Proyecto Integrado V - Línea de Énfasis (Entrega 1 y 2).

Julio Cesar Muñoz Rios - 1116436782

Ingeniería de Software y Datos, Institución Universitaria Digital de Antioquia

PREICA2501B020128 – Proyecto Integrado V - Línea de Énfasis

Andrés Felipe Callejas

Mayo de 2025

Título

Automatización de la Recolección Histórica de Datos de NVIDIA: Persistencia, Control de Versiones y Preparación para el Análisis Predictivo

Resumen

Este documento presenta un proyecto orientado a automatizar la recolección continua de datos históricos del comportamiento bursátil de NVIDIA Corporation. Se implementa un sistema que extrae datos desde Yahoo Finance, los limpia, almacena en una base de datos SQLite y los exporta a un archivo CSV versionado, todo bajo un entorno de ejecución automatizada con GitHub Actions. Además, se enriquece la información con indicadores financieros, se desarrolla un modelo predictivo para estimar el precio de cierre del día siguiente y se crea un panel de control para visualizar los KPIs fundamentales. Los resultados del modelo muestran un RMSE de 0.9889, MAE de 0.2538 y R^2 de 0.9985. Esta fase busca garantizar la persistencia, trazabilidad y calidad de los datos, estableciendo la base para futuras etapas centradas en la toma de decisiones basadas en analítica predictiva.

Introducción

NVIDIA Corporation es una de las empresas líderes en tecnología de procesamiento gráfico y computación de alto rendimiento. Fundada en 1993, comenzó a cotizar en bolsa en enero de 1999 bajo el tíker NVDA en el Nasdaq. Su desempeño bursátil ha sido un referente del sector tecnológico y del auge de la inteligencia artificial. La evolución histórica de su precio de acción, volumen de transacciones y otras variables

permite calcular indicadores financieros clave (KPIs) que sirven para evaluar tendencias, volatilidad y retornos.

La automatización de la recolección de esta información es crucial para análisis continuo y reproducible. En esta primera fase se establece la base de datos histórica, se enriquece con métricas financieras, se modela el precio de cierre y se implementa un dashboard interactivo.

Objetivo General

Automatizar la recolección continua de datos históricos del indicador bursátil de NVIDIA Corporation, garantizando su persistencia, trazabilidad y preparación para análisis predictivo y visualización de KPIs.

Objetivos Específicos

Obtener automáticamente datos históricos de NVIDIA desde Yahoo Finance mediante web scraping.

1. Preprocesar y limpiar los datos eliminando inconsistencias y formatos erróneos.
2. Calcular y almacenar KPIs financieros clave.
3. Entrenar y validar un modelo predictivo para estimar el precio de cierre del día siguiente.
4. Desplegar un dashboard interactivo que muestre los KPIs y resultados del modelo.

Metodología

La metodología se articula en cinco fases:

1. Extracción: Uso de requests y BeautifulSoup para capturar datos de Yahoo Finance.
2. Transformación y Limpieza: Conversión de fechas, normalización de formatos numéricos y eliminación de valores faltantes.

3. Persistencia: Almacenamiento en SQLite y CSV bajo control de versiones en GitHub Actions.

4. Enriquecimiento: Cálculo de KPIs con enricher.py.

5. Automatización: Orquestación completa en .github/workflows.

Proceso de Enriquecimiento

El archivo enricher.py realiza las siguientes tareas:

- Convierte la columna Date a formato datetime.
- Limpia las columnas de precios (Open, High, Low, Close) eliminando comas de miles y conservando puntos decimales.
- Normaliza Volume eliminando comas y convirtiendo a entero.
- Calculamos cinco KPIs y los adicionamos a los datos obtenidos en la fase de recolección de datos:

1) **Daily_Return:** Variación porcentual diaria del cierre, que permite medir el rendimiento instantáneo.

2) **MA_7_Close:** Media móvil de 7 días del precio de cierre, útil para suavizar la serie y destacar tendencias.

3) **STD_7_Close:** Desviación estándar móvil de 7 días, indicador de volatilidad reciente.

4) **Cumulative_Return:** Retorno acumulado desde el inicio de la serie, que muestra el crecimiento total.

5) **Momentum_7:** Ratio del precio actual frente al de hace una semana, útil para identificar sobrecompra o sobreventa.

- Redondea cada KPI a 4 decimales para evitar notación científica.
- Guarda el CSV enriquecido en src/static/data/enriched_data.csv.

Proceso de Modelado.

El archivo modeller.py define la clase ModelTrainer con dos métodos:

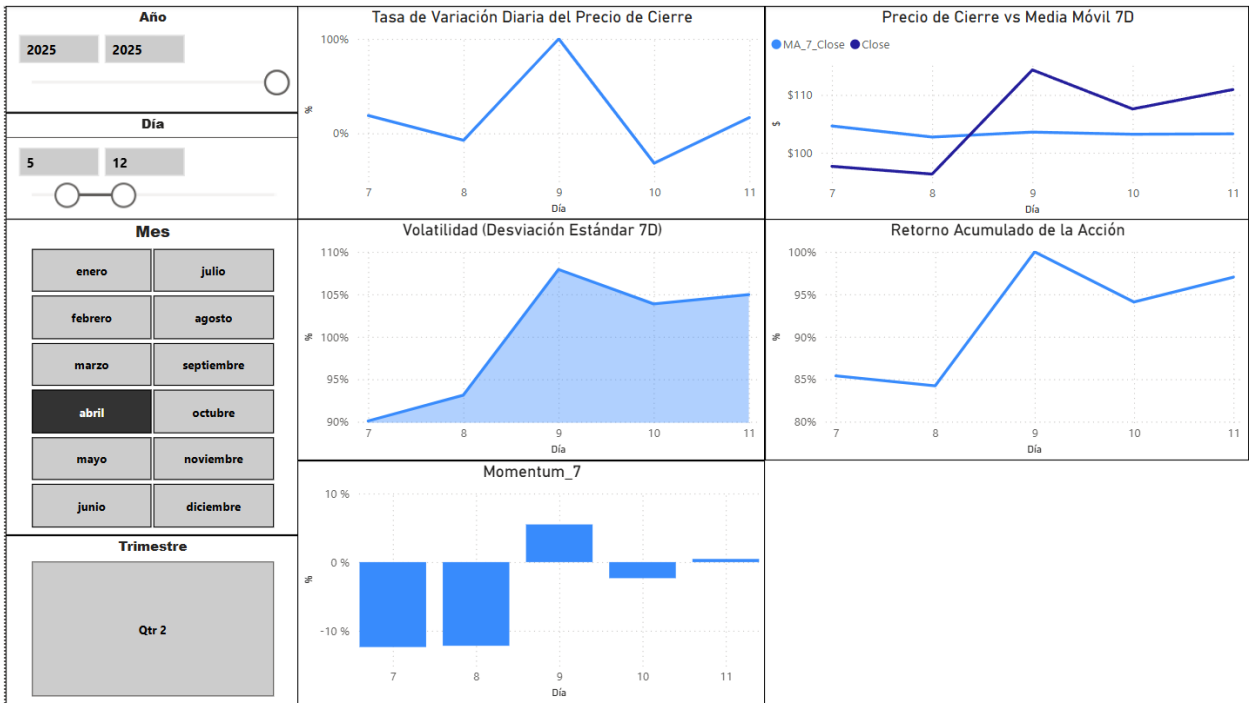
- **entrenar():** Carga enriched_data.csv, ordena por fecha, define las características (Open, High, Low, Volume) y el objetivo (Close del día siguiente), entrena un modelo de regresión lineal, guarda el artefacto en src/static/models/model.pkl y calcula métricas de desempeño (RMSE, MAE, R^2), almacenándolas en src/static/models/metrics.csv.
- **predecir():** Carga el modelo entrenado y genera predicciones sobre un nuevo DataFrame, retornando una serie de precios estimados.

Dashboard y Visualización de KPIs

Se implementó un dashboard en Power BI que consume enriched_data.csv y presenta:

- Tasa de Variación Diaria: Daily_Return para mostrar fluctuaciones diarias.
- Media Móvil (7 días): Para identificar tendencias.
- Volatilidad (STD 7 días): Para observar períodos de alta inestabilidad.
- Retorno Acumulado: Para visualizar el crecimiento histórico.
- Momentum (7 días): Para detectar señales de impulso.

Cada visual incluye segmentadores por año, trimestre, mes y día.



Justificación de la Métrica

Se eligió RMSE como métrica principal porque penaliza errores grandes y mantiene unidades homogéneas (precio). MAE se reporta como complemento por su interpretabilidad directa. R^2 indica la proporción de varianza explicada por el modelo.

Resultados del Modelo

RMSE: 0.9889

MAE: 0.2538

R^2: 0.9985

Estos resultados muestran un ajuste excelente, con errores promedio menores a una unidad monetaria y casi el 99.85% de la varianza explicada.

Conclusión

La integración de los procesos de recolección, enriquecimiento y modelado ha permitido construir un pipeline robusto y reproducible. Los KPIs calculados ofrecen una visión clara del comportamiento bursátil de NVIDIA, y el modelo predictivo demostró alta precisión. El dashboard en Power BI facilita la interpretación para la toma de decisiones.

Referencias

Yahoo Finance. (2025). NVIDIA Historical Data. Recuperado de https://finance.yahoo.com/quote/NVDA/history

Microsoft. (2021). Power BI documentation. Recuperado de https://docs.microsoft.com/power-bi

GitHub, Inc. (2024). GitHub Actions Documentation. Recuperado de https://docs.github.com/en/actions

The Git Project. (2024). Git Documentation. Recuperado de https://git-scm.com/doc

Reback, J., McKinney, W., & Wes McKinney. (2020). pandas: powerful Python data analysis toolkit. Recuperado de https://pandas.pydata.org/

Python Software Foundation. (2024). Python Language Reference. Recuperado de https://www.python.org