

Unidad 3

Actividad:

Proyecto Integrado V - Línea de Énfasis (Entrega 3).

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 financiera 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, finalmente se realiza una revisión y ajuste de los KPI.

2 OBJETIVO

Integrar los resultados finales del proyecto y comunicar su funcionamiento y hallazgos a través de un video explicativo, demostrando claridad técnica y síntesis.

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 Shanghai (<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.

- Media Móvil Exponencial 20 días (EMA_20d) (KPI adicional)

Definición: Media móvil que da más peso a los datos recientes, calculada con un “span” de 20 días.

Justificación:

La EMA reacciona más rápido que la SMA ante cambios recientes de precio. Al incorporar la EMA-20d complementas la visión de tendencia media (SMA-50d) con una señal más sensible al momentum de corto

plazo. Esto ayuda a detectar aceleraciones o desaceleraciones tempranas antes de que la SMA lo refleje.

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.

		Fecha	Abrir	Max	Min	Cerrar	Cierre_ajustado	Volumen	Retorno	Volatilidad_30d	SMA_50d	EMA_20d	Volumen_20d_avg	Volume_Ratio	Drawdown	Senal_Mercado
1	2	2024-06-13	3038.09	3040.4	3022.62	3028.92	3028.92	297900				3028.92			0.0	
	3	2024-06-14	3020.96	3037.91	3011.58	3032.63	3032.63	348700	0.001224659025699386			3029.2733333333335			0.0	
4	4	2024-06-17	3017.72	3026.62	3012.04	3015.89	3015.89	295800	-0.0055199612217778915			3027.9967301587303			-0.0055199612217778915	
5	5	2024-06-18	3015.07	3031.98	3012.52	3030.25	3030.25	273600	0.004761446869746555			3028.2131368102796			-0.0007847973541118135	
6	6	2024-06-19	3029.1	3030.23	3015.33	3018.05	3018.05	261300	-0.004026070456232933			3027.2452190168246			-0.004807708160903219	
7	7	2024-06-20	3015.72	3028.51	3001.77	3005.44	3005.44	295500	-0.004178194529580348			3025.1685314932234			-0.008965815150545913	
8	8	2024-06-21	2998.24	3011.91	2985.93	2996.14	2996.14	255000	-0.002428928875638947			3022.5943856367253			-0.011372966698872014	
9	9	2024-06-24	2982.83	2990.64	2960.13	2963.1	2963.1	298700	-0.01168724609257743			3016.9282536713226			-0.022927294130836995	
10	10	2024-06-25	2962.76	2971.65	2935.54	2950.0	2950.0	273200	-0.004421045526644352			3010.5541342740535			-0.02724697704632616	
11	11	2024-06-26	2943.45	2976.67	2933.33	2972.52	2972.52	282400	0.007633896305084763			3006.9318357717625			-0.019821079393134055	
12	12	2024-06-27	2962.7	2963.06	2944.9	2945.85	2945.85	270500	-0.008972185216583939			3001.1145180792137			-0.028615426214210173	
13	13	2024-06-28	2940.3	2965.47	2939.98	2967.4	2967.4	315200	0.007315375867746221			2997.9036115954707			-0.0215509382944836664	
14	14	2024-07-01	2965.25	2996.41	2961.86	2994.73	2994.73	297800	0.008210062900855896			2997.6013628721			-0.012487403244048924	
15	15	2024-07-02	2980.86	3004.99	2989.47	2997.01	2997.01	281400	0.000761337416060881			2997.5450425855664			-0.011745580568681272	
16	16	2024-07-03	2995.66	2998.54	2978.07	2982.38	2982.38	246300	-0.004881531926620393			2996.1007528272744			-0.01656977606895647	
17	17	2024-07-04	2984.83	2995.57	2955.29	2957.57	2957.57	274300	-0.008318859844411463			2992.4311573199147			-0.02475078386539075	
18	18	2024-07-05	2954.37	2957.11	2920.21	2949.93	2949.93	272300	-0.0025832017500855287			2988.383428051351			-0.02272705952144715	
19	19	2024-07-08	2939.84	2947.61	2917.23	2922.45	2922.45	277500	-0.009315475282464347			2982.1040539512223			-0.03633150104035121	
20	20	2024-07-09	2919.57	2963.78	2904.04	2959.37	2959.37	315100	0.012633235912417709			2979.93895955868			-0.0243157249647995375	
21	21	2024-07-10	2940.03	2958.86	2935.33	2939.36	2939.36	301500	-0.0067615742539796075			2976.0742482457854	286700.0	1.0516219044297175	-0.030735482864708164	Volumen normal
22	22	2024-07-11	2957.25	2971.36	2947.4	2970.39	2970.39	328400	0.010556719830167038			2975.532891269996	288225.0	1.13938763119091	-0.020523440050385387	Volumen normal
23	23	2024-07-12	2965.61	2977.14	2963.3	2971.29	2971.29	303700	0.0030329905163969282			2975.1288063871193	285975.0	1.061980942390089	-0.020226667941687524	Volumen normal
24	24	2024-07-15	2963.85	2977.31	2959.13	2974.01	2974.01	269000	0.0009154273059850393			2975.022253397888	284635.0	0.9450700019322993	-0.01932973667894551	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 Shanghai 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.

Coefficiente 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 simultánea 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.

R² modelo vs R² naive: 0.9726 vs 0.9731.

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

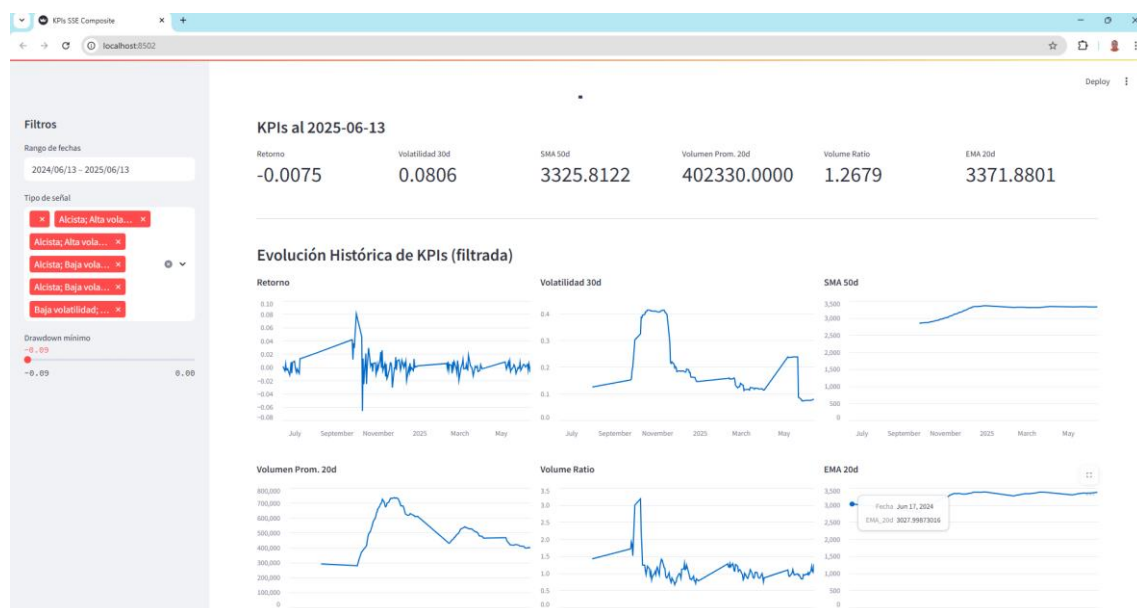
Lo anterior indica que, con solo el cierre de hoy como variable, la regresión lineal no mejora realmente la predicción frente a “mañana será igual que hoy”. En finanzas cortoplacistas, el precio tiende a comportarse de forma muy

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 acción 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/>