, \qquad \forall s \in S, p \in P, t \in T \quad (3)$$

$$z_{spt} \geq 0, \qquad \forall s \in S, p \in P, t \in T \quad (4)$$

El modelo incorpora varias restricciones clave para garantizar su operatividad y realismo. La primera restricción se enfoca en la dinámica de actualización del inventario en cada periodo y en cómo se atiende la demanda utilizando las cantidades disponibles de cada producto. La segunda restricción asegura coherencia entre la cantidad de vehículos necesarios para el transporte desde un proveedor y la cantidad total de producto adquirido. La tercera restricción, no menos importante, establece que la cantidad de producto comprada no puede exceder la cantidad disponible en el proveedor. Finalmente, se impone que la cantidad de producto a comprar debe ser siempre un valor positivo. Estas restricciones, combinadas con la formulación previamente descrita, permiten construir adecuadamente tanto el ambiente como el problema en cuestión.

En cuanto a la función de recompensas para el modelo, se define de la siguiente manera:

$$R(s_t, a_t) = \sum_{p\in p} \sum_{s\in S_p} \left(x_p z_{spt} - \max\left(k_p\left(d_{pt} - z_{spt}\right), h_p\left(z_{spt} - d_{pt}\right)\right) - p_{spt} z_{spt}\right) - \frac{c}{v} \sum_{s\in S} l_s n_{st} \quad (5)$$

Esta ecuación tiene como objetivo maximizar las ganancias obtenidas de la venta de productos, descontando los costos asociados a mantener el inventario, no suplir la demanda, la compra de productos y el transporte de estos. La función de recompensas es crucial para modelar la eficacia de las acciones tomadas por el agente dentro del ambiente, evaluando el rendimiento de estas decisiones en términos de rentabilidad. En este sentido se busca un agente que logre maximizar estas recompensas individualmente para cada periodo. Esto no significa que el agente tomara decisiones únicamente pensando en el periodo en el que esté tomando la decisión. Esta búsqueda de las decisiones que maximicen las recompensas durante todos los periodos se representa como la búsqueda de una política π^*, la cual define las acciones tomadas por el agente, que permita

maximizar las recompensas obtenidas durante todos los periodos como se describe en Gomez et al. (2023)

$$\pi^* = argmax_{\pi \in \Pi} \mathbb{E}\left(\sum_{t \in T} R(s_t, a_t^{\pi}) \mid S_0\right) \qquad (6)$$

Por lo que, resumiendo el problema y la formulación en tanto a este, se busca obtener un agente que tenga la política optima que maximice las recompensas obtenidas a través de los episodios. Este agente deberá desarrollar maneras de aproximarse a esta política optima de distintas maneras logrando en muchos casos soluciones poco adecuadas para este ambiente.

4 DESARROLLO DEL DISEÑO

En el proceso de diseño de este sistema, se adoptó un enfoque multifacético para crear agentes con propósitos variados. Inicialmente, se desarrollaron dos agentes inspirados en políticas de manejo de inventarios existentes en la literatura. Estos agentes, fundamentados en decisiones 'greedy' o codiciosas, establecen políticas base que actúan como referencia para evaluar y mejorar las estrategias de los agentes subsecuentes.

Además, se diseñó un modelo de optimización entero mixto. Este modelo, que se alinea con la formulación del MDP previamente descrita, es esencial para determinar una cota superior teórica, representando el ideal o el mejor resultado posible que un agente pudiera alcanzar en un escenario específico. Este enfoque ofrece un punto de comparación valioso para medir la efectividad de las políticas implementadas por los agentes.

En lo que respecta a los agentes diseñados específicamente para abordar y optimizar el problema, se desarrollaron tres modelos distintos: dos utilizando el aprendizaje Q-Learning y uno empleando Deep Q-Networks (DQN). Cada uno de estos agentes se diferencia en términos de la técnica utilizada y su formulación específica, permitiendo un análisis detallado y comparativo de sus respectivas eficacias y particularidades.

4.1 Recolección de Información

Las decisiones de diseño fundamentales para la construcción de los diversos agentes y modelos se centraron en la utilización y el tratamiento de los datos empleados para instanciar el ambiente del problema. La naturaleza de estos datos fue determinante en la aplicación de distintas técnicas de diseño e implementación. Una parte significativa de estos datos proviene del trabajo realizado por Gomez et al. (2023), quienes proporcionaron información detallada y específica sobre el ambiente que utilizaron en su investigación.

Para ciertos elementos, especialmente aquellos relacionados con el ruteo, se emplearon datos distintos o con pequeños cambios. Por ejemplo, se realizaron cálculos adicionales para determinar variables necesarias, como la distancia. La disponibilidad y el procesamiento de estos datos fueron cruciales para tomar decisiones clave en el diseño, tales como la discretización de variables continuas, la formulación de penalizaciones para los agentes y otros aspectos relevantes.

4.2 Políticas de manejos de inventarios

En el desarrollo de las políticas de manejo de inventarios, se optó por diseñar dos agentes basados en las conocidas políticas deterministas R,Q (también llamada s,Q) y S,s. La política R,Q implica realizar un pedido de tamaño Q cuando el nivel de inventario de un producto cae por debajo del umbral R. Por otro lado, la política S,s rellena el inventario hasta alcanzar un nivel S cuando este desciende por debajo de una cantidad crítica s. Estas políticas, a pesar de su simplicidad, son ampliamente reconocidas y utilizadas tanto en la literatura académica como en la práctica (fuente: https://repositorio.uniandes.edu.co/server/api/core/bitstreams/cd4ea271-aac3-4aca-aa5c-0fa131138e87/content).

Dado el contexto del problema definido, que contempla múltiples productos y proveedores, se decidió adaptar estas políticas a una aproximación 'greedy'. Así, para ambas políticas, la decisión de realizar un pedido se basa en el nivel actual del inventario de cada producto. Sin embargo, la selección del proveedor se realiza considerando el precio de venta de cada uno de ellos. En el caso de que la cantidad disponible en el proveedor con el precio más bajo sea insuficiente para cumplir con el pedido requerido, se procede entonces con el siguiente proveedor que ofrezca el menor precio, y así sucesivamente, hasta completar la cantidad deseada. Esta adaptación permite la implementación efectiva de estas políticas en el ambiente diseñado para el estudio.

4.3 Modelo mixto-entero de optimización

El modelo mixto-entero de optimización difiere significativamente de las políticas de manejo de inventarios previamente mencionadas. Mientras que estas últimas fueron diseñadas como agentes que interactúan con el ambiente de manera secuencial, el modelo de optimización se concibe como una herramienta comparativa y no como un agente per se. Este modelo se nutre de información simulada correspondiente a todos los periodos del estudio, incluyendo datos sobre la demanda, precios y cantidades, entre otros. A partir de esta información, el modelo es capaz de generar los resultados óptimos y máximos posibles para el escenario en cuestión.

Aunque podría parecer que este modelo no tiene una aplicación directa en la práctica, resulta ser de gran valor para comprender los puntos débiles de los agentes y para identificar áreas de mejora. Al comparar los resultados de los agentes con los del modelo de optimización, es posible discernir dónde y cómo se pueden ajustar las estrategias de los agentes para aumentar sus recompensas. Así, la formulación de este modelo de optimización se convierte en un componente crucial para el análisis y la mejora continua del sistema.

$$\max \sum_{t\in T}\sum_{p\in p}\sum_{s\in S_p}\left(x_p z_{spt} - p_{spt} z_{spt}\right) - \frac{c}{v}\sum_{s\in S} l_s n_{st} \qquad (7)$$

$$I_{pt} = I_{pt-1}(1-\phi) + \sum_{s\in S_p} z_{spt} - d_{pt}\,, \qquad \forall p \in P, t \in T | t > 0 \qquad (8)$$

$$\sum_{p\in P} z_{spt} \leq Q n_{st}, \qquad \forall s \in S, t \in T \qquad (9)$$

$$z_{spt} \leq q_{spt}, \qquad \forall s \in S, p \in P, t \in T \qquad (10)$$

$$z_{spt} \geq 0, \qquad \forall s \in S, p \in P, t \in T \qquad (11)$$

$$n_{st} \in \mathbb{Z}^+, \qquad \forall, p \in P \qquad (12)$$

$$I_{p0} = 0, \qquad \forall, p \in P \qquad (13)$$

Este Modelo de Programación Entero Mixto (MIP) representa una versión simplificada del problema, incorporando ciertos supuestos basados en las discusiones previas. Un aspecto notable en esta versión es la omisión de penalizaciones. La razón detrás de esta simplificación radica en la naturaleza de los datos utilizados para la implementación del modelo, los cuales permiten hallar una solución factible sin necesidad de aplicar penalizaciones. En esencia, esto implica que siempre existen cantidades adecuadas de productos disponibles para satisfacer completamente la demanda.

Dado este contexto, el modelo proporciona una perspectiva más específica del escenario ideal, esencial para realizar ejercicios de 'backtesting'. Este enfoque permite evaluar cómo se comportarían los agentes en condiciones óptima, ofreciendo así una referencia clara para comprender el potencial máximo de rendimiento de un agente.

4.3 Agentes Q-Learning

5 IMPLEMENTACION

Describir los detalles de la solución implementada.

5.1 Descripción de la implementación

Describir cómo el proyecto se divide en etapas. Presentar detalle de las etapas.

5.2 Resultados esperados

Describir y justificar las formas de implementar modelos y soluciones, así como las herramientas empleadas. Evaluar la precisión del desempeño esperado considerando el efecto de soluciones aproximadas y errores de medición esperados.

6 VALIDACION

6.1 Métodos

Describir qué pruebas de validación fueron realizadas para evaluar que los resultados satisficieran las especificaciones (de manera concisa, cuantitativa y precisa). Explique las razones para cada prueba y su significancia, así como el protocolo para ella. Documenteestos pasos para permitir la reproducción de las pruebas.

6.2 Validación de resultados

Presentar los resultados de las validaciones (verificación, *testing*) mediante herramientas apropiadas (lógica, estadística, gráficas, tablas, etc.). Incluir datos completos y resultados intermedios. Identificar y cuantificar posibles fuentes de errores de medición.

7 CONCLUSIONES

7.1 Discusión

Resumen del trabajo, discutiendo el desempeño y las limitaciones; problemas encontrados y cómo pudieron o podrían resolverse; lo que falta por hacer.

Validación de los resultados de forma cuantitativa y cualitativa.

7.2 Trabajo futuro

Lista de sugerencias para trabajos futuros. Enfatizar aquellos resultados que merezcan consideración especial (v.gr., casos especiales que deban ser tenidos en cuenta, necesidad de proteger la propiedad intelectual, etc.).

List of suggestion for future work. Emphasize any results that merit special consideration (e.g., special cases to be treated, need to protect intellectual property, etc.).

8 REFERENCIAS

Gutiérrez, A. (2021). Estudio de la cadena de suministro [PDF document]. Recuperado de https://repositorio.ulima.edu.pe/bitstream/handle/20.500.12724/13303/Gutierrez_Estudio-cadena-suministro.pdf?sequence=1

EU Food Information Council. (2021). Los beneficios y la sostenibilidad de las cadenas de suministro de alimentos cortas. Recuperado de https://www.eufic.org/es/produccion-de-alimentos/articulo/Los-beneficios-y-la-sostenibilidad-de-las-cadenas-de-suministro-de-alimentos-cortas

Kaizen Institute. (2021). Importancia de la optimización de la cadena de suministro. Recuperado de https://kaizen.com/es/insights-es/importancia-optimizacion-cadena-suministro/

IBM. (2021). Optimización de la cadena de suministro. Recuperado de https://www.ibm.com/mx-es/topics/supply-chain-optimization

Cozowicz, M. (2021). Reinforcement Learning in Supply Chain [LinkedIn article]. Recuperado de https://www.linkedin.com/pulse/reinforcement-learning-supply-chain-markus-cozowicz/

Hugging Face. (n.d.). Offline vs Online Learning. Recuperado de https://huggingface.co/learn/deep-rl-course/en/unitbonus3/offline-online

Hubbs, C., Heitz, G., & Dilkina, B. (2019). Or-gym: A Reinforcement Learning Environment for Operations Research Problems. Recuperado de http://egon.cheme.cmu.edu/Papers/Hubbs_or_gym_9_11.pdf

van Hasselt HP (2010) Double Q-Learning. In: Advances in Neural Information Processing Systems, The MIT Press, vol 23

https://libstore.ugent.be/fulltxt/RUG01/002/790/831/RUG01-002790831_2019_0001_AC.pdf

https://arno.uvt.nl/show.cgi?fid=159101

http://www.incompleteideas.net/book/RLbook2020.pdf

https://repositorio.uniandes.edu.co/server/api/core/bitstreams/cd4ea271-aac3-4aca-aa5c-0fa131138e87/content

Apéndices

Datos relevantes que puedan ser consultados para soportar el diseño, la implementación y / o los resultados.