# Entrega 1. Deep Learning

Andrea Sánchez Castrillón
Alejandro Vargas Ocampo


Universidad de Antioquia
Facultad de Ingeniería
2025

**Contexto de aplicación.**

NVIDIA Corporation es una de las compañías tecnológicas más influyentes a nivel mundial, reconocida principalmente por el desarrollo de procesadores gráficos (GPU) y chipsets. En agosto de 2025, NVIDIA alcanzó un hito histórico al convertirse en la empresa más valiosa del mundo por capitalización bursátil, con un valor de 4.4 billones de dólares, superando a gigantes tecnológicos tradicionales. La capitalización bursátil, entendida como el valor total de las acciones en circulación de una empresa cotizada, se ha convertido en una medida clave para dimensionar la magnitud de NVIDIA en el mercado financiero global.

Dada esta relevancia, el análisis de sus acciones es de gran interés tanto para inversionistas como para investigadores financieros. En este proyecto se propone abordar el problema de la predicción del precio de cierre de las acciones de NVIDIA como una tarea de series de tiempo, utilizando redes neuronales convolucionales (CNNs). Aunque estas redes son comúnmente aplicadas en visión por computador, han demostrado una alta efectividad en la detección de patrones locales en datos secuenciales, lo que las convierte en una alternativa sólida para el forecasting financiero.

**Objetivo de machine learning**

El objetivo del modelo de *Machine Learning* es clasificar la dirección futura del precio de la acción de NVIDIA (NVDA) a partir de su historial bursátil. Para ello, se utilizarán secuencias de datos diarios desde 1999 hasta 2025, compuestas por las variables *open*, *high*, *low*, *close*, *adj_close* y *volume*. La variable objetivo (target) será una etiqueta binaria que indicará si el precio de cierre del siguiente dia subirá (1) o bajará (0) respecto al valor actual, esta etiqueta se calculará comparando el precio de cierre “close” con el del siguiente dia. Con el fin de capturar patrones tanto locales como temporales en la serie, se empleará una arquitectura basada en redes neuronales convolucionales combinadas con redes recurrentes (CNN-RNN), lo que permitirá detectar estructuras de corto plazo en ventanas temporales y modelar dependencias secuenciales de largo plazo para mejorar la capacidad predictiva del sistema.

**Descripción del dataset**

El dataset utilizado corresponde al historial bursátil de la acción de NVIDIA Corporation (NVDA), con registros diarios desde 1999 hasta 2025. Cada instancia –En total, 6648– contiene siete variables (Tabla 1):

| Variable | Descripción | Tipo de dato |
|---|---|---|
| Date | Fecha de la transacción | Date |
| Open | Precio de apertura | Float |
| High | Precio máximo del día | Float |
| Low | Precio mínimo del día | Float |
| Close | Precio de cierre | Float |
| Adj_close | Precio de cierre ajustado (considerando dividendos y splits) | Float |
| Volume | Volumen total negociado | Int |

Tabla 1. Descripción columnas del dataset. Elaboración propia.

El tamaño en disco aproximado es de 2.88 MB, el dataset no cuenta con valores faltantes.

Se calcula la variable target comparando valores, esta columna es de tipo binario e indica si el precio subió(1) o bajó(0) con respecto al día anterior, una vez elaborado esto, encontramos que la distribución de las clases (tabla 2) es relativamente equilibrada, lo que indica que el modelo no requiere técnicas avanzadas de balanceo y podrá aprender de manera simétrica patrones de subida y bajada del precio.

| Proporcion de clases en Target | | |
|---|---|---|
| Valor | Distribución | Proporción |
| 0 | 3231 | 51,60% |
| 1 | 3452 | 48,30% |

Tabla 2. Proporción de clases en la variable “Target”. Elaboración propia.

En el análisis exploratorio también se evaluaron otros factores, por ejemplo se observa que el volumen de transacciones presenta una distribución relativamente uniforme, sin concentraciones extremas. En cuanto al precio, la evolución de “close” revela una clara tendencia alcista sostenida en el largo plazo, lo que confirma que NVIDIA ha pasado de valores bajos a precios significativamente altos con el tiempo. La matriz de correlación evidencia una relación casi perfecta entre las variables de precio (open, high, low, close y adj_close), por lo que muchas de ellas son redundantes y podrían eliminarse para evitar multicolinealidad, siendo suficiente conservar sólo *close* o *adj_close*.

Finalmente, el gráfico de autocorrelación de precio de cierre (gráfico 1) muestra una correlación muy alta en los primeros lags, cercana a 1, lo que indica que el precio de cierre de NVIDIA está fuertemente influenciado por los valores recientes, es decir, si hoy el precio es alto, mañana también tiende a ser alto. Sin embargo, esta correlación cae rápidamente conforme aumenta el lag, estabilizándose alrededor de 0. Esto significa que los precios pierden memoria en el mediano y largo plazo, volviéndose prácticamente independientes después de cierto número de días, lo cual es una buena noticia para el modelo CNN que tenemos pensado aplicar, ya que estos pueden aprovechar esa dependencia de corto plazo

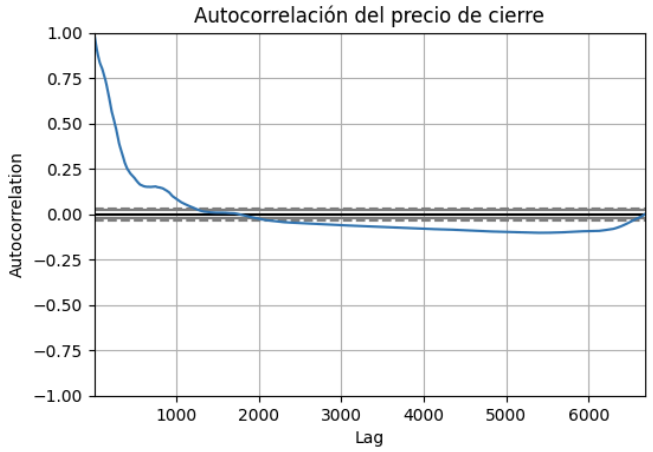


Gráfico 1. Autocorrelación de precio de cierre. Elaboración propia.

**Métricas de desempeño**

Para evaluar el desempeño del modelo CNN–RNN se utilizarán dos métricas:

1. **AUC-ROC**:  permitirá medir la capacidad global del sistema para distinguir entre días alcistas y bajistas independientemente del umbral de decisión, siendo especialmente útil ante posibles desbalances en la etiqueta binaria.
2. **F1-score**: combinará precisión y recall en un único indicador, favoreciendo un equilibrio entre evitar señales falsas y no omitir oportunidades reales.

Estas métricas permitirán validar la capacidad del modelo para distinguir correctamente la dirección del mercado desde una perspectiva estadística y de clasificación. Si bien no garantizan por sí mismas su eficacia operativa en escenarios reales, proporcionan una base sólida para determinar si el modelo posee un desempeño suficientemente robusto como para ser considerado en etapas posteriores de evaluación práctica o integración en sistemas automatizados de apoyo a la toma de decisiones financieras.

**Resultados previos**

Varios participantes abordaron el dataset con enfoques cuantitativos y de aprendizaje automático. Pessoa [1] analizó estrategias de mean reversion mediante cointegración entre acciones del índice MAG7, logrando rentabilidad positiva aunque atribuida al azar; aun así, su trabajo destacó el valor del dataset para evaluar métricas como Sharpe ratio y drawdown. En la línea predictiva, Yadav [2] aplicó clasificación binaria con un 56 % de precisión pero bajo recall para clases negativas, evidenciando limitaciones del enfoque. Waskar [3] comparó modelos de regresión, destacando que la regresión lineal alcanzó un MSE de $1.29\times10^{-5}$ y un $R^2$ de 0.9997, pero aun así concluyó que LSTM fue superior por su mayor estabilidad secuencial (MSE de 0.0019), sugiriendo sobreajuste en el modelo lineal.

Otros trabajos complementan esta perspectiva: Risk [4] obtuvo un RMSE de 0.375 en una predicción sencilla tras un análisis exploratorio, mientras que Delong [5] identificó que el precio presenta una tendencia logarítmica frente a un volumen estable, resultando en un crecimiento progresivo del daily dollar volume. En conjunto, los resultados previos indican que el dataset es especialmente útil para estrategias basadas en series temporales profundas, reversión a la media y análisis de liquidez.

**Referencias**

[1] S. Pessoa, “Conclusion (Mean Reversion & Cointegration study on MAG7 stock),” *Kaggle Notebook*, 2024.

[2] R. Yadav, “SLT-Final_project,” *Kaggle Notebook*, 2024.

[3] S. Waskar, “Stock Prediction LSTM, Linear regression, Xgboost,” *Kaggle Notebook*, 2024.

[4] B. Risk, “NVIDIA Stock Trends and Predictions 1999–2024,” *Kaggle Notebook*, 2024.

[5] M. Delong, “NVDA price and volume EDA with scatter plots,” *Kaggle Notebook*, 2024.