TRANSFORMACIÓN, ENRIQUECIMIENTO Y MODELADO DE DATOS HISTÓRICOS BTC-EUR

LAURA ARBOLEDA GALLEGO

GIORDAN JESE RICARDO PARRA

PROYECTO INTEGRADO V

ANDRÉS FELIPE CALLEJAS

INGENERÍA DE SOFTWARE Y DATOS

IU DIGITAL DE ANTIOQUIA

MEDELLÍN

2025

Resumen

Este proyecto consiste en la creación de un sistema automatizado en la recolección y almacenamiento de datos históricos del precio del Bitcoin en Euros, desde su creación el 14 de septiembre del 2014 hasta que nuestros días. Estos datos son obtenidos de Yahoo Finanzas y para su estudio se utilizaron herramientas como Python y la implementación de técnicas como el *Web Scraping*. El proceso de datos se desarrolló mediante Pandas y la información fue almacenada en dos formatos diferentes: CSV y en una tabla de base de datos SQLite. Como resultado de esta práctica, se desarrolló una aplicación funcional capaz de consultar y almacenar datos financieros, todo esto integrado a un Main que es capaz de correr mediante flujos automatizado, escalando a entornos de producción y ejecución, como por ejemplo, en la plataforma de GitHub Actions.

Introducción

En el presente trabajo, vamos a presentar un sistema automatizado de recopilación, almacenamiento y procesamiento de datos históricos del precio del Bitcoin (BTC) en euros (EUR), utilizando el lenguaje de programación Python. La motivación principal de este proyecto se centra en el análisis de datos financieros en tiempo real, recopilando la información para el desarrollo de balances que nos permitan tener un acercamiento con fines de obtener resultados que amplíen más la información, encontrando estrategias y obteniendo respuestas más acertadas en referencia al cómo se mueve este sistema financiero en estos días.

Metodología

La metodología aplicada en esta primera parte del proyecto se basó en el desarrollo de un script automatizado que permite la extracción, transformación y almacenamiento de datos históricos de BTC-EUR. Este enfoque se dividió en las siguientes etapas:

- **Recolección de datos:** Se utilizó la biblioteca **request** para hacer una solicitud HTTP al sitio web de Yahoo Finanzas, simulando un navegador mediante un encabezado **User-**

Agent. Posteriormente, se usó **BeautifulSoup** para analizar los datos del HTML de la página y localizar la tabla con los datos históricos.

- **Procesamiento de datos:** Una vez los datos son extraídos, comienza su procesamiento y estructuración en un dataframe en **Pandas,** donde se renombran las columnas y se validan los registros, de manera tal que la tabla no tenga datos nulos, incompletos o no estructurados completamente.
- **Registro de eventos:** Se implementó un sistema de **logging** personalizado para monitorear la ejecución del programa, registrando eventos relevantes, como lo son el inicio de clases, los errores de conexión o la creación exitosa de los archivos en la carpeta data.
- **Almacenamiento:** Los datos procesados se guardaron en dos formatos. **CSV** y **SQLite.**
- **Estructura modular del código:** Se organizaron las funcionalidades del código mediante clases independientes, las cuales son **Logger, Collector** y **main,** permitiendo la modularidad que facilita la mantenibilidad, pruebas y futuras mejoras. El código en general se corre mediante el main, creando una carpeta de **Logs** donde se nos va a mostrar los datos de conexión, también se va a generar automáticamente la carpeta **static** y **data,** donde se van a guardar los archivos **CSV** y **SQLite.**

Transformación y enriquecimiento

En este paso, ya con los datos históricos recolectado mediante Yahoo Finanzas, procederemos con un proceso de enriquecimiento a los datos que ya tenemos para poder aumentar la productividad y el análisis al momento de desarrollar el modelado de datos y las predicciones de los valores. Este enriquecimiento se verá reflejado en un dashboard ejecutado en Power Bi, donde mediante KPI podremos ver el comportamiento del bitcoin en función al estado del euro. Para esto se añadieron las siguientes columnas:

- **Retorno logarítmico diario (retorno_log_diario):** Se encarga de calcular el cambio

porcentual del precio de cierre entre un día y el siguiente a través del uso de algoritmos. Este recurso sirve para análisis financieros, ya que captura los cambios relativos de manera simétrica y reduce el efecto de valores extremos.

- **Media móvil de 7 días (media_movil_7d):** Se encarga de calcular el promedio de los precios de cierre de los últimos siete días, suavizando las fluctuaciones diarias e identificando tendencias a corto plazo.

- **Media móvil de 30 días (media_movil_30d):** Se encarga de calcular el promedio de precios de cierre en los últimos 30 días, se utiliza para la observación de tendencias a un largo plazo.

- **Volatilidad de 7 días (volatilidad_7d):** Se encarga de estudiar el promedio de siete días, detectando si hay una dirección alta o baja en el mercado. Lo que hace la volatilidad es medir riesgos, mide cuánto varían los precios cada día respecto al promedio, detentando si el mercado permanece estable o si es muy cambiante.

- **Volatilidad de 30 días (volatilidad_30d):** Se encarga de estudiar el promedio de treinta días, detectando si hay una dirección alta o baja en el mercado. Lo que hace la volatilidad es medir riesgos, mide cuánto varían los precios cada día respecto al promedio, detentando si el mercado permanece estable o si es muy cambiante.

Modelado de los datos

El código está usando un **modelo de regresión,** que permite predecir el valor de cierre del bitcoin en euros en la columna **cerrar,** mediante características técnicas y macroeconómicas. En este caso, el modelo que se está utilizando es **Random Forest Regressor** (Bosques Aleatorios de Regresión), es un modelo de aprendizaje supervisado que se basa en árboles de decisión y sirve para predecir un valor numérico (Regresión) combinando muchos árboles diferentes (Determinados como "El bosque"), para tener predicciones más precisas y estables.

El modelo predice los valores de la columna **cerrar,** los cuales sin el precio de cierre del

bitcoin para cada día. El modelo usa también las columnas del log diario, media móvil de 7 días, media móvil 30 días, volatilidad de siete días y volatilidad de treinta días. El modelo evalúa dos métricas de error:

- **RMSE (Root Mean Squared Error):** Penaliza los errores grandes, es buena para saber qué tan lejos están las predicciones en promedio.
- **MAE (Mean Absolute Error):** Es el promedio de los valores absolutos, haciéndolos más fáciles de interpretar. Si por ejemplo el MAE es de 480, significa que en promedio el modelo se puede equivocar por 480 euros.

Observaciones de las predicciones

Precio de cierre real VS predicción

a. **Alta correlación entra la predicción y los resultados:** Las predicciones se muestran demasiados precisas, siguiendo la dirección general del mercado, aunque suaviza ciertas fluctuaciones que pueden considerarse más extremas. No obstante, llama la atención tanta precisión.

b. **Volatilidad inherente:** Se reflejan picos que pueden significar un comportamiento volátil en cómo el bitcoin dentro del euro va a teniendo ciclos donde puede estar en auge o muestra una caída.

c. **Tendencia alcista a largo plazo:** Pese a las caídas, la dirección general del precio permanece en ascenso, lo que da una visión optimista en el mercado a un largo plazo.

d. **Utilidad de la predicción:** Puede ser útil para identificar tendencias generales, pero no necesariamente para prever movimientos a corto plazo debido a la alta volatilidad.

e. **Crecimiento exponencial (2020-2024):** A partir de mediados de 2020, el precio se

dispara con varios picos y correcciones. Este comportamiento coincide con el auge del interés institucional y la adopción masiva de criptomonedas.

Serie diferenciada de cierre_ajustado

a. **Naturaleza de la serie:** La serie mostrada es diferenciada, lo que significa que se ha transformado para analizar los cambios entre periodos consecutivos del precio ajustado, en lugar del precio en sí.

b. **Comportamiento visual:** La serie fluctúa alrededor de cero, lo cual es típico en una serie diferenciada. No se observan tendencias claras ni patrones estacionales evidentes, lo que sugiere que la diferenciación ha sido efectiva.

c. **Estacionariedad confirmada:** El valor del estadístico ADF (-12.48) es mucho menor que todos los valores críticos (1%, 5%, 10%). El p-value extremadamente bajo (de < 0.05) indica que rechazamos la hipótesis nula de que la serie tiene una raíz unitaria, lo que significa que no es estacionaria.

Comportamientos mensuales y semanales

a. **Ciclo semanal:** Se menciona que presenta más ruido tanto en la tendencia como en los residuos, lo cual sugiere que los datos semanales tienen mayor variabilidad y menos claridad en los patrones subyacentes. Por tanto, es probable que factores de corto plazo afecten más este ciclo.

b. **Ciclo mensual:** Se indica que revela mejor la tendencia y muestra una estacionalidad más estable. Lo cual implica que al agrupar los datos mensualmente, se suavizan las fluctuaciones y se destacan mejor los patrones de largo plazo. Por otro lado, la estacionalidad mensual puede reflejar comportamientos recurrentes más claros, como ciclos económicos o patrones de inversión.

c. **Recomendación:** Si el objetivo es construir modelos predictivos o entender el comportamiento estructural del mercado, es preferible trabajar con datos mensuales.

Predicciones ARIMA

a. **Serie Original (Línea azul sólida):** Representan los datos reales desde el 2015 hasta el año 2025. Muestra una evolución con alta volatilidad y varios picos, especialmente en los últimos años.

b. **Predicción ARIMA (1,1,1) (Línea naranja discontinua):** Contiene un componerte autorregresivo (AR), una diferenciación para hacer la serie estacionaria y un componente de media móvil. La línea, por tanto sigue la tendencia de la serie original pero suavizando ciertas fluctuaciones externas, captando bien la tendencia original sin reproducir picos, lo que la hace ideal para las series estacionarias. Por esto, se puede asumir que la serie original no era estacionaria y su transformación fue adecuada.

Predicciones SARIMA

a. **Serie original (Línea azul):** Representa los valores reales desde el 2016 hasta el 2024. Se puede observar una evolución con variaciones notables, mostrando cierta estacionalidad.

b. **Predicción SARIMA (Línea naranja):** El modelo SARIMA incluye componentes estacionales, lo que lo hace adecuado para series con patrones repetitivos a lo largo del tiempo.

c. Se puede ver un intento del modelo SARIMA en modelar patrones estacionales, lo cual puede ser útil porque el modelo traba con ciclos anuales y mensuales. También muestra un ajuste moderado, aunque no lo logra tan bien como lo hace el modelo ARIMA, mostrando métricas de error más altas. Este aspecto se puede mejorar ajustando los

parámetros estacionales.