CIENCIAS DE DATOS Y ANALÍTICA

TRADING ALGORÍTMICO

Proyecto elaborado por:

Allison Quintero

Pedro Villegas

Jesús Samuel Benjumea

Hassan Amid Chedraui

Dilger Felipe Becerra

Agenda

1. Problema a resolver
2. Análisis de datos
3. Desarrollo y evaluación de los modelos
4. Análisis de resultados
5. Tecnologías de Cloud Services
6. Conclusiones

¿Qué problema estamos resolviendo?

Decisiones de trading a corto plazo influenciadas por sesgos cognitivos.

Objetivo

Usar datos objetivos para mejorar el timing en decisiones de entrada y salida.

Pregunta problema

¿Es posible diseñar un modelo algorítmico en el que utilizando datos financieros objetivos en tiempo casi real, reduzca la influencia de los sesgos cognitivos y mejore el timing en las decisiones de trading?

Hipótesis

El uso de indicadores técnicos integrados en modelos automatizados permite tomar decisiones de trading más eficientes que las basadas en intuición, al disminuir la influencia de los sesgos cognitivos.

Marco Teórico

🧠 **Finanzas Conductuales:**

- **Sesgos comunes**: Confirmación, Anclaje, Sobreconfianza, Aversión a la pérdida.
- **Impacto:** Decisiones impulsivas e inconsistentes.

📉 **Hipótesis de Mercados Eficientes (EMH):**

- Los precios no siempre reflejan toda la información.
- Las ineficiencias permiten oportunidades para modelos sistemáticos.

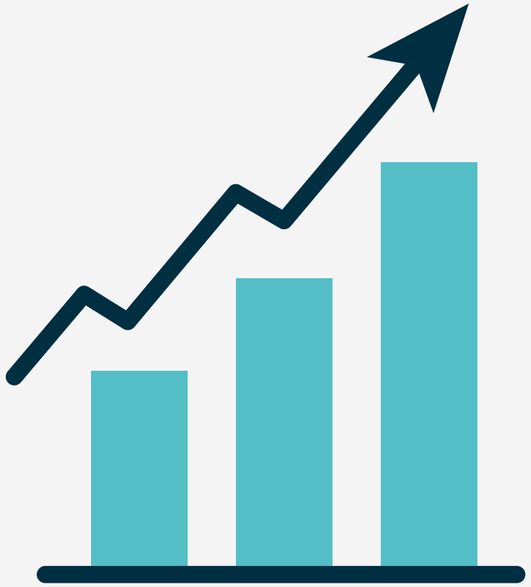
📊 **Análisis Técnico:**

- **Indicadores:** SMA, EMA, RSI, MACD, Bandas de Bollinger.
- **Problema:** Interpretación subjetiva.
- **Solución:** Automatización de señales.

Marco Teórico

📈 Estadística Financiera:

- **Herramientas:** Regresiones, pruebas de hipótesis, modelos GARCH.
- **Rol:** Validar reglas y detectar patrones reales.



🧮 Álgebra Lineal:

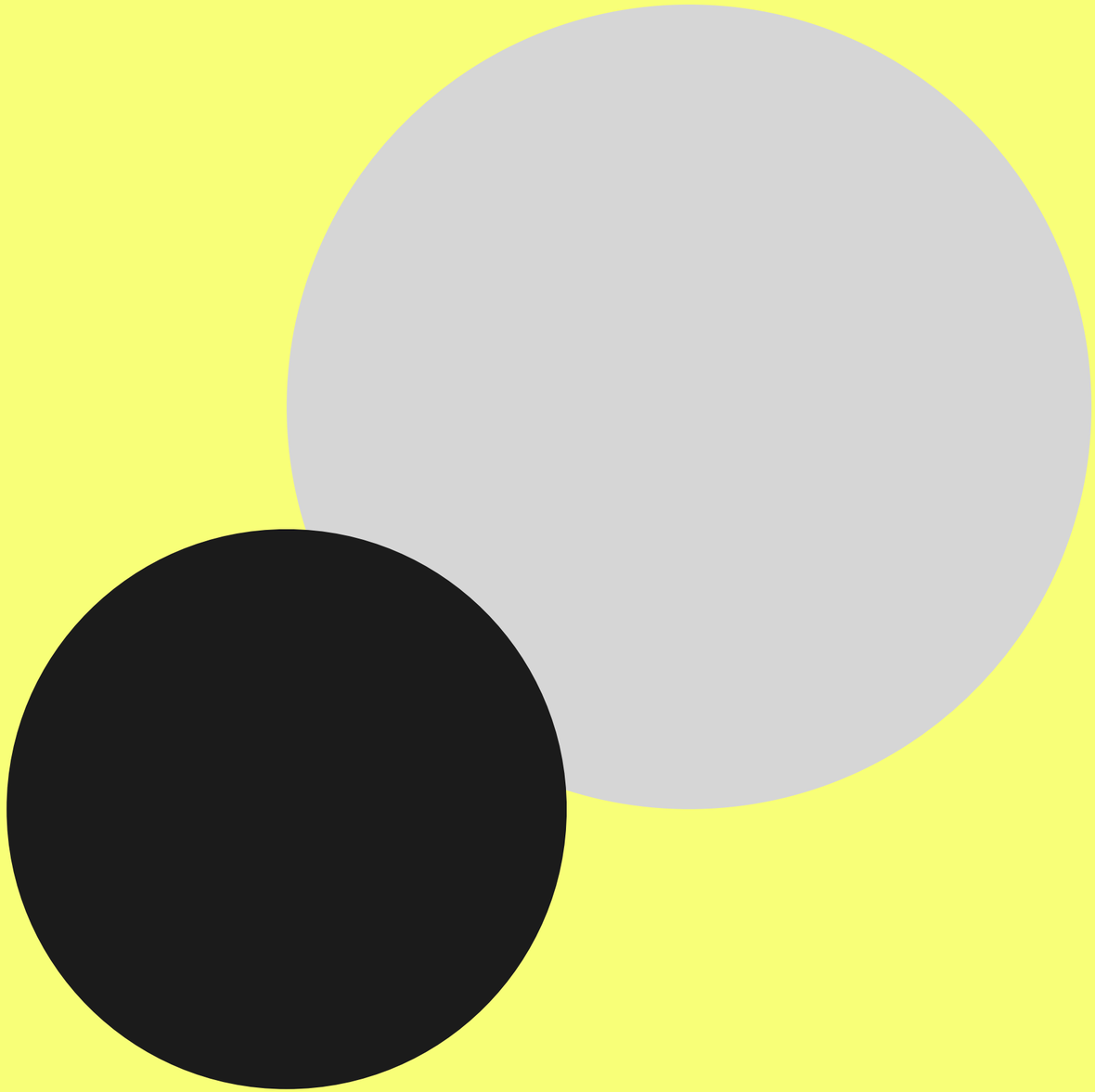
- PCA, SVD, matrices de correlación.
- **Aplicación:** Reducción de dimensionalidad y comparación de activos.

💻 Ciencia de Datos & CRISP-DM:

- **Enfoque estructurado:** desde la comprensión del problema hasta el despliegue.
- **Herramientas:** K-means, modelos predictivos, dashboards con backtesting.

¿Con qué datos trabajamos?

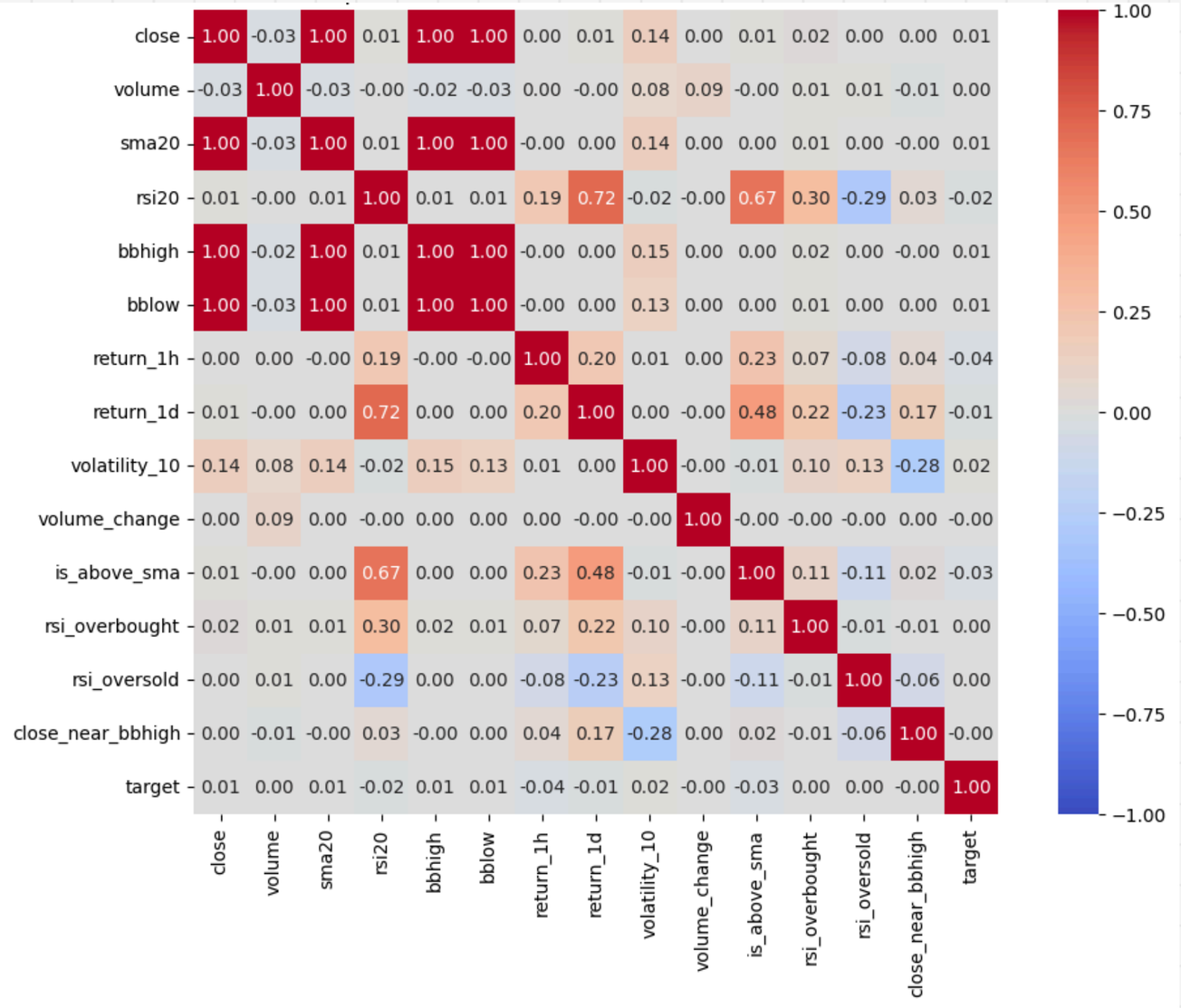
- Datos minuto a minuto desde agosto de 2022 Apple, Microsoft, Google, Amazon, Meta, Nvidia y Tesla.
- Limpieza y filtrado por calidad.
- Construcción de variables técnicas (SMA, RSI, Bandas de Bollinger).
- Creación de la variable target como señal de compra.



¿Qué nos dicen los datos?

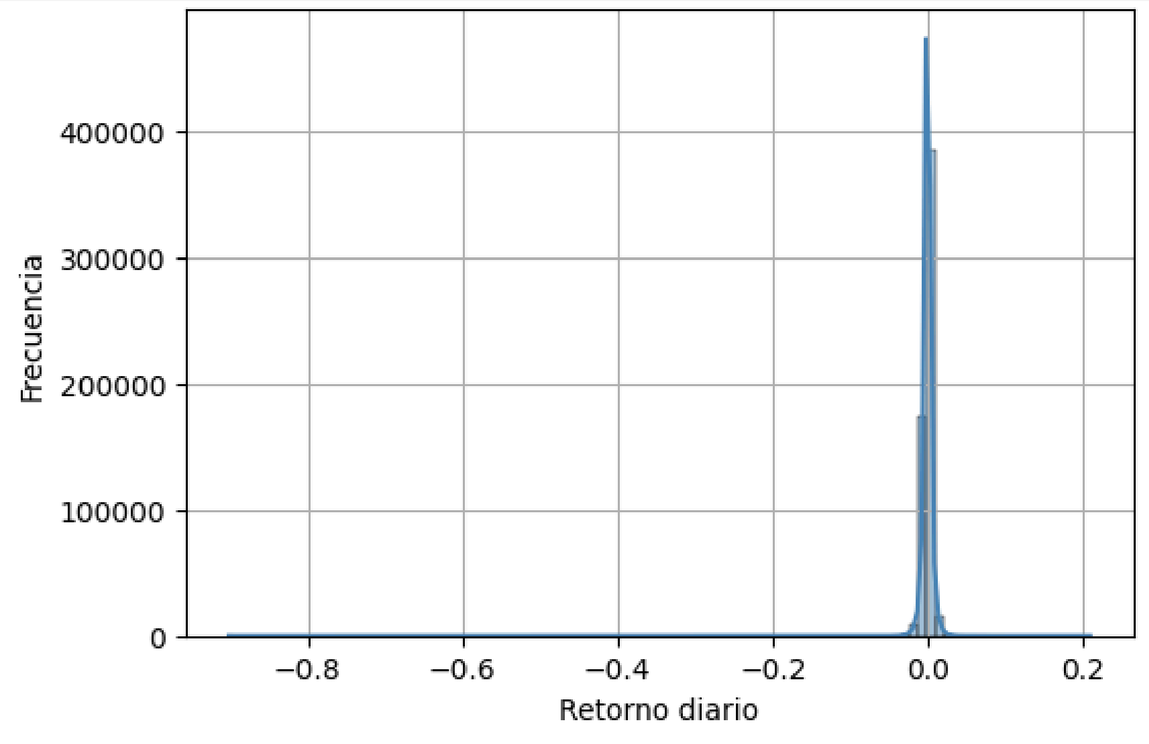
- Alta correlación entre close, sma20, bbhigh, bblow.
- RSI correlacionado con return_1d en algunos activos.
- Variables binarias (target, rsi_overbought, etc.) balanceadas y útiles para clasificación.

Correlación entre variables técnicas – GOOGL



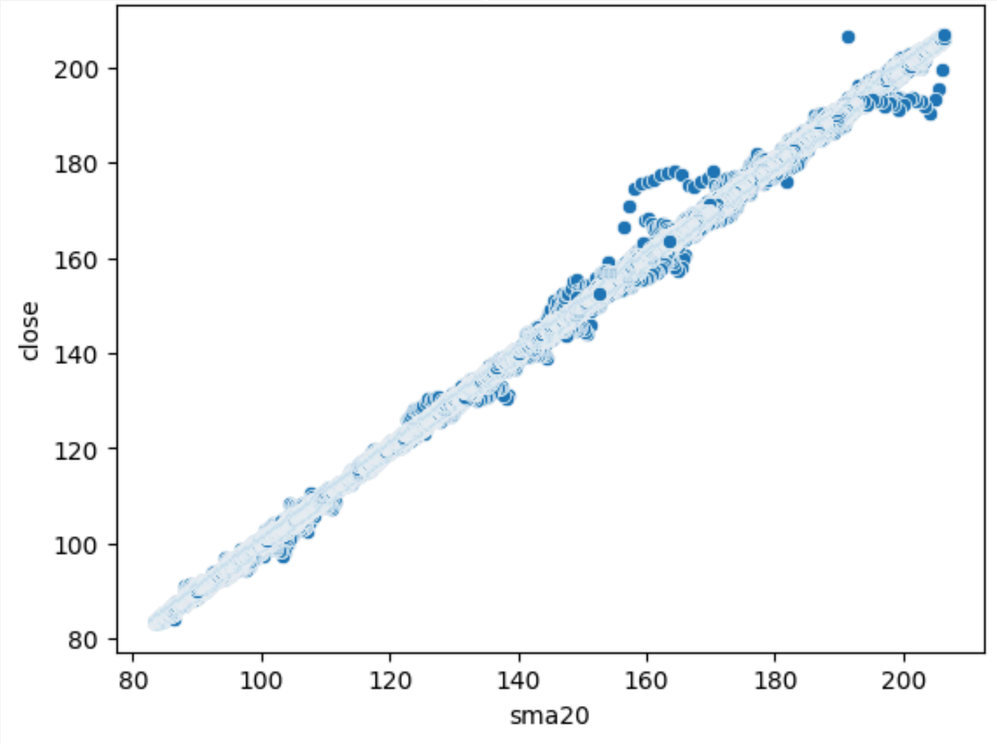
Destacan correlaciones fuertes entre indicadores derivados del precio y correlaciones relevantes entre RSI y retornos diarios.

Distribución de retornos diarios – NVDA



Distribución centrada en cero con colas largas, típica de retornos financieros

Relación entre precio y media móvil – GOOGL



Confirma que la SMA es útil como insumo para detectar tendencias.

Modelos que trabajamos

Y aquellos que son mas optimos para nuestro problema

Regresion Logistica

Razones de uso

- Modelo base y rápido.
- Interpretable y fácil de calibrar.

Decision Tree

Razones de uso

- Permite aprendizaje de reglas claras (if-then).
- Capaz de modelar relaciones no lineales.

Random Forest

Razones de uso

- Ensamble de múltiples árboles (bagging).
- Mejora generalización y reduce overfitting.

¿Como los entrenamos?

Q1

Proceso de entrenamiento
Variables utilizadas:

- close, volume, sma20, rsi20, bbhigh, bblow
- return_1h, return_1d, volatility_10, volume_change
- Condiciones booleanas: is_above_sma, rsi_oversold, etc.

Q2

Target binario:

- target = ((close.shift(-1) / close - 1) > 0.003).astype(int)

- 1 si sube al menos +0.3%, 0 si no.

Q3

Configuración del entrenamiento:

- train_test_split 80% / 20% (sin shuffle, por orden temporal)
- Scaler: StandardScaler()
- Hiperparámetros:
  - LogisticRegression: max_iter=1000
  - DecisionTree: max_depth=5
  - RandomForest: n_estimators=100, max_depth=7

Resultados Entrenamiento

Modelo	Accuracy	Precision	Recall	F1	AUC	MCC
Regresión Log.	0.5268	0.4977	0.0386	0.0716	0.5164	0.0096
Árbol Decisión	0.5282	0.5061	0.1072	0.1769	0.5330	0.0221
Random Forest	0.5284	0.5102	0.0759	0.1322	0.5345	0.0205

¿Cómo los evaluamos?

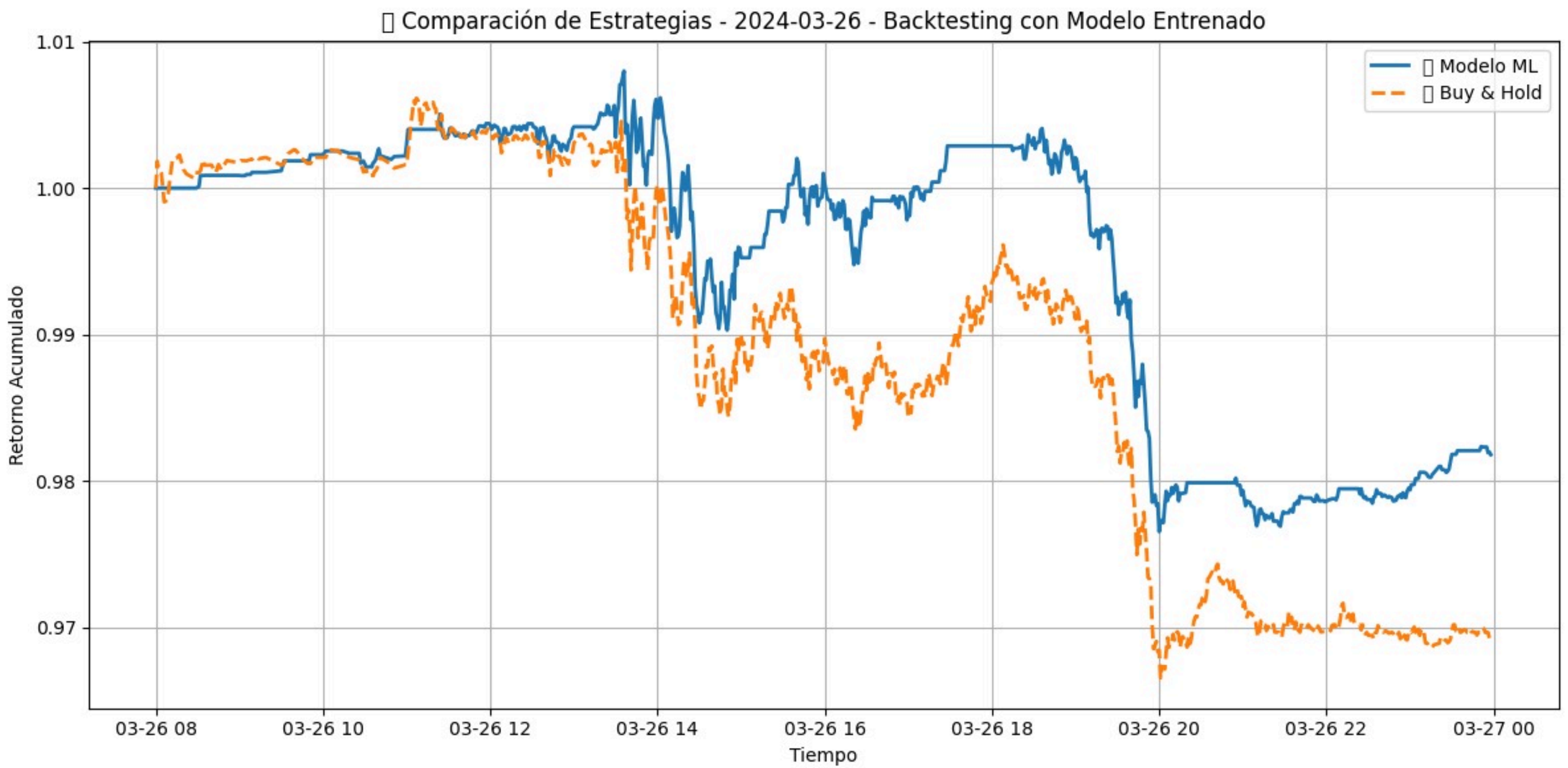
Métrica	¿Qué mide?
Accuracy	Porcentaje general de aciertos
Precision	Señales de compra que realmente subieron
Recall	Subidas que fueron correctamente detectadas
F1 Score	Balance entre precision y recall
AUC ROC	Capacidad de separar subidas de no subidas
MCC	Correlación binaria balanceada (ideal para clases desbalanceadas)
Hamming Loss	Porcentaje de errores en clasificación
Lift Curve	Cuánto mejor que el azar identifica los positivos el modelo en los primeros percentiles

Backtesting

- Simulación de portafolio con capital inicial ($10,000)
- Compra/venta realista con señales del modelo
- Comisiones del 2% sobre ganancia
- Comparación contra estrategia Buy & Hold
- Curvas de rendimiento: portafolio vs mercado

Random Forest, Mejor resultado

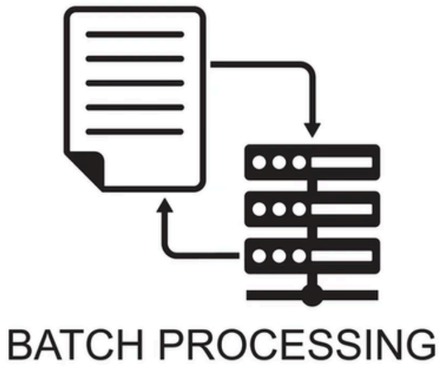
Resultados Backtesting



Tecnologías de nube

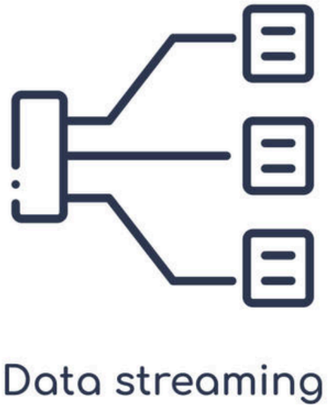
Dentro de un ambiente de Amazon Web Services

Fase 1



Desarrollo, entrenamiento y procesamiento de datos históricos estructurados

Fase 2



Utilización, despliegue y producción de decisiones por medio de modelos ML

Arquitectura Batch

Alpaca

Ingesta - Procesamiento - Modelado

Ingesta

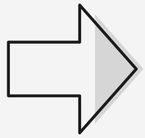
Raw Zone

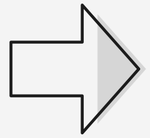
Procesamiento

Trusted Zone

Modelado

Refined Zone





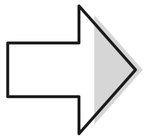
Arquitectura Streaming

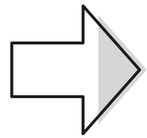
Ingesta - Procesamiento - Serving

Ingesta

Procesamiento

Serving

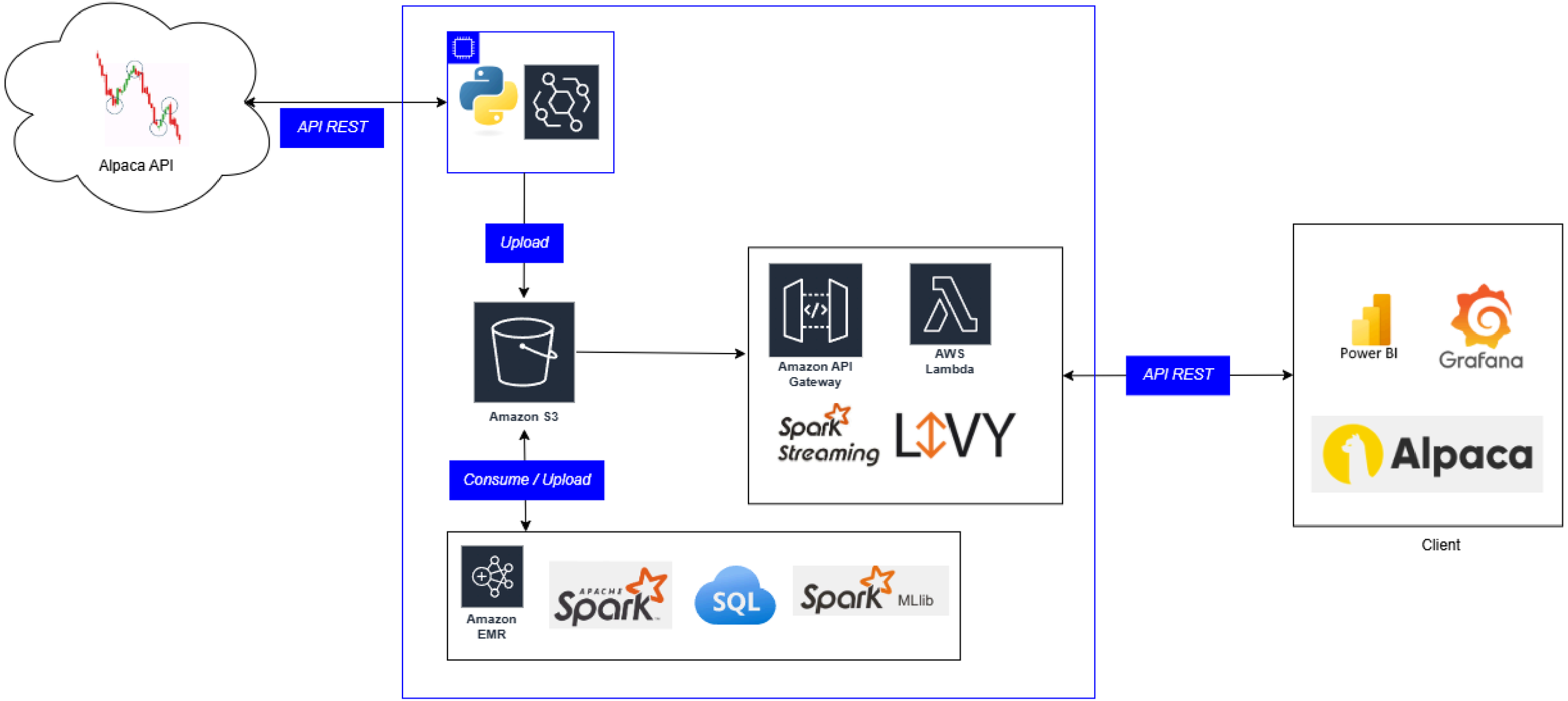




Automatización

Automatización

Diagrama de Arquitectura



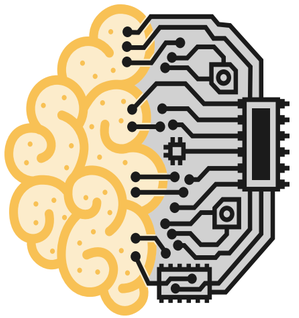
Conclusiones: Transformando Datos en Decisiones de Trading

De la teoría a la práctica:

Logramos materializar un sistema algorítmico para analizar las "7 Magníficas" y generar señales de compra.

Combatiendo Sesgos Humanos:

El proyecto ofrece una alternativa objetiva y basada en datos para las decisiones de trading, minimizando errores comunes.

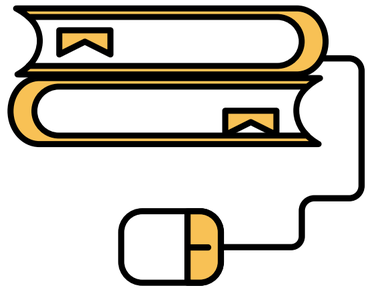


Potencial Real y Escalable:

Contamos con una base funcional y una arquitectura en la nube (AWS) preparada para operar y evolucionar.

Valor Educativo y Aplicado:

Más allá de la rentabilidad inicial, el proyecto integró exitosamente múltiples disciplinas de la ciencia de datos.

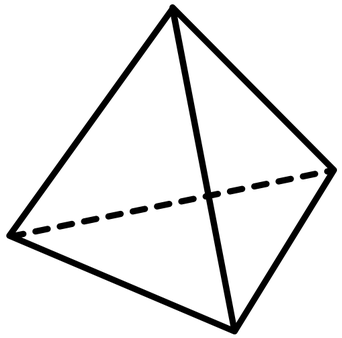


Fundamentos de Ciencias de datos

- Guió el ciclo de vida completo: desde definir el problema de los sesgos en el trading hasta la evaluación de modelos de Machine Learning.
- Fundamental para el entendimiento, preparación y limpieza de los datos financieros.
- Aplicación de la metodología de zonas (RAW, TRUSTED, REFINED) para asegurar la calidad y gobernanza de los datos.

Algebra en ciencias de datos

- Esencial en el manejo de datos como tablas (matrices) y variables (vectores) como precio, volumen, SMA20, RSI20.
- Base de los modelos de Machine Learning (ej. Regresión Logística) para calcular pesos de variables.
- Aplicada en la transformación y escalado de características para el entrenamiento de modelos.

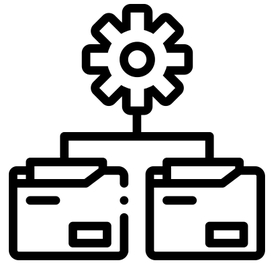


Estadística en analítica

- Crucial para el análisis exploratorio: entender distribuciones, correlaciones e identificar inconsistencias.
- Base de indicadores técnicos como Medias Móviles Simples, RSI y Bandas de Bollinger.
- Utilizada para la evaluación del rendimiento de modelos con métricas como Accuracy, Precision, Recall, F1-Score y AUC-ROC.
- Consideraciones sobre validación temporal y data drift en series de tiempo financieras.

Almacenamiento y recuperación

- Clave para diseñar el manejo de grandes volúmenes de datos históricos y en tiempo real.
- Implementación de Data Lake en Amazon S3 con zonas RAW, TRUSTED y REFINED.
- Uso de Amazon EMR con Spark para procesamiento distribuido y planificación de AWS Lambda y Timestream para datos en tiempo real.
- Catalogación con Hive SQL o Glue Data Catalog para facilitar la consulta.



Puntos de Mejora

- Comportamiento humano impredecible
- Sector volátil como el mercado accionario americano (información asimétrica)
- Más variables de otros enfoques (análisis de sentimientos)
- Evaluar otros subsectores de aplicación (sectores o empresas especificas)
- Modelos más avanzados (RNN o LSTM)

Referencias

Imágenes:

- https://www.shutterstock.com/es/search/batch-processing-icon
- https://www.istockphoto.com/es/vector/icono-de-streaming-de-datos-lineales-de-seguridad-de-internet-y-colecci%C3%B3n-de-gm1134332835-301391318
- https://es.wikipedia.org/wiki/AWS_Lambda
- https://worldvectorlogo.com/es/logo/amazon-s3-simple-storage-service
- https://stackademic.com/blog/spring-boot-eventbridge
- https://www.thewealthmosaic.com/vendors/alpaca/
- https://es.wikipedia.org/wiki/Apache_Spark
- https://medium.com/rv-data/my-first-foray-into-spark-mllib-2907dde75f73
- https://www.datacamp.com/es/blog/all-about-power-bi
- https://en.wikipedia.org/wiki/Grafana
- https://cloud.in28minutes.com/aws-certification-devops-in-aws-cloudformation-vs-codepipeline-vs-opsworks
- https://towardsaws.com/amazon-api-gateway-95b7f711b14a?gi=1bc04aa1e35e
- https://digitalfactoryalliance.eu/components/apache-livy/