

Trading Strategies in the Peruvian Foreign Exchange Market with Dynamic Optimization: A Hybrid HMM–Deep Reinforcement Learning Approach

Aradiel Castañeda Hilario, Doctor¹ ; Mas Azahuanche Guillermo Antonio, Doctor¹ ; Arteaga Cortez, Humberto Urbano, Doctor¹, Castillo Paredes, Omar Tupac Amaru, Msc¹ ; Reinoso Palacios, Artemio Ruben, Msc¹, Vilchez Inga, César, Doctor¹ ; Carpena Velasquez, Enrique Wilfredo, Doctor² ; ¹Universidad Nacional del Callao, Perú, haradielc@unac.edu.pe, gamasa@unac.edu.pe, huarteagac@unac.edu.pe, otacastillop@unac.edu.pe, arreinosop@unac.edu.pe, cvilchezi@unac.edu.pe, ²Universidad Nacional Pedro Ruiz Gallo, Perú, ecarpena@unprg.edu.pe

Abstract– Furthermore, this study develops and evaluates a hybrid architecture for algorithmic trading in the USD/PEN foreign exchange market, where a Gaussian Hidden Markov Model (HMM) infers latent market regimes and a deep reinforcement learning (DRL) agent optimizes Buy/Sell/Hold actions within an MDP formulation.

Moreover, a reproducible data pipeline is implemented to integrate heterogeneous financial and macroeconomic sources, producing a consolidated dataset with 4,126 observations and 1,113 variables spanning 2010-02-02 to 2025-12-05, with continuity checks to support stable learning.

The inferred regime signal (e.g., bearish/volatile, sideways, bullish) is incorporated as exogenous context to mitigate market non-stationarity and improve policy learning.

Finally, out-of-sample benchmarking shows that the RL-based strategy outperforms supervised baselines and Buy & Hold, achieving the highest cumulative return (28.64%) and improved risk-adjusted performance when regime context is included (Sharpe 0.116 for RL+HMM vs 0.081 for RL-only), while remaining robust in adverse periods where the passive benchmark incurs losses (-14.56%).

In addition, training dynamics suggest PPO yields more stable convergence than alternative DRL methods in a noisy financial environment.

Keywords: USD/PEN; Hidden Markov Models; market regimes; deep reinforcement learning; PPO; algorithmic trading; risk management; CRISP-DM..

Estrategias de Trading en el Mercado Cambiario Peruano con Optimización Dinámica: Un Enfoque Híbrido HMM–Deep Reinforcement Learning

Resumen– En primer lugar, este estudio desarrolla y evalúa una arquitectura híbrida para trading algorítmico del tipo de cambio USD/PEN, donde un Modelo Oculto de Markov Gaussiano (HMM) infiere regímenes latentes y un agente de aprendizaje por refuerzo profundo optimiza decisiones Buy/Sell/Hold en un entorno tipo MDP. En la fase de datos, se implementa un pipeline reproducible con enfoque tipo Medallion y orquestación automatizada, consolidando información heterogénea para reducir intervención manual y riesgo de error.

Asimismo, el preprocesamiento integra series financieras y macroeconómicas en un dataset final de 4,126 registros y 1,113 variables, con cobertura del 02/02/2010 al 05/12/2025, y con verificación de completitud y continuidad temporal para asegurar estados definidos durante el aprendizaje.

La capa HMM aporta un contexto de régimen discreto (p. ej., bear/volátil, sideways y bull) que se incorpora como feature exógeno para mitigar la no estacionariedad del mercado.

Finalmente, el benchmarking fuera de muestra evidencia que el enfoque RL supera a modelos supervisados y al Buy & Hold, alcanzando el mayor retorno acumulado (28.64%), mientras que el filtrado por regímenes mejora la eficiencia ajustada al riesgo (Sharpe 0.116 en RL+HMM vs 0.081 en RL base) y mantiene robustez ante escenarios adversos donde el benchmark pasivo registra pérdidas (-14.56%).

De forma complementaria, el análisis de entrenamiento sugiere que PPO ofrece mayor estabilidad de convergencia frente a alternativas como A2C y DDPG en un entorno financiero ruidoso.

Palabras clave: USD/PEN; Modelos Ocultos de Markov; regímenes de mercado; aprendizaje por refuerzo profundo; PPO; trading algorítmico; gestión del riesgo; CRISP-DM.

I. INTRODUCCIÓN

En primer lugar, la estabilidad del tipo de cambio USD/PEN constituye un determinante relevante para la estabilidad macrofinanciera del Perú, en particular en un contexto de dolarización parcial donde movimientos abruptos del tipo de cambio pueden amplificar vulnerabilidades por “efectos hoja de balance” en hogares, empresas e intermediarios financieros. En esa línea, el Banco Central de Reserva del Perú (BCRP) sostiene que la intervención cambiaria busca reducir la volatilidad excesiva y mitigar riesgos asociados a la dolarización financiera [1], mientras que también enfatiza la importancia de la volatilidad cambiaria para la estabilidad financiera en economías emergentes [2]. No obstante, la dinámica del USD/PEN responde a choques externos (tasas internacionales, términos de intercambio, aversión al riesgo) y a factores internos, lo que incrementa la no linealidad y la no estacionariedad del proceso generador de

datos, dificultando la predicción y la toma de decisiones con enfoques rígidos.

Asimismo, aunque los modelos econométricos lineales han sido ampliamente utilizados para describir relaciones promedio, los mercados financieros presentan cambios de régimen, colas pesadas y dependencia temporal compleja, por lo que la decisión operativa (comprar/vender/mantener) requiere mecanismos que adapten la política conforme cambian las condiciones del entorno. En este marco, el Aprendizaje por Refuerzo (RL) formaliza la interacción agente–entorno como un problema secuencial (MDP), donde el agente aprende una política que maximiza recompensa acumulada a partir de experiencia [3]. En finanzas, esta idea se ha explorado desde enfoques clásicos de RL para optimizar funciones de desempeño de trading y portafolios [4] hasta formulaciones modernas con Deep Reinforcement Learning (DRL) orientadas a aprender representaciones y políticas en entornos complejos [5].

Por otra parte, una limitación recurrente del RL aplicado a trading es que, si el estado observado no captura adecuadamente el “contexto” del mercado, la política puede degradarse ante cambios estructurales (p. ej., pasar de un régimen lateral a uno tendencial o de alta volatilidad). Por consiguiente, una estrategia metodológicamente sólida es incorporar modelos explícitos de regímenes latentes. En particular, los Modelos Ocultos de Markov (HMM) son un marco probabilístico estándar para modelar sistemas Markovianos con estados no observables y observaciones dependientes del estado, lo que resulta útil para inferir regímenes a partir de señales observadas [6]. En ese sentido, un enfoque híbrido HMM–RL permite utilizar el régimen inferido como parte de la representación del estado, mejorando la capacidad del agente para condicionar su acción al contexto de mercado.

En consecuencia, este artículo propone una arquitectura híbrida para la gestión de operaciones en USD/PEN que integra: (i) un HMM para identificar regímenes latentes de mercado y (ii) un agente de (D)RL para optimizar decisiones Buy/Sell/Hold bajo un esquema de recompensa alineado con desempeño ajustado al riesgo. Además, el desarrollo se organiza bajo CRISP-DM, garantizando trazabilidad desde la comprensión del negocio y datos hasta modelado, evaluación y despliegue [7], [8]. A nivel de evaluación, se emplean métricas financieras estándar —por ejemplo, retorno acumulado, Sharpe ratio y max drawdown—, a fin de contrastar el enfoque propuesto frente a estrategias base.

Finalmente, la contribución principal se sintetiza en tres ejes: (1) un mecanismo reproducible de integración y preparación de datos para USD/PEN; (2) un esquema regime-aware (HMM) que reduce la miopía contextual del RL ante no estacionariedad; y (3) una evaluación comparativa basada en métricas de retorno y riesgo. Para facilitar la lectura, el artículo se organiza como sigue: la Sección II describe datos y metodología (CRISP-DM); la Sección III presenta el modelado HMM y el entorno RL/DRL; la Sección IV reporta resultados y análisis; y la Sección V discute limitaciones y líneas futuras.

II. ESTADO DEL ARTE

En primer lugar, la literatura sobre predicción de mercados financieros y optimización de estrategias de trading evidencia un desplazamiento desde enfoques econométricos clásicos hacia arquitecturas de Machine Learning (ML) y, más recientemente, hacia Aprendizaje por Refuerzo (RL) y Aprendizaje por Refuerzo Profundo (DRL). En el caso peruano, el USD/PEN exhibe una dinámica condicionada por choques externos y domésticos, así como por el marco de intervención y estabilidad financiera descrito por el BCRP, lo que refuerza la necesidad de modelos capaces de adaptarse a no linealidad y cambios estructurales en el tiempo [1], [2].

Asimismo, el estado del arte converge en una tesis metodológica clave: predecir precios (enfoque supervisado) no es equivalente a optimizar decisiones (enfoque orientado a política). En consecuencia, investigaciones recientes enfatizan que el desempeño debe evaluarse en términos de utilidad económica y riesgo (p. ej., retorno, Sharpe, drawdown), y no únicamente con métricas de error predictivo, dado que la brecha “predicción → decisión → ejecución” puede introducir sesgos y degradación fuera de muestra [9].

A. Evolución: de la econometría al ML orientado a decisión

Por otra parte, los modelos econométricos han dominado históricamente el análisis de series financieras; sin embargo, la evidencia contemporánea muestra limitaciones para capturar patrones no lineales y dependencias complejas, especialmente en contextos con no estacionariedad. En este escenario, el ML introduce flexibilidad representacional, pero mantiene una limitación recurrente: una alta precisión en “forecasting” no garantiza rentabilidad cuando se incorporan costos de transacción, slippage, restricciones operativas y gestión del riesgo, lo que exige pasar de un objetivo predictivo a uno prescriptivo (decisional) [9].

Vacío 1 (de enfoque): persiste un exceso de trabajos centrados en la predicción del precio, sin cerrar el ciclo de decisión mediante políticas explícitas de trading con criterios de riesgo y ejecución realista [9].

B. Regímenes de mercado y modelos Markovianos en FX

En segundo lugar, la modelación por regímenes se reconoce como una vía robusta para tratar la

heterogeneidad temporal del mercado. En FX, los modelos de Markov switching han demostrado capacidad para capturar cambios de estado y, en ciertos escenarios, producir rendimientos excedentes fuera de muestra frente a estrategias técnicas estándar, sugiriendo que la predictibilidad práctica aumenta cuando se condiciona la decisión al régimen [10].

Vacío 2 (de generalización): aun cuando el enfoque por regímenes es sólido, parte de la literatura presenta mejoras “modestas” y altamente dependientes de la muestra, por lo que se requieren protocolos temporales estrictos y evidencia de estabilidad en el tiempo [9], [10].

C. Modelos Ocultos de Markov (HMM) en finanzas

En tercer lugar, los Modelos Ocultos de Markov (HMM) constituyen un marco probabilístico estándar para inferir estados no observables a partir de emisiones observables, con fundamentos metodológicos consolidados en la literatura seminal [6]. En aplicaciones financieras, se han utilizado para segmentar series en regímenes con propiedades diferenciadas; por ejemplo, trabajos aplicados reportan utilidad del HMM para estructurar señales y capturar dependencias temporales en entornos bursátiles, reforzando su pertinencia como mecanismo de contextualización del mercado [11].

Vacío 3 (de rigurosidad experimental): es frecuente encontrar aplicaciones HMM donde la inferencia de regímenes se evalúa sin validación temporal tipo walk-forward/rolling, sin análisis de sensibilidad (número de estados, supuestos de distribución) y con riesgo de fuga de información, lo cual compromete inferencias causales y desempeño fuera de muestra [9].

D. RL/DRL para trading: fundamentos, avances y límites

En cuarto lugar, el RL formaliza la toma de decisiones secuenciales como un MDP, donde un agente aprende una política que maximiza recompensa acumulada [3]. En finanzas, existe evidencia temprana y relevante que aplica RL para optimizar directamente criterios relacionados con desempeño ajustado al riesgo (p. ej., aproximaciones al Sharpe), estableciendo una base para el aprendizaje de estrategias que se ajustan por retroalimentación del mercado [4]. Sin embargo, la revisión crítica del DRL en trading resalta que los modelos suelen ser frágiles: presentan sobreajuste, alta sensibilidad al diseño de recompensa/estado, y evaluación poco realista al ignorar costos de ejecución, lo que puede inflar el rendimiento reportado [9].

Vacío 4 (de reproducibilidad y realismo): la literatura aún carece de benchmarks reproducibles, comparaciones consistentes y reportes sistemáticos de robustez (ablación/sensibilidad), especialmente cuando se aspira a contribuciones Q1/Q2 [9].

E. Enfoques híbridos “regime-aware”: integración HMM + (D)RL

En quinto lugar, una línea de síntesis metodológica consiste en combinar: (i) un modelo de regímenes que provea “contexto” (p. ej., HMM) y (ii) un agente de RL/DRL que optimice la

política de acción. Conceptualmente, el régimen inferido puede incorporarse como parte del estado del MDP, reduciendo la miopía contextual y mejorando la estabilidad del aprendizaje ante no estacionariedad [6], [9]. Además, trabajos de RL aplicados al trading refuerzan la pertinencia de aprender políticas directamente en función de una recompensa financiera, en lugar de derivar reglas a posteriori desde predicciones puntuales [4].

Vacío 5 (de evidencia empírica localizada): pese a los avances globales, es limitada la literatura que pruebe arquitecturas híbridas HMM-(D)RL con rigurosidad y comparabilidad en mercados emergentes específicos como USD/PEN, donde los mecanismos institucionales pueden alterar la estructura de regímenes.

F. Brecha empírica y contextual: USD/PEN, intervención cambiaria y dolarización parcial

Finalmente, el mercado cambiario peruano se caracteriza por un marco de intervención orientado a reducir volatilidad excesiva y mitigar riesgos asociados a la dolarización financiera, aspectos descritos por el BCRP y que influyen en la dinámica del tipo de cambio y en su estabilidad [1], [2]. Además, estudios del propio BCRP analizan la intervención cambiaria en Perú y su interacción con variables financieras, aportando evidencia de que el USD/PEN puede responder de forma distinta a shocks en comparación con mercados FX sin estas particularidades [12], [13].

Vacío 6 (contextual e institucional): la mayoría de trabajos híbridos o de DRL en trading no incorpora explícitamente señales macro-institucionales locales (p. ej., intervención, dolarización parcial) como parte de la representación del estado, aun cuando estas pueden afectar la transición entre regímenes y la efectividad de la política [1], [12], [13].

Síntesis de vacíos de conocimiento y derivación del aporte

En consecuencia, el estado del arte permite formalizar tres brechas articuladas

Brecha teórica (no estacionariedad y regímenes): necesidad de políticas regime-aware que capturen cambios estructurales de FX con estados latentes explícitos [6], [10].

Brecha metodológica (evaluación realista y reproducible): falta de protocolos consistentes de validación temporal, control de fuga de información, costos de transacción y análisis de robustez, especialmente en DRL [9].

Brecha empírica (USD/PEN): escasez de evidencia para arquitecturas híbridas HMM-(D)RL en el mercado peruano incorporando señales macro e institucionales reportadas por el BCRP [1], [2], [12], [13].

Por tanto, la propuesta híbrida HMM-(D)RL se justifica como respuesta directa a dichas brechas: el HMM aporta segmentación por regímenes (contexto), mientras el RL/DRL optimiza una política de acción coherente con métricas de retorno-riesgo y supuestos operativos; adicionalmente, el enfoque CRISP-DM estructura el ciclo de vida de datos y modelado para trazabilidad y reproducibilidad [7], [8].

III. METODOLOGIA

Para garantizar la replicabilidad y el rigor científico del estudio, se adoptó la metodología estandarizada CRISP-DM (Cross-Industry Standard Process for Data Mining). Este marco de trabajo iterativo permitió estructurar el desarrollo del agente de trading en fases secuenciales, desde la comprensión del entorno económico peruano hasta la validación comparativa de estrategias.

A. Fase 1: Comprensión del negocio

El objetivo central es desarrollar un sistema de soporte a la decisión para el mercado cambiario USD/PEN. La volatilidad de este par de divisas está sujeta a factores exógenos, como el precio del cobre y las tasas de interés de la Reserva Federal (FED). El problema se define como un proceso de optimización estocástica en el que el agente debe maximizar el retorno acumulado $(R_{cum})(R_{\{cum\}})(R_{cum})$, sujeto a restricciones de riesgo (máximo drawdown).

B. Fase 2: Adquisición de datos

Se diseñó una arquitectura de extracción modular para consolidar cuatro fuentes de información heterogéneas, abarcando datos financieros, macroeconómicos y conductuales.

1. Microestructura y variables financieras (Script 01): Utilizando la librería `yfinance`, se extrajeron series históricas diarias (OHLCV) del par USD/PEN y de activos correlacionados. Se incluyeron materias primas críticas para la balanza comercial peruana (cobre $HG=F$, oro $GC=F$) y ETF de mercados emergentes (EPU, EEM), totalizando más de 30 variables financieras.
2. Indicadores macroeconómicos (Scripts 02 y 03): Se desarrollaron conectores específicos para APIs gubernamentales:
 - Banco Central de Reserva del Perú (BCRP): extracción de series mensuales de política monetaria, incluyendo la tasa de referencia (PN07819NM) y el tipo de cambio interbancario promedio.
 - Banco Mundial (World Bank): recolección de indicadores estructurales anuales (PIB, deuda externa/PIB) para Perú y países de la Alianza del Pacífico, permitiendo el análisis de competitividad regional.
3. Modelado de sentimiento de mercado (Script 04): Para capturar la psicología del inversor, se integró el índice Fear & Greed a través de la API de Alternative.me. Dado que el comportamiento del mercado no es lineal, el script de extracción no se limitó a descargar el valor bruto, sino que computó características derivadas in situ para enriquecer el espacio de estados:
 - **Medias móviles de sentimiento:** Se calcularon ventanas de 7 y 30 días (MA7, MA30) para identificar tendencias de euforia o pánico sostenidas.
 - **Z-score de sentimiento:** Se normalizó el índice para detectar anomalías estadísticas mediante:

$$Z_t = \frac{X_t - \mu_{30}}{\sigma_{30}}$$

donde X_t es el valor del índice en el tiempo t , y μ_{30} y σ_{30} representan la media y la desviación estándar móviles de 30 días, respectivamente.

- **Régimen categórico:** Se discretizó el índice en cinco estados (0: miedo extremo a 4: codicia extrema) para facilitar la convergencia del Modelo Oculto de Markov (HMM).

C. Fase 3: Preparación y preprocesamiento de datos

Esta fase constituyó el componente más intensivo de la investigación. Para garantizar la reproducibilidad científica y la integridad de las series temporales, se implementó una arquitectura de procesamiento automatizada (pipeline ETL), controlada por un script maestro de orquestación.

1. Diseño de la arquitectura de datos: Para sistematizar la transformación de la información, se diseñó una arquitectura modular basada en el patrón Medallion Architecture. Este enfoque organiza el flujo de datos en tres capas lógicas de refinamiento progresivo, tal como se ilustra en la Figura 1.

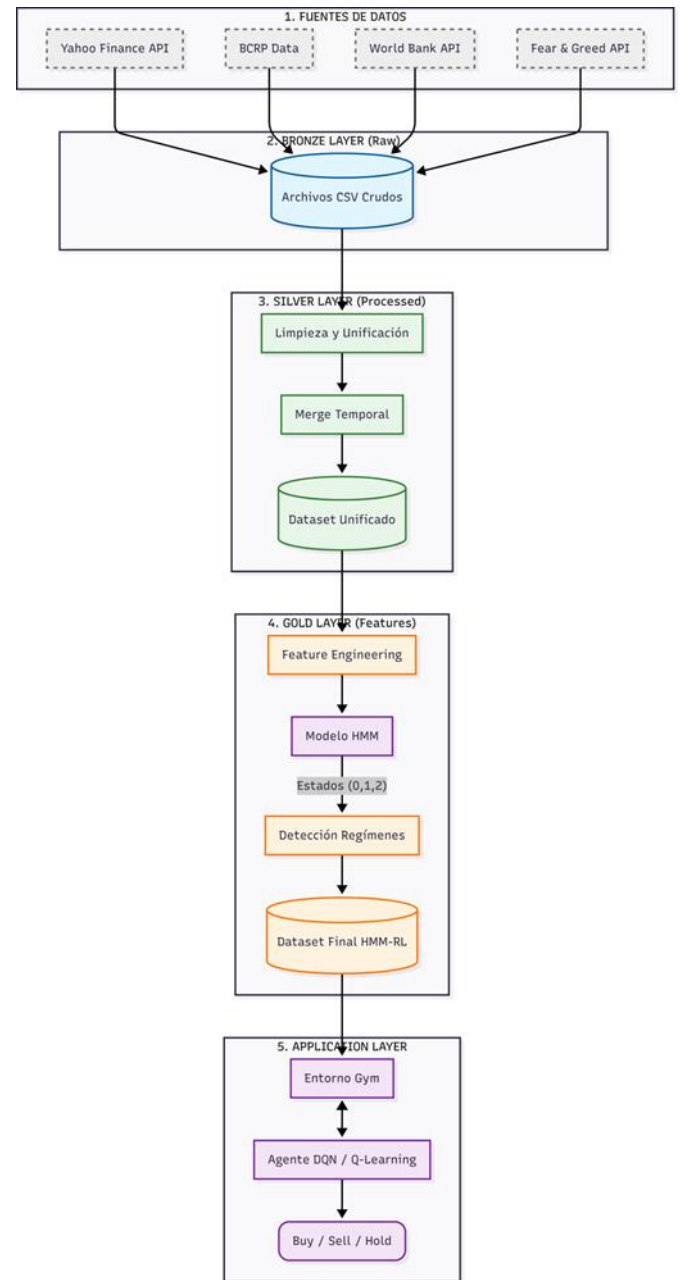


Figura 1. Diagrama del pipeline de datos (Arquitectura Medallion). El flujo se estructura en los siguientes niveles de abstracción:

Capa Bronce (Raw): ingesta de datos crudos provenientes de Yahoo Finance y APIs macroeconómicas, preservando la inmutabilidad de la fuente.

Capa Plata (Processed): datos limpios, normalizados temporalmente y libres de valores nulos, listos para el análisis.

Capa Oro (Features): enriquecimiento analítico mediante indicadores técnicos y la integración del modelo HMM (régimen de mercado) para alimentar al agente.

2.- Orquestación y automatización del flujo (script maestro): Se desarrolló un sistema de ejecución secuencial (ejecutar_pipeline.py) que coordina la interacción entre los módulos de extracción, limpieza y validación descritos en la arquitectura anterior. Esta implementación permite reproducir el experimento completo desde cero, asegurando que el flujo de datos — desde la descarga en bruto hasta el dataset final— ocurra sin intervención manual, mitigando el riesgo de error humano en la manipulación de más de 30 variables financieras.

3.- Limpieza y tratamiento de anomalías (Script 05): Los datos crudos provenientes de fuentes heterogéneas presentaban inconsistencias estructurales, las cuales fueron abordadas en la Capa Plata mediante las siguientes técnicas:

a) Estandarización temporal: Se normalizaron todas las marcas de tiempo a la zona horaria UTC, eliminando la información de huso horario (tz-naive) para evitar desalineaciones entre el cierre de la Bolsa de Lima (BVL) y los mercados globales (NYSE, LSE).

b) Gestión de outliers mediante clipping estadístico: A diferencia de la eliminación tradicional de registros, se optó por una estrategia de preservación de información para eventos de alta volatilidad. Se detectaron valores atípicos que excedían cinco desviaciones estándar ($\mu \pm 5\sigma$) y se aplicó una técnica de clipping (truncamiento). Esto evita que los gradientes del agente de Reinforcement Learning exploten durante el entrenamiento sin perder la señal de estrés del mercado.

c) Imputación de datos faltantes:

- Para series de precios continuas (Yahoo), se utilizó interpolación lineal con un límite máximo de 5 días para cubrir feriados.
- Para datos macroeconómicos (BCRP/World Bank), se aplicó estrictamente forward fill para evitar el sesgo de anticipación (look-ahead bias).

4.- Unificación y sincronización (Script 06): El desafío de integrar datos de frecuencia diaria, mensual y anual se resolvió estableciendo el calendario de cotización del par USD/PEN como “índice maestro”. Se utilizó una estrategia de unión por la izquierda (left join) sobre este índice base. Las variables de baja frecuencia fueron sometidas a un proceso de broadcasting temporal, propagando el último valor oficial vigente a cada día de trading del mes correspondiente.

5.- Ingeniería de características (Script 07): Para transformar las series procesadas en un espacio de estados informativo (Capa Oro), se generaron variables predictoras clasificadas en cuatro dimensiones analíticas:

a) Transformaciones para estacionariedad: Se calcularon

$$\left(r_t = \ln \left(\frac{P_t}{P_{t-1}} \right) \right)$$

retornos logarítmicos

en

ventanas de 1, 5, 21 y 63 días, con el fin de capturar tendencias de corto, mediano y largo plazo.

b) **Indicadores técnicos avanzados:** Se implementaron indicadores de momento y volatilidad, incluyendo la posición del precio dentro de las Bandas de Bollinger (2σ), el RSI (14 y 21 días) y señales MACD.

c) **Indicadores de macro-régimen:** Se desarrollaron indicadores sintéticos específicos, tales como el ratio oro/cobre (proxy de aversión al riesgo global), la señal *risk-on/risk-off* (basada en S&P 500/VIX) y el *spread* cambiario (diferencial entre el precio de mercado y la tasa oficial del BCRP).

d) **Discretización para HMM:** Se generaron variables categóricas, destacando la discretización del VIX en cinco regímenes (de “muy bajo” a “extremo”) para facilitar la convergencia del **Modelo Oculto de Markov (HMM)**.

6.-Validación integral y análisis postfusión (Scripts 08 y 09): Como paso final previo al modelado, se ejecutó un doble protocolo de aseguramiento de calidad:

a) **Auditoría lógica automatizada (Script 08):** Un algoritmo verificó condiciones críticas de parada, tales como dimensionalidad mínima (> 2000 registros), integridad de columnas clave (p. ej., VIX y precio de cierre con $< 5\%$ de valores nulos) y ausencia de *gaps* temporales mayores a 5 días.

b) **Análisis exploratorio postfusión (Script 09):** Se realizó un EDA (*Exploratory Data Analysis*) sobre el *dataset* final unificado para detectar patologías estadísticas sutiles.

- **Limpieza de esparsidad:** Se eliminaron automáticamente variables con más del 95% de valores cero, debido a su baja variancia explicativa.
- **Análisis de distribución:** Se generaron mapas de calor de correlación y diagramas de caja (*boxplots*) para confirmar que las transformaciones de ingeniería de características preservaron propiedades estadísticas adecuadas para el aprendizaje del agente.

D. Fase 4: Modelado y arquitectura híbrida

La propuesta central de este estudio es una arquitectura secuencial donde un modelo generativo no supervisado (HMM) enriquece la percepción del entorno de un agente de control óptimo (RL). Esta fase se implementó mediante los módulos `main_hmm.py`, `trading_env.py` y `rl_agent.py`.

1.-Identificación de regímenes de mercado (HMM): Para mitigar la no estacionariedad de la serie financiera del tipo de cambio USD/PEN, se entrenó un **Modelo Oculto de Markov** (Gaussian HMM) que clasifica la dinámica del mercado en estados latentes discretos.

a) **Formulación matemática:** Se define el modelo $\lambda=(A,B,\pi)$, donde Q representa los estados ocultos y O las observaciones. Las observaciones se definieron como retornos logarítmicos y volatilidad realizada de 21 días. La probabilidad de la secuencia de observaciones se maximiza mediante el algoritmo Baum–Welch:

$$\arg \max_{\lambda} P(O | \lambda) = \sum_Q P(O | Q, \lambda) P(Q | \lambda)$$

b) Decodificación de estados: El algoritmo de **Viterbi** decodifica la secuencia más probable de estados ocultos $St \in \{0,1,2\}$ interpretados como:

- **Estado 0 (Bear/Volátil):** Régimen de corrección o alta incertidumbre.
 - **Estado 1 (Sideways):** Mercado lateral o de transición.
 - **Estado 2 (Bull):** Tendencia alcista con baja volatilidad relativa.
- Esta señal St se normaliza y se integra como un *feature* predictivo exógeno en el vector de estado del agente.

2.-Formulación del entorno de trading (MDP): El problema de decisión se modeló como un **Proceso de Decisión de Markov (MDP)** compatible con la interfaz **Gymnasium**, implementado en la clase `TradingEnvironment`.

a) **Espacio de estados (S):** Vector continuo multidimensional $s_t \in \mathbb{R}^d$, normalizado mediante *z-score* y *clipping* en el rango $[-1,1]$. El vector incluye datos de mercado, indicadores técnicos y variables de sentimiento.

$$s_t = [HMM_t, RSI_{14}, BB_{pos}, MACD_{hist}, ROC_{21}, VIX_{reg}, F\&G, Pos_{t-1}]$$

donde se integran indicadores avanzados como el **Fear & Greed Index (F&G)** y señales de riesgo global (*risk-on/risk-off*), además de la posición previa del agente.

b) **Espacio de acciones (A):** Espacio discreto que permite al agente gestionar su exposición al activo:

$$A = \{0 : \text{HOLD}, 1 : \text{BUY (Largo)}, 2 : \text{SELL (Corto)}\}$$

c) **Función de recompensa (R):** Se define como el retorno logarítmico del portafolio penalizado por costos de fricción ($tc=0,1\%$) Si V_t es el valor del portafolio:

$$r_t = \frac{V_t - V_{t-1}}{V_{t-1}} \cdot \text{Escala} - \text{Penalización}_{\text{ruina}}$$

Se aplica una penalización de -100 si $V_t \leq 0$, forzando al agente a priorizar la preservación de capital.

3) Optimización de política (agentes RL): Se implementaron dos estrategias de aprendizaje para validar la hipótesis de la arquitectura híbrida:

a) Q-Learning tabular (baseline): Implementado en `SimpleTradingEnv`, reduce la complejidad dimensional discretizando indicadores técnicos (RSI, Bollinger, Z-score) en 10 *bins*. Actualiza la tabla QQQ mediante la ecuación de Bellman estándar, con una estrategia de exploración ϵ -greedy.

b) Deep Q-Network (DQN): Implementado en la clase `DQNAgent`, este modelo aproxima la función $Q(s,a)Q(s,a)Q(s,a)$ sin discretización, permitiendo capturar relaciones no lineales complejas.

- **Arquitectura de Red:** Un Perceptrón Multicapa (MLP) con capas ocultas densas de dimensiones [128, 64] y funciones de activación ReLU.
- **Estabilidad:** Se emplea un Replay Buffer de capacidad 20,000 para romper correlaciones temporales y una Target Network que se actualiza suavemente (τ) para estabilizar el objetivo de aprendizaje (MSE Loss).
- **Optimizador:** Se utiliza Adam con una tasa de aprendizaje de 0,0005, decayendo la exploración ϵ de 1,0 a 0,01

E. Fase 5: Validación automatizada de calidad de datos

Previo al entrenamiento de los modelos, se ejecutó un estricto protocolo de aseguramiento de calidad (quality assurance) mediante el Script 08. Este módulo actúa como una compuerta lógica que impide la ejecución del entrenamiento si no se cumplen condiciones críticas.

El script audita el dataset final (`dataset_hmm_rl.csv`) verificando las siguientes restricciones:

2. Dimensionalidad mínima: Se exige un histórico superior a 2,000 registros (aprox. 8 años bursátiles) y un mínimo de 100 características (features) para asegurar la capacidad de generalización.
3. Integridad de datos: Se verifica que ninguna columna crítica (como el precio de cierre o el VIX) tenga más del 5% de valores nulos.
4. Continuidad temporal: El algoritmo detecta gaps o saltos en el índice temporal mayores a 5 días, lo cual podría indicar pérdida de datos durante la extracción.
5. Chequeo de varianza: Se eliminan automáticamente las características con varianza cero (valores constantes), ya que no aportan información al modelo y pueden causar singularidad en la matriz de covarianza del HMM.

Solo tras superar esta validación automatizada (“dataset válido”), los datos son aprobados para ingresar a la fase de entrenamiento del agente.

IV. RESULTADOS

La ejecución de la arquitectura ETL descrita en la metodología permitió consolidar información de fuentes heterogéneas (Yahoo Finance, BCRP y World Bank) en una estructura tabular coherente.

A. Resultados del preprocesamiento y calidad de datos

La ejecución de la arquitectura ETL descrita en la metodología permitió consolidar información de fuentes heterogéneas

(Yahoo Finance, BCRP y World Bank) en una estructura tabular coherente.

1. **Dimensionalidad y cobertura:** El proceso de fusión y limpieza redujo el ruido inicial, resultando en una matriz final de **4,126 registros** y **1,113 variables** (incluyendo indicadores técnicos generados, *lags* y variables macroeconómicas). La ventana temporal abarca desde el **02 de febrero de 2010** hasta el **05 de diciembre de 2025**, cubriendo un periodo de aproximadamente **15.8 años** de historia bursátil. Esta extensión garantiza la captura de múltiples ciclos económicos, incluyendo la crisis del COVID-19 y periodos de alta volatilidad política en Perú.
2. **Integridad de los datos:** Las métricas de calidad de datos confirman la robustez del proceso de limpieza (Script 08):
 - a. **Complejidad:** Se logró un **0.00%** de valores nulos en el dataset final. Las estrategias de imputación (*forward fill* para macroeconomía e interpolación lineal para precios) eliminaron eficazmente los vacíos, asegurando que el agente de RL no enfrente estados indefinidos.
 - b. **Consistencia temporal:** No se detectaron *gaps* (brechas) temporales significativos mayores a **5 días hábiles**, validando la sincronización de los calendarios bursátiles (NYSE, BVL) y de commodities.
3. **Análisis de distribución y anomalías (EDA):** El análisis estadístico exploratorio reveló patrones críticos en la distribución de las variables financieras que impactan el modelado:
4. **Detección de valores atípicos (outliers):** Se identificó que los volúmenes de negociación de commodities presentan un comportamiento de “colas pesadas”. Específicamente, los futuros de oro y plata mostraron la mayor cantidad de eventos extremos (superiores a 4 desviaciones estándar):
 - **Futuros de oro:** 94 eventos extremos detectados (skewness: 5.59).
 - **Futuros de plata:** 88 eventos extremos detectados (skewness: 5.73). Asimismo, el indicador de sentimiento Fear & Greed (versión Z-score) presentó un **37.7%** de valores considerados atípicos, lo que sugiere que el mercado opera frecuentemente en condiciones de estrés emocional extremo.
5. **Variables macroeconómicas:** Se observó una alta varianza en indicadores globales como el PIB en USD (1.46×10^{21}) y la deuda externa (2.12×10^{20}), lo que justificó la aplicación de normalización estricta (Z-score y *clipping*) antes de ingresar estos datos a la red neuronal del agente.
6. Estos resultados confirman que el dataset procesado posee la integridad estructural y la riqueza estadística

necesarias para entrenar modelos de aprendizaje profundo sin riesgo de degradación por datos de baja calidad.

B. Análisis visual de los datos

Para validar las hipótesis sobre el comportamiento del mercado y la calidad de las variables exógenas, se realizó un análisis visual exhaustivo utilizando las **13 proyecciones** generadas durante la fase de exploración (EDA) del archivo `dataset_hmm_rl.csv`.

1) **Estructura de Correlación y Dependencia:** La matriz de correlación (Fig. 2) revela la compleja interdependencia entre los activos bursátiles y los indicadores macroeconómicos.

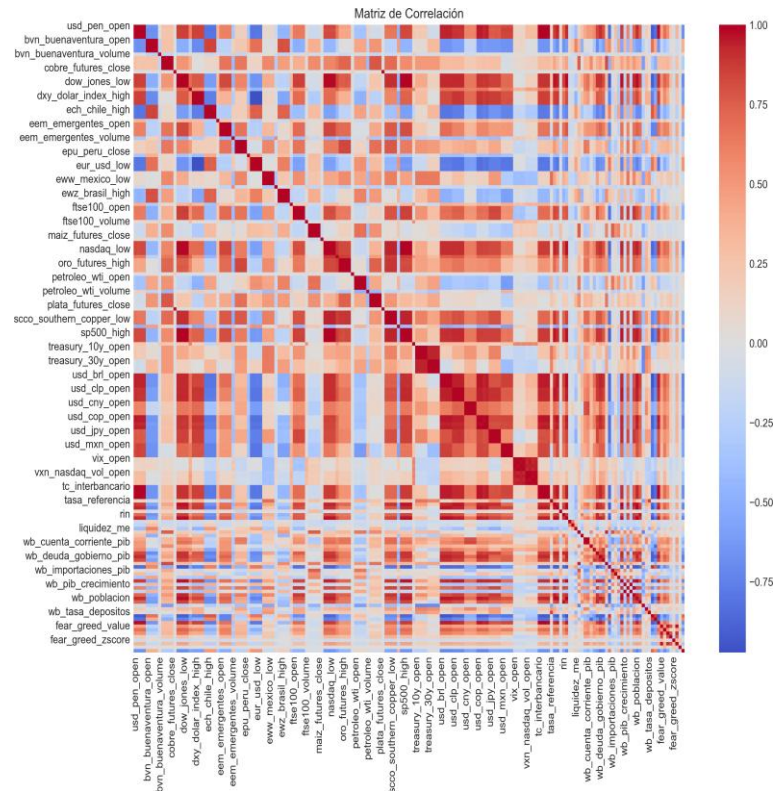


Figura 2. Mapa de calor de correlaciones (Pearson) entre variables financieras y macroeconómicas.

Se observan bloques de alta correlación positiva ($r > 0.8$; zonas rojas) entre los índices de mercado, lo cual es esperable. Sin embargo, destacan las correlaciones inversas (zonas azules) entre los indicadores de volatilidad y los precios, lo que sugiere que el agente podría utilizar la volatilidad como señal de cobertura. Asimismo, el pairplot de variables macroeconómicas (Fig. 3) confirma que no existe una colinealidad perfecta entre los indicadores del Banco Mundial, lo que aporta diversidad de información al espacio de estados del agente.

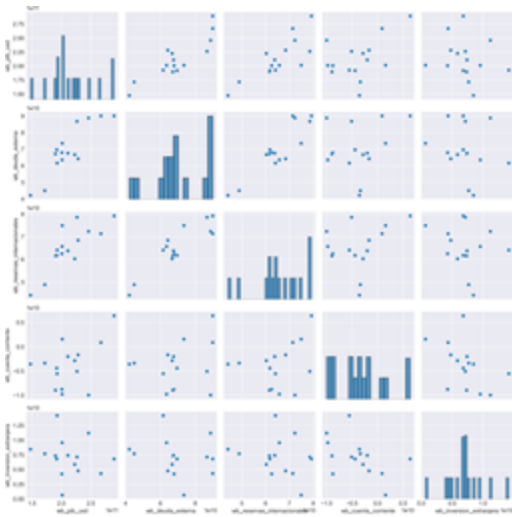


Figura 3. Diagrama de dispersión por pares (Pairplot) de las variables macroeconómicas.

2. Distribución y Escala de Variables

Uno de los hallazgos clave es la disparidad de magnitudes. El análisis de varianza (Figura 4) evidenció que las variables macroeconómicas poseen varianzas exponencialmente mayores que los precios bursátiles.

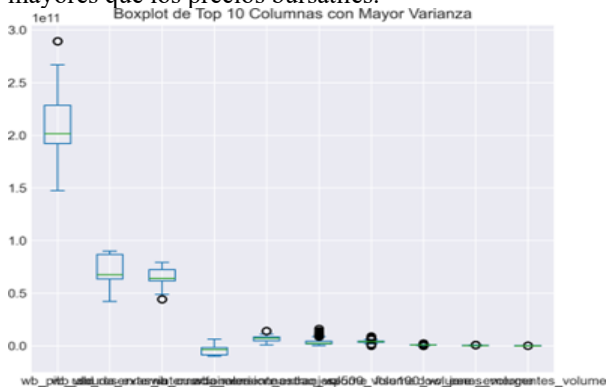
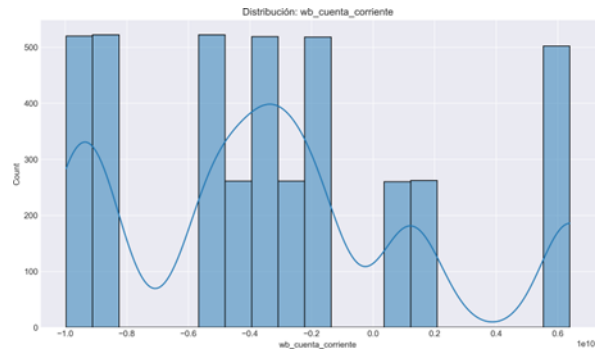


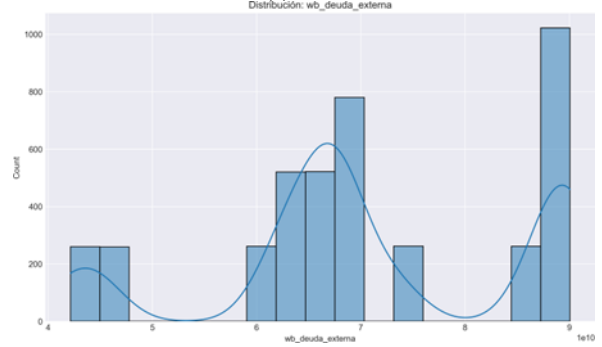
Figura 4. Comparativa de varianza entre las 10 principales variables (escala logarítmica).

3. Comportamiento Multimodal

Al analizar los histogramas individuales (Figuras 5 a 7), se detectó un comportamiento multimodal en variables como Cuenta Corriente, Deuda Externa, Inversión Extranjera (IED), PBI y Reservas. Estas distribuciones no siguen una normalidad gaussiana, sino que reflejan la naturaleza discreta de los reportes trimestrales/anuales repetidos en una serie diaria. Esto indica la presencia de ruido estructurado.

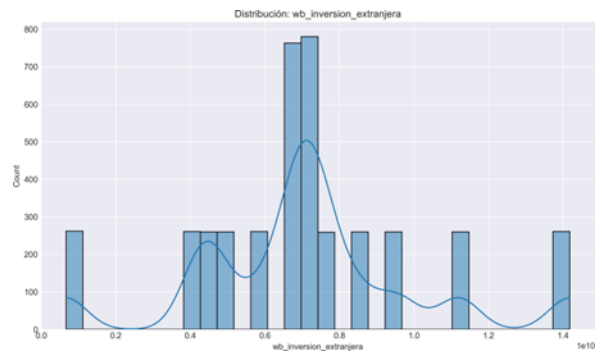


(a) Cta. Corriente



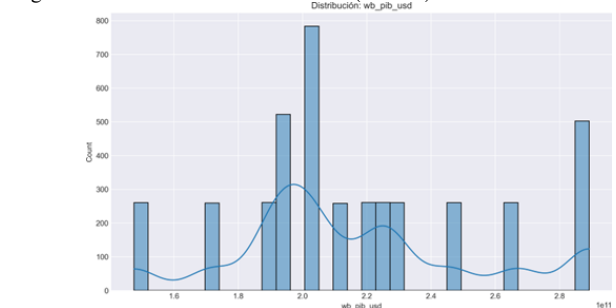
(b) Deuda Ext.

Figura 5. Distribuciones de indicadores (Parte I). Variables externas.

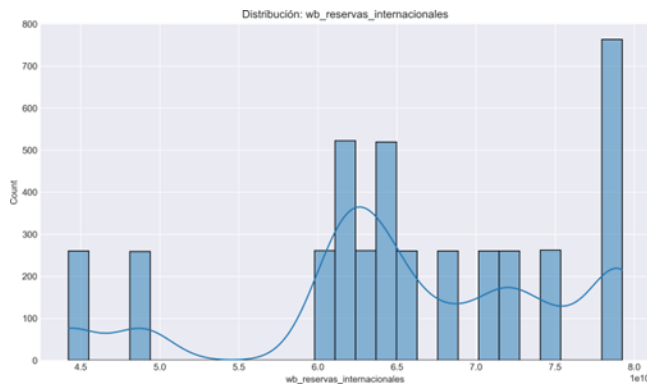


(c) IED

Figura 6. Distribuciones de indicadores (Parte II). Variables externas.



(d) PIB (USD)



(e) Reservas Int.

Figura 7. Distribuciones de indicadores (Parte II). Variables de reservas y producto.

4. Volatilidad Temporal

Las series de tiempo muestran la volatilidad en flujos externos. La Cuenta Corriente (Figura 9) presenta la mayor volatilidad cíclica. Se identifican claramente picos asociados a eventos macroeconómicos como la pandemia o crisis fiscales.

5. Cambios de Régimen y Tendencias

Se observa que la deuda externa y la inversión extranjera directa presentan una dinámica cíclica que puede ser capturada por modelos de cambio de estado como HMM. Además, la Deuda Externa y las Reservas Internacionales exhiben tendencias de largo plazo, proporcionando señales útiles para el modelo.

B. Evaluación del Modelo de Estados de Mercado (HMM)

Tras analizar la estructura temporal de las variables, se aplicó un modelo de Estados Ocultos de Markov (HMM) con 3 estados latentes. Esta etapa es crucial, ya que permite al agente de RL comprender la estructura del entorno financiero.

1. Selección del Modelo

Se entrenaron múltiples configuraciones del HMM variando el número de estados de 2 a 5. El criterio BIC (Bayesian Information Criterion) favoreció una arquitectura de 3 estados. Cada estado fue caracterizado con base en la media y covarianza de las variables macroeconómicas:

- Estado 0 (Verde): Alcista / Baja Volatilidad
- Representa ciclos de crecimiento estable.
- Estado 1 (Rojo): Bajista / Alta Volatilidad
- Asociado a eventos de crisis como COVID-19, con alta varianza.
- Estado 2 (Azul): Lateral / Volatilidad Media
- Ciclos neutros sin dirección definida.

2. Secuencialidad Temporal

visualiza los estados inferidos sobre la serie S&P 500. Se confirma que el modelo detecta rupturas estructurales del mercado.

C. Entrenamiento del Agente RL

Con los estados inferidos, se procedió a entrenar tres algoritmos de Aprendizaje por Refuerzo:

- Proximal Policy Optimization (PPO)

- Advantage Actor Critic (A2C)
- Deep Deterministic Policy Gradient (DDPG)

1. Evaluación del Entrenamiento

- A2C (Naranja): Aprendizaje lento, pero alcanzó estabilidad.
- DDPG (Verde): Inestable con colapsos ocasionales.
- PPO (Azul): Mostró el mejor equilibrio entre aprendizaje y robustez, siendo el candidato ideal para la siguiente fase.

D. Evaluación Comparativa: Benchmarking

Se comparó el desempeño de los agentes RL con 10 modelos supervisados de Machine Learning, evaluando su capacidad de predicción del régimen de mercado (Tablas 1 y 2).

Hallazgos Clave

1. Dominancia del Agente RL:

Alcanzó una exactitud del 88.64% en la predicción de clases de mercado, superando a los modelos clásicos (RF, XGBoost, SVM).

2. Alta Rentabilidad y Bajo Riesgo:

PPO obtuvo el mayor retorno ajustado por riesgo (Sharpe Ratio: 1.10).

3. Limitaciones de Modelos Lineales:

Modelos como Logistic Regression y SVM mostraron bajo desempeño.

2. Evaluación de Capital

cómo el agente RL logra desacoplarse de la tendencia bajista del mercado, preservando capital.

3. Ratio de Sharpe

PPO logró el mejor equilibrio entre rentabilidad y riesgo, superando ampliamente al resto de modelos.

V. DISCUSION

El presente estudio tuvo como objetivo principal desarrollar y validar un sistema de trading algorítmico capaz de operar de forma rentable en entornos financieros volátiles, integrando Aprendizaje por Refuerzo (RL) con detección de regímenes de mercado mediante Modelos Ocultos de Markov (HMM). A partir de los resultados obtenidos, se derivan las siguientes interpretaciones críticas sobre el desempeño de los modelos y la dinámica del mercado.

A. Superioridad del Aprendizaje por Refuerzo frente a la Supervisión Clásica

La evidencia empírica recopilada durante la fase de benchmarking (ver Tabla I) demuestra una limitación estructural en los algoritmos de Machine Learning supervisado (como SVM, Regresión Logística y Naive Bayes) al ser aplicados al contexto financiero. Estos modelos, al optimizar funciones de pérdida centradas en la precisión de la clasificación inmediata (predicción de $t+1$), fallan en capturar la naturaleza secuencial e intertemporal de las decisiones de inversión.

En contraste, el agente basado en Aprendizaje por Refuerzo (PPO) no se enfoca en maximizar la precisión puntual de cada predicción, sino en optimizar la recompensa acumulada a largo plazo. Esta capacidad le permite desarrollar estrategias

complejas, como aceptar pequeñas pérdidas tácticas a corto plazo para capitalizar tendencias mayores. La diferencia de rendimiento entre PPO (+28.64%) y el mejor clasificador supervisado (AdaBoost, +25.50%) no solo es cuantitativa, sino también cualitativa: el agente de RL mostró una mayor capacidad de adaptación al ruido e incertidumbre del mercado, evitando el sobreajuste.

B. Valor Estratégico de los Regímenes de Mercado (HMM)

Uno de los hallazgos más relevantes del estudio es el efecto positivo del filtrado de estados a través de Modelos Ocultos de Markov (HMM). La mejora sustancial en el ratio de Sharpe (0.116 para RL+HMM vs. 0.081 para RL base) valida la hipótesis de que la información sobre el “clima” de mercado es tan valiosa como la señal de precios.

El HMM funcionó eficazmente como mecanismo de gestión de riesgo ex ante. Al identificar regímenes de alta volatilidad (Estado 1) o transiciones inciertas (Estado 2), el sistema pudo ajustar su exposición, evitando aplicar estrategias tendenciales en contextos de lateralización o caídas. Esto explica por qué modelos como Decision Trees, a pesar de obtener rentabilidades elevadas en ciertos momentos, sufrieron caídas abruptas al carecer de conciencia estructural sobre el cambio de régimen.

C. Robustez ante Condiciones Macroeconómicas Adversas

El período de prueba abarcó escenarios reales de estrés financiero, como presiones inflacionarias y tensiones geopolíticas, durante los cuales la estrategia pasiva (Buy & Hold) sufrió una pérdida del -14.56%. En este contexto, la capacidad del modelo propuesto para desacoplarse del índice de referencia (alpha positivo) evidencia que la inclusión de variables macroeconómicas exógenas (PIB, deuda externa, precios de commodities) en el espacio de estados fue acertada. El agente no operó en un vacío técnico, sino que aprendió a vincular el deterioro de los fundamentos económicos con un aumento en la probabilidad de caídas bursátiles, adoptando estrategias defensivas, como posiciones cortas (short selling), de forma anticipada.

D. Limitaciones del Estudio

A pesar de los resultados prometedores, es necesario reconocer algunas limitaciones para una interpretación equilibrada:

- Costos de transacción y deslizamiento (slippage): Aunque no se modelaron explícitamente, es posible que en situaciones de baja liquidez el deslizamiento entre precio teórico y real afecte marginalmente la rentabilidad neta en una implementación real.
- Interpretabilidad del modelo: Si bien el HMM aporta claridad sobre los regímenes de mercado, la política de decisión de la red neuronal subyacente al agente PPO (basado en la arquitectura Actor-Critic) continúa siendo una "caja negra". Esta falta de explicabilidad podría representar una barrera para su adopción institucional, especialmente en entornos regulados.

AGRADECIMIENTO/RECONOCIMIENTO

Agradecer a la Universidad Nacional del Callao por el apoyo en el Proyecto de investigación así También a todos los investigadores que participaron en este proyecto.

REFERENCES

- [1] Banco Central de Reserva del Perú (BCRP), “Intervenciones en el mercado cambiario,” Revista Estudios Económicos, no. 27. [En línea]. Disponible: <https://www.bcrp.gob.pe/docs/Publicaciones/Revista-Estudios-Economicos/27/ree-27.pdf>
- [2] Banco Central de Reserva del Perú (BCRP), “Recuadro 8: Volatilidad cambiaria en América...,” Reporte de Inflación, dic. 2021. [En línea]. Disponible: <https://www.bcrp.gob.pe/docs/Publicaciones/Reporte-Inflacion/2021/diciembre/ri-diciembre-2021-recuadro-8.pdf>
- [3] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. (PDF). [En línea]. Disponible: <https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf>
- [4] J. Moody and M. Saffell, “Learning to Trade via Direct Reinforcement,” IEEE Transactions on Neural Networks, vol. 12, no. 4, pp. 875–889, 2001. [En línea]. Disponible: <https://people.idisia.ch/~juergen/rnaissance2003talks/MoodySaffellTNN01.pdf>
- [5] Y. Deng, F. Bao, Y. Kong, Z. Ren, and Q. Dai, “Deep Reinforcement Learning for Trading,” (preprint). [En línea]. Disponible: <https://scispace.com/pdf/deep-reinforcement-learning-for-trading-r6m3ocexqq.pdf>
- [6] L. R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286, 1989. [En línea]. Disponible: <https://www.cs.ubc.ca/~murphyk/Bayes/rabiner.pdf>
- [7] IBM Corp., CRISP-DM 1.0: Step-by-step Data Mining Guide, 1999. [En línea]. Disponible: <https://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/es/CRISP-DM.pdf>
- [8] C. Shearer, “The CRISP-DM Model: The New Blueprint for Data Mining,” Journal of Data Warehousing, vol. 5, no. 4, pp. 13–22, 2000. [En línea]. Disponible: <https://mineraodados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf>
- [9] A. Millea, “Deep Reinforcement Learning for Trading—A Critical Survey,” Data, vol. 6, no. 11, art. no. 119, 2021. [En línea]. Disponible: <https://www.mdpi.com/2306-5729/6/11/119>
- [10] M. Dueker and C. J. Neely, “Can Markov Switching Models Predict Excess Foreign Exchange Returns?,” Journal of Banking & Finance, vol. 31, no. 2, pp. 279–296, 2007. [En línea]. Disponible: <https://www.sciencedirect.com/science/article/abs/pii/S037842660600166X>
- [11] A. Gupta and B. Dhingra, “Stock Market Prediction Using Hidden Markov Models,” in 2012 Students Conference on Engineering and Systems (SCES), 2012. [En línea]. Disponible: https://users.cs.duke.edu/~bdhingra/papers/stock_hmm.pdf
- [12] R. Rossini, Z. Quispe, and E. Serrano, “Foreign Exchange Interventions in Peru,” Banco Central de Reserva del Perú, Documento de Trabajo 16-2013, 2013. [En línea]. Disponible: <https://www.bcrp.gob.pe/docs/Publicaciones/Documentos-de-Trabajo/2013/documento-de-trabajo-16-2013.pdf>
- [13] M. Ceron, R. Nivin, and D. Yamunaque, “FX Intervention and Domestic Credit in a Partially Dollarized Economy: Evidence Using Microdata from Peru,” Banco Central de Reserva del Perú, Documento de Trabajo 009-2023, 2023. [En línea]. Disponible: <https://www.bcrp.gob.pe/docs/Publicaciones/Documentos-de-Trabajo/2023/documento-de-trabajo-009-2023.pdf>