Unidad 2

Actividad:

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

Julio César Cárdenas Veloth

Módulo – Proyecto integrado V

Grupo PREICA2501B020128

Profesor

Andrés Felipe Callejas

Institución Universitaria Digital de Antioquia

Ingeniería de Software y Datos

Medellín

2025

1 INTRODUCCIÓN

De acuerdo con el origen de datos seleccionado para la realización de todo el proceso de captura , transformación y enriquecimiento de datos financieros, obtenidos en específico de la página web https://es.finance.yahoo.com/quote/000001.SS/history/, en este segundo documento se realiza la definición de 5 Kpi que permitirán ayudar a la toma de decisiones a las personas interesadas en realizar algún tipo de operación financiara en la bolsa de Shanghái, de igual forma se construye un modelo que ayuda a predecir cual sería el comportamiento del precio de cierre al día siguiente del índice de Shanghái.

2 OBJETIVO

Demostrar competencias en enriquecimiento de datos, desarrollo y validación de un modelo predictivo desacoplado, y construcción de un dashboard interactivo que permita analizar resultados de forma efectiva

3 METODOLOGIA

Metodología para la descarga automatizada de datos desde Yahoo Finanzas

3.1 Definición del alcance y fuente de datos

Se seleccionó como objeto de estudio el índice de Shanghái (https://es.finance.yahoo.com/quote/000001.SS/history/) y su histórico de precios y volúmenes, disponible en la URL indicada:

El período de interés abarca del 9 de mayo de 2024 al 9 de mayo de 2025 para el momento de la realización de esta documentación..

3.2 Configuración del entorno de Big Data y automatización

Se emplea Python como lenguaje de scripting, con entornos virtuales para gestionar dependencias.

Se instalan bibliotecas clave:

Selenium para controlar un navegador en modo “headless” y reaccionar ante elementos dinámicos (botón de cookies, carga de tablas).

BeautifulSoup para extraer el HTML resultante.

Pandas para estructurar y transformar los datos tabulares.

SQLite como gestor de base de datos ligero y escalable para series temporales.

3.3 Extracción

La función de extracción se automatizó mediante un script de Python con Selenium, el funcionamiento se explica a continuación:

El script abre la página en un navegador controlado por Selenium, asegurando la ejecución de JavaScript y la carga completa del contenido.

Se detecta y pulsa automáticamente el botón de “Aceptar cookies” si aparece, garantizando acceso sin interferencias.

Se espera a que el elemento <table> con el histórico esté presente en el DOM y se realiza la captura de cada uno de los campos.

3.4 Transformación

Se captura el HTML con la tabla completa y se envía a BeautifulSoup para extraer cabeceras y filas.

Las cabeceras se limpian de etiquetas anidadas y caracteres especiales, normalizándose a ASCII y reemplazando espacios por guiones bajos.

Las fechas en español (p. ej. “30 abr 2025”) se descomponen en día, mes y año; el mes se traduce a su número correspondiente y la fecha se convierte a formato ISO (YYYY-MM-DD).

Los valores numéricos eliminan separadores de miles, convierten coma decimal a punto y se escriben adecuadamente: precios como float y volúmenes como int.

3.5 Carga incremental

Se conecta a la base SQLite (historical.db) y se obtienen las fechas ya almacenadas.

Solo los registros con fecha nueva se insertan en modo append, evitando recargar datos anteriores y garantizando un delta incremental.

Se genera además un CSV (historical.csv) con todo el conjunto de datos, listo para uso en otros entornos.

3.6 Definición de los Kpi

Con el objetivo de que se puedan realizar unos análisis de mercado con mayor profundidad, a continuación se define cada uno de los Kpi seleccionados y su justificación.

- Retorno

  Definición: Cambio porcentual diario sobre el cierre anterior.

  Justificación: Mide la rentabilidad diaria y permite comparar variaciones independientemente del nivel absoluto.

- Volatilidad_30d

  Definición: Desviación estándar anualizada de los retornos en una ventana de 30 días.

  Justificación: Captura la variabilidad reciente del índice y la anualiza para compararla con otros benchmarks de riesgo.

- SMA_50d
  Definición: Media simple de los precios de cierre de los últimos 50 días.

  Justificación: Suaviza el ruido de corto plazo y ayuda a identificar la tendencia dominante.

- Volumen_20d_avg

  Definición: Volumen medio de los últimos 20 días.

  Justificación: Establece un punto de referencia para determinar si el volumen de un día es excepcional en relación con la actividad reciente.

- Volume_Ratio

  Definición: Razón entre el volumen del día y su media a 20 días.

  Justificación: Destaca días con flujo de transacciones inusualmente alto o bajo, señalando posibles movimientos de precio.

- Drawdown

  Definición: Caída porcentual desde el máximo histórico acumulado hasta la fecha.

  Justificación: Mide la magnitud de las pérdidas desde picos previos, informando sobre el riesgo en retrocesos.

3.7 Enriquecimiento y transformación

Con los Kpi definidos se procedió a desarrollar el scripts (función) enricher.py, la cual se llama desde el scripts collector.py y se encarga de adicionar las nuevas columnas para los Kpi y su respectivo cálculo.

Con este mismo script se adiciona una columna que se encarga de calificar el comportamiento del índice bursátil basado en los Kpi.

Retorno	Volatilidad_30d	SMA_50d	Volumen_20d_avg	Volume_Ratio	Drawdown	Senal_Mercado
0.0009154273	0.1015681858	[NULL]	284,635	0.9450700019	-0.0621667219	Baja volatilidad; Volumen normal
0.0007700041	0.1016274451	[NULL]	283,995	0.9183260269	-0.0614445863	Baja volatilidad; Volumen normal
-0.0045190337	0.1008452467	[NULL]	285,045	0.9903699416	-0.0656859499	Baja volatilidad; Volumen normal
0.0048196838	0.1001740933	3,041.5732	284,170	0.9782876447	-0.0611828516	Bajista; Baja volatilidad; Volumen normal
0.0017399307	0.0995484573	3,038.133	285,465	0.9840085475	-0.0595493748	Bajista; Baja volatilidad; Volumen normal
-0.0060657678	0.1006205944	3,034.3264	284,465	0.9797338864	-0.06525393	Bajista; Baja volatilidad; Volumen normal
-0.0164798834	0.1086833726	3,029.6734	285,620	1.0373923395	-0.0806584362	Bajista; Baja volatilidad; Volumen normal
-0.0046031893	0.1082972682	3,024.797	286,805	1.0672756751	-0.0848903395	Bajista; Baja volatilidad; Volumen normal
-0.0052413033	0.1087809627	3,020.1338	287,010	0.9567610885	-0.0896867067	Bajista; Baja volatilidad; Volumen normal
0.0014410719	0.1088311764	3,015.5038	285,170	0.9762597749	-0.0883748798	Bajista; Baja volatilidad; Volumen normal
0.0003251583	0.1083056052	3,010.26	283,125	0.9073730684	-0.0880784573	Bajista; Baja volatilidad; Volumen normal
-0.0043363395	0.107027416	3,004.423	282,180	0.9302572826	-0.0920328587	Bajista; Baja volatilidad; Volumen normal
0.0206473796	0.1247498581	3,000.0386	290,500	1.4206540448	-0.0732857165	Bajista; Baja volatilidad; Volumen elevado
-0.0021641855	0.1244201113	2,995.5156	294,115	1.1784506061	-0.0752912981	Bajista; Baja volatilidad; Volumen normal
-0.0092245574	0.126735214	2,991.2946	296,110	1.054337915	-0.0838213266	Bajista; Baja volatilidad; Volumen normal
-0.015364811	0.1298360817	2,986.7312	300,010	1.1849605013	-0.0978982388	Bajista; Baja volatilidad; Volumen normal
0.0023001363	0.1298211826	2,981.596	299,120	0.9939154854	-0.0958232818	Bajista; Baja volatilidad; Volumen normal
0.0008893446	0.1274109299	2,976.8012	297,200	0.8852624495	-0.0950191571	Bajista; Baja volatilidad; Volumen normal
0.0000243917	0.1252551504	2,971.9788	295,235	0.9792199434	-0.0949970831	Bajista; Baja volatilidad; Volumen normal
-0.0026865048	0.1229288856	2,967.389	294,265	0.9661359659	-0.0974283777	Bajista; Baja volatilidad; Volumen normal
-0.0013905436	0.1189219747	2,962.817	293,025	0.8333759918	-0.0986834429	Bajista; Baja volatilidad; Volumen normal
0.0034077272	0.1196098264	2,958.6062	291,570	0.7946633741	-0.095612002	Bajista; Baja volatilidad; Volumen normal
-0.0060321833	0.1199436475	2,953.7952	288,350	0.7556788625	-0.1010674361	Bajista; Baja volatilidad; Volumen normal
0.0093697929	0.1221125851	2,950.0344	288,045	0.9439497301	-0.0926446242	Bajista; Baja volatilidad; Volumen normal
0.0007194095	0.1220892172	2,946.6472	287,185	0.9182234448	-0.0919918642	Bajista; Baja volatilidad; Volumen normal
0.0049454232	0.1204076584	2,943.495	286,335	0.9139644123	-0.0875013796	Bajista; Baja volatilidad; Volumen normal
-0.0093341673	0.1166753903	2,940.2674	284,705	0.9262218788	-0.0960187944	Bajista; Baja volatilidad; Volumen normal
-0.0035162872	0.1156658648	2,936.6496	280,860	0.8160649434	-0.099197452	Bajista; Baja volatilidad; Volumen normal

3.8 Modelo de predicción

Para realizar la modelación de predicción se utilizó la predicción del precio del cierre, el proceso parte de la idea de que el precio de cierre de hoy contiene gran parte de la información necesaria para anticipar el cierre de mañana. Primero, se recopilan en orden cronológico todos los cierres históricos del índice de Shanghái y, a partir de ellos, se crea una “tarea” para el modelo: predecir el siguiente cierre conociendo únicamente el cierre anterior.

Para enseñarle al modelo, los datos se dividen en dos grupos: uno para que aprenda (entrenamiento) y otro para comprobar qué tan bien lo ha aprendido (prueba). El modelo que se usa, una recta ajustada entre el precio de hoy y el precio de mañana, intenta minimizar la diferencia entre sus predicciones y los valores reales.

Al final, se mide qué tanto se ha acercado el modelo a la realidad usando tres indicadores sencillos:

Error medio absoluto (MAE): cuánto se equivoca, en promedio, el modelo cada día.

Error cuadrático medio (RMSE): qué tan frecuentes y grandes son los errores más importantes.

Coeficiente de determinación (R^2): qué parte de la variación diaria del cierre logra explicar.

Además, para saber si el modelo realmente aporta valor, se compara con una estrategia muy simple que asume que el cierre de mañana será exactamente igual al de hoy. Si el modelo no mejora esa suposición básica, significa que no está captando ningún patrón adicional.

Por último, una vez entrenado y validado, el modelo se guarda en un archivo. De ese modo, cada vez que se necesite una predicción, basta con cargarlo y pasarle el último precio de cierre para obtener inmediatamente la estimación del siguiente día, sin tener que repetir todo el proceso de entrenamiento.

De forma simultanea se corrieron dos modelos con diferentes metodologías y se obtuvieron los siguientes resultados:

=== Evaluación del modelo 1===

MAE modelo: 22.8550

RMSE modelo: 31.7024

R2 modelo: 0.9726

=== Predictor ingenuo 2 ===

MAE naive: 22.2957

RMSE naive: 31.4058

R2 naive: 0.9731

Los resultados muestran que el modelo lineal no está superando la estrategia más simple de “persistencia” (predecir que mañana cierra igual que hoy):

MAE modelo vs MAE naive: 22.8550 vs 22.2957.

El modelo comete un error medio ligeramente mayor (~0.56 puntos) que la predicción ingenua.

RMSE modelo vs RMSE naive: 31.7024 vs 31.4058.

De nuevo, la regresión lineal arroja desviaciones cuadráticas medias un poco peores que la técnica trivial.

R^2 modelo vs R^2 naive: 0.9726 vs 0.9731.

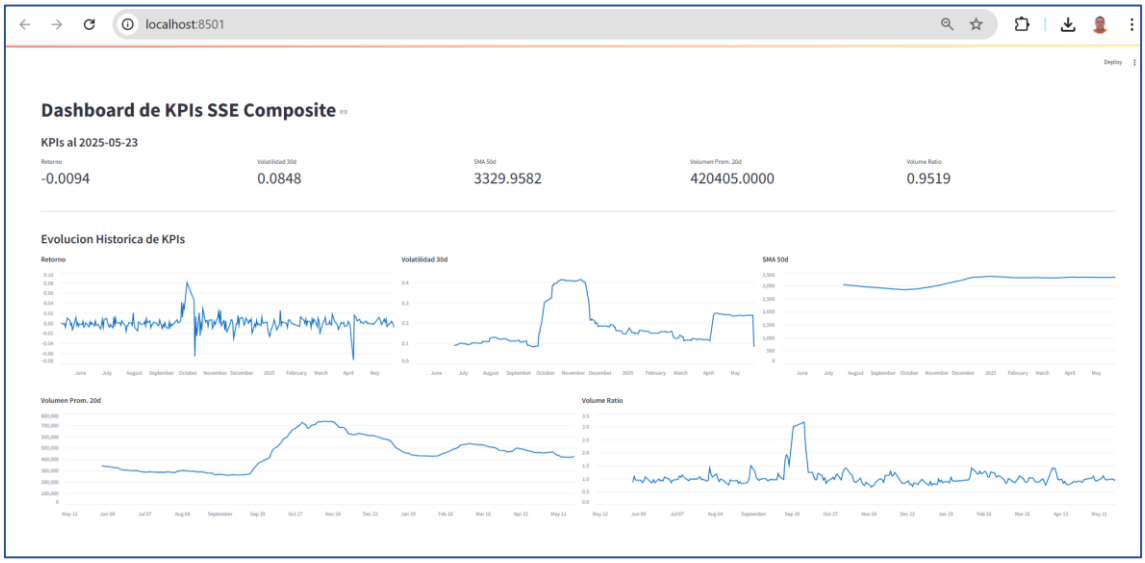
El coeficiente de determinación cae del 97.31 % al 97.26 % cuando uso el modelo frente al naive.

Lo anterior indica que, con solo el cierre de hoy como variable, la regresión lineal no mejora realmente la predicción frente a “mañana será igual que hoy”. En finanzas cortoplacistas, el precio tiende a comportarse de forma muy autocorrelacionada, de modo que el enfoque de persistencia actúa como un benchmark muy duro de superar.

Con un MAE de ~22.86 y un RMSE de ~31.70, el modelo está cometiendo un error medio diario de entre 23 y 32 puntos sobre un índice que ronda los 3300–3400 puntos. Eso equivale a un error relativo del 0.7 %–1 %, que, para un modelo tan sencillo (regresión lineal univariante), puede considerarse aceptable como línea de base.

3.9 Dashboard

Para realizar el Dashboard se creó un scripts en Python que utiliza la librería streamlit, con fue posible realizar las gráficas del comportamiento de los Kpi y generar un tablero que se despliega en un navegador Web como se observa en la siguiente imagen.



Debido a que la operación que se realiza desde github actión no permite realizar el despliegue directamente a un navegador Web, se implementó el guardado de la página completa en la carpeta del repositorio con el nombre de src\SSE_Composite\artifacts.

4 Conclusiones y recomendaciones

La implementación de un flujo automatizado que abarca la captura, normalización y enriquecimiento de los datos históricos del índice de Shanghái—con métricas como retorno, volatilidad, medias móviles, ratios de volumen y drawdown—y la incorporación de modelos predictivos evaluados contra benchmarks ingenuos, ha optimizado la eficiencia operativa, elevado la precisión y consistencia de las decisiones financieras, y garantizado la escalabilidad y reproducibilidad del proceso; además, al centralizar la información en una base de datos y habilitar la integración continua, se facilita el monitoreo en tiempo real y la generación ágil de insights, consolidando una infraestructura de Big Data robusta que potencia el análisis estratégico del mercado.

5 Bibliografía

Yahoo Finanzas. (s. f.). Historial de cotizaciones de 000001.SS – Índice de Shanghái Composite. Recuperado 11 de mayo de 2025, de https://es.finance.yahoo.com/quote/000001.SS/history/