FASE 0 - PROYECTO DE BIG DATA E INTELIGENCIA ARTIFICIAL

Curso de Especialización en IA y Big Data
Curso académico 2025/2026

Alumnos:

Cristian Cantero López (cristiancanterolopez@gmail.com)

Èric Garcia Dalmases (ericgada05@gmail.com)

Claudia Tello Calles (tellocallesclaudia@gmail.com)

Jesús García Quesada (jesusgarcia.301004@gmail.com)

Docente:

José Miguel Ruiz Guevara

ÍNDICE

1. TÍTULO PROVISIONAL DEL PROYECTO

CryptoPredict. Enlace a GitHub:
https://github.com/TFM-Master-en-IA-y-Big-Data/proyecto-tfm

2. ELEVATOR PITCH

El mercado de criptomonedas es altamente volátil y carece de herramientas integrales que combinen análisis masivo de datos y estimación cuantitativa del riesgo. Este proyecto desarrolla un sistema end-to-end que procesa datos con Spark, entrena modelos de Machine Learning en Python y expone predicciones mediante una API. Se complementa con dashboards en Power BI y una interfaz web interactiva. Aporta una solución completa y reproducible para analizar el histórico y estimar probabilísticamente subidas, bajadas y caídas severas.

3. EL TEMA Y EL PROBLEMA

Actualmente, el mercado de criptomonedas, como Bitcoin y Ethereum, se caracteriza por una elevada volatilidad, cambios bruscos de tendencia y un comportamiento altamente no lineal. Aunque existen múltiples fuentes de datos históricos y plataformas de visualización, la información suele encontrarse fragmentada y orientada principalmente a mostrar gráficos o indicadores técnicos básicos. No existe una integración clara entre procesamiento masivo de datos, modelado predictivo y estimación probabilística de riesgo dentro de un sistema estructurado y reproducible.

Esta situación genera una dificultad real en la toma de decisiones. Los usuarios deben consultar diferentes herramientas, interpretar manualmente indicadores y, en muchos casos, actuar sin una estimación cuantitativa clara del riesgo de caídas severas. Dado que los movimientos en criptomonedas pueden ser abruptos y significativos en periodos muy cortos, la falta de herramientas integrales puede traducirse en pérdidas importantes o en análisis poco rigurosos desde el punto de vista académico y profesional.

Por ello, resulta relevante desarrollar un sistema que combine procesamiento Big Data, Machine Learning y visualización interactiva en una arquitectura unificada. Un enfoque integral permite transformar datos complejos en información cuantitativa accionable, estimar probabilísticamente escenarios de riesgo y ofrecer una visión estructurada del comportamiento del mercado. Además de su valor académico, este tipo de solución sienta las bases para aplicaciones fintech más avanzadas y para una mejor comprensión y gestión del riesgo en entornos financieros altamente volátiles.

4. OBJETIVO PRINCIPAL DEL PROYECTO

Construir un sistema integral de análisis y predicción de criptomonedas que permita procesar grandes volúmenes de datos históricos, entrenar modelos de Machine Learning y exponer

predicciones en tiempo real para lograr una estimación probabilística y visual del riesgo de mercado.

5. ALCANCE INICIAL (IN / OUT)

IN:

- Elección de criptomonedas principales: Se seleccionan cinco criptomonedas representativas del mercado en términos de capitalización, volatilidad y modelo de negocio, con el objetivo de permitir tanto un análisis comparativo sólido como el entrenamiento de modelos de predicción más generalizables.
- Elaboración de series temporales históricas: se procesarán ventanas deslizantes de datos (por ejemplo, últimos 30 días) para calcular métricas relevantes y preparar el dataset para entrenamiento de modelos de Machine Learning.
- Desarrollo de features y métricas técnicas: se generarán diferentes indicadores durante el procesamiento de los datos, permitiendo capturar el comportamiento del mercado y sus patrones de riesgo.
- Modelo de clasificación binaria (MVP): predicción de subida o bajada a corto plazo de cada criptomoneda, utilizando los datos procesados y las features generadas.
- Dashboard histórico y comparativo: se visualizará la evolución de cada criptomoneda, su volatilidad y métricas técnicas, facilitando la comparación y análisis exploratorio de datos en Power BI.
- API de inferencia y frontend multiplataforma: permitirá recibir solicitudes sobre cualquier criptomoneda incluida, calcular automáticamente las features correspondientes y devolver predicciones en tiempo real, además de mostrar resultados de manera interactiva.
- Modelo adicional para caídas severas (>10%): desarrollo opcional que estimará la probabilidad de una caída importante, a implementar solo si el tiempo disponible lo permite.

OUT:

- Predicción del precio estimado (o rango de precio) de las criptomonedas.
- Análisis de criptomonedas minoritarias con bajo volumen o liquidez.
- Sistemas de trading automático.
- Modelos de riesgo ultra-avanzados que no estén contemplados en los IN (solo posibles extensiones futuras).

6. USUARIOS / PARTES INTERESADAS

El sistema está orientado principalmente a inversores minoristas en criptomonedas que desean disponer de una estimación cuantitativa del riesgo antes de tomar decisiones. También está dirigido a analistas financieros junior, estudiantes y perfiles interesados en mercados volátiles, así como a entusiastas de Data Science y Machine Learning que quieran estudiar un caso práctico aplicado a series temporales financieras reales.

A nivel académico, los stakeholders clave son la dirección del TFM, el tribunal evaluador y el propio Máster en IA y Big Data, ya que el proyecto pretende servir como caso demostrativo end-to-end, reproducible y técnicamente coherente. En el ámbito técnico, las partes interesadas incluyen el equipo de desarrollo, la comunidad open-source (mediante repositorio en GitHub) y posibles iniciativas fintech futuras que puedan reutilizar la arquitectura propuesta.

7. PROPUESTA DE SOLUCIÓN (ALTO NIVEL)

Se propone una arquitectura modular y escalable dividida en cuatro capas. En la capa de datos, se obtendrán datos históricos de mercado mediante APIs públicas, que serán procesados con Apache Spark para su limpieza, normalización y generación de features técnicas. Los datos se almacenarán en formato estructurado optimizado (Parque) para facilitar su reutilización.

En la capa de modelado, se construirá un dataset supervisado mediante ventanas temporales deslizantes y se entrenarán modelos de clasificación binaria (Regresión Logística, Random Forest y Gradient Boosting/XGBoost, a probar según cuál resulte más efectivo). La evaluación se realizará mediante métricas como Accuracy y/o F1-score, aplicando validación temporal para garantizar rigor metodológico.

La capa de servicio consistirá en una API REST desarrollada con FastAPI, que permitirá recibir el símbolo de una criptomoneda, generar automáticamente las features actuales y devolver la probabilidad de subida o bajada. Finalmente, la capa de visualización incluirá un dashboard en Power BI para análisis histórico y comparativo, junto con una interfaz web que permitirá consultar predicciones y visualizar el riesgo de forma interactiva.

La arquitectura completa sigue un flujo claro: fuente de datos → procesamiento con Spark → feature engineering → modelo ML, con salida hacia dashboard y API, manteniendo modularidad y reproducibilidad.

8. DATOS (ESTIMACIÓN INICIAL)

Se utilizarán APIs públicas como CoinGecko y Binance (aún pendiente de realizar un estudio exhaustivo sobre diferentes fuentes de datos) para la obtención de datos históricos de mercado. Los datos incluirán precios de apertura, cierre, máximo y mínimo, volumen, capitalización y timestamp con frecuencia diaria en una primera fase.

El proyecto se centrará en un conjunto reducido de criptomonedas representativas del mercado, seleccionadas en función de su capitalización, liquidez y relevancia histórica. Se trabajará con datos históricos de varios años y frecuencia diaria, lo que permitirá disponer de un volumen suficiente de registros para el desarrollo del MVP. La arquitectura estará diseñada para ser escalable, de modo que, si el rendimiento lo permite, pueda ampliarse posteriormente a frecuencias temporales más finas.

9. VIABILIDAD TÉCNICA

El proyecto se desarrollará con Python, Apache Spark, FastAPI y Power BI, utilizando GitHub para control de versiones y Docker de forma opcional para despliegue reproducible.

Desde el punto de vista técnico, el procesamiento de series temporales presenta complejidad media-alta, mientras que el modelado supervisado binario y el desarrollo de una API REST son abordables dentro del marco del TFM. El dashboard presenta complejidad baja-media. No se requiere infraestructura cloud obligatoria, las APIs son gratuitas y un entorno local con 16GB de RAM es suficiente. En consecuencia, el proyecto se considera técnicamente viable dentro del calendario previsto.

10. RIESGOS PRINCIPALES Y PLAN DE MITIGACIÓN

El principal riesgo es el bajo rendimiento predictivo debido a la eficiencia del mercado. En ese caso, el enfoque se reorientará hacia estimación probabilística de riesgo y análisis estadístico del comportamiento del mercado, incorporando modelos baseline comparativos.

Existe también riesgo de overfitting por la alta volatilidad; se mitigará mediante regularización, validación temporal estricta y posible reducción de features. Problemas con APIs o limitaciones de acceso se resolverán mediante descarga masiva inicial y almacenamiento local de los datos. Finalmente, si no hubiera tiempo suficiente, el modelo de predicción de caídas severas se mantendrá como extensión futura documentada, priorizando el modelo binario principal.

11. REPARTO DE TAREAS

El equipo se organizará en cuatro áreas alineadas con la arquitectura del sistema:

- Claudia→ Data: ingesta de datos desde APIs, procesamiento con Spark, limpieza, almacenamiento estructurado y análisis exploratorio (EDA).
- Cristian → Machine Learning: construcción del dataset supervisado, entrenamiento y evaluación de modelos de clasificación, validación temporal y selección del modelo final.
- Jesús → Platform: desarrollo de la API con FastAPI, integración del modelo para inferencia y preparación del entorno de despliegue.

- Eric → Business Intelligence: diseño del dashboard en Power BI, definición de KPIs y visualización de métricas clave y comparativas.

12. PRÓXIMOS PASOS INMEDIATOS

1. Confirmar y documentar las criptomonedas definitivas que formarán parte del proyecto.
2. Definir formalmente la variable objetivo (horizonte temporal y criterio de subida/bajada).
3. Implementar el script de descarga automática de datos históricos desde las APIs seleccionadas.
4. Ejecutar una primera carga completa de datos y almacenarlos en formato estructurado (Parquet).
5. Realizar un análisis exploratorio inicial (EDA) para validar calidad, detectar valores nulos y analizar distribuciones.
6. Implementar la lógica de ventanas deslizantes para la construcción del dataset supervisado.
7. Generar las primeras features técnicas (retornos, volatilidad, momentum).
8. Entrenar un modelo baseline inicial (Regresión Logística) para validar que el pipeline completo funciona end-to-end.

ANEXO - EVIDENCIA DE EQUIPO: “ACUERDO DE TRABAJO”

La comunicación principal del equipo se realizará a través de **WhatsApp** para coordinación ágil y mensajes rápidos, y mediante **Discord** para reuniones formales y sesiones técnicas compartiendo pantalla. **GitHub** será el canal central para la gestión técnica del proyecto (issues, pull requests y control de versiones).

En cuanto a disponibilidad, los miembros del equipo mantendrán comunicación activa durante la semana en horario de tarde, con flexibilidad según compromisos individuales. Se establecerá una frecuencia aproximada de dos reuniones semanales de 30–45 minutos, orientadas a seguimiento de tareas, resolución de bloqueos y revisión de avances. En fases críticas (entrenamiento de modelos, integración API o preparación de defensa) se podrán añadir reuniones extraordinarias si es necesario.

Para garantizar orden y calidad en el desarrollo, se establecen las siguientes reglas básicas de trabajo:

- No se subirá código directamente a la rama main sin revisión previa mediante Pull Request.
- Todo trabajo deberá estar asociado a una issue previamente definida, asegurando trazabilidad.
- Las decisiones técnicas relevantes (elección de modelo, cambios de arquitectura, definición de target, etc.) deberán documentarse brevemente en el repositorio.
- Se respetará el reparto de responsabilidades acordado, pero cualquier miembro del equipo deberá apoyar a otro en caso de bloqueo técnico o sobrecarga puntual.

Este marco organizativo busca mantener profesionalidad, trazabilidad y eficiencia, asegurando una dinámica colaborativa alineada con prácticas reales de desarrollo en entornos tecnológicos.