
UNIVERSIDAD SAN FRANCISCO

**FUNDAMENTOS DE CIENCIA DE DATOS**

# PROYECTO
# MODELO PREDICTIVO PARA ACCIONES DE LAS EMPRESAS QUE ESTAN DENTRO DEL S&P 500

**GABRIELA ZUMARRAGA – 345769**

**QUITO 10 DE ABRIL 2025**

# PROYECTO

# MODELO PREDICTIVO PARA ACCIONES DE LAS EMPRESAS QUE ESTAN DENTRO DEL S&P 500

**Tema:** Análisis de los movimientos de las acciones de las empresas que se encuentran en el S&P (Standard and Poor's) 500 para realizar un modelo predictivo utilizando datos históricos del mercado bursátil

## 1. BUSINESS UNDERSTANDING

### a. Contexto

El índice S&P 500, (Standard & Poor's 500), es un indicador bursátil sobre el rendimiento de las acciones de las 500 empresas lideres que cotizan en la bolsa de valores de Estados Unidos (por capitalización de mercado). Este es un indicador clave de la salud económica de los Estados Unidos ya que abarca información sobre diferentes tipos de industrias, como tecnología, atención médica, servicios financieros, bienes de consumo y energía, seleccionadas por su tamaño, liquidez y representatividad económica.

Hipótesis principal: Los movimientos históricos de precios pueden revelar tendencias repetitivas o correlaciones significativas que permitan prever cambios en los precios de las acciones en el mercado bursátil.

Qué necesitas comprobar o resolver: Identificar patrones relevantes en los precios de las acciones que puedan ayudar a entender el comportamiento de los precios de las acciones en el mercado para tomar decisiones de inversión y mejorar el rendimiento.

Pregunta de investigación: **¿**Qué patrones de precios se existen en las acciones de las empresas del S&P 500, y cómo pueden usarse para anticipar movimientos futuros?

## b. Objetivo

El objetivo principal de este proyecto es desarrollar un modelo predictivo que permita anticipar si el precio de las acciones de las empresas que se encuentran dentro del S&P 500 subirá o bajará (diariamente) en función de datos históricos, como precios de apertura, cierre, cambios diarios de precios y otras variables derivadas.

## c. Visión

Este análisis puede ayudar a los stakeholders, inversionistas, analistas financieros o incluso instituciones financieras a:

- Tomar decisiones informadas: sobre la compra o venta de acciones para mejorar sus rendimientos al conocer el comportamiento o la tendencia de las acciones de interés.
- Gestión de riesgos: permitirá identificar patrones que podrían reducir el riesgo asociado con las inversiones este mercado.
- Optimizar estrategias: facilita la creación de estrategias de trading automatizadas basadas en predicciones de movimientos del mercado.

## d. Alcance

El alcance de este proyecto incluye los siguientes puntos:

- EDA: realizar un análisis inicial para entender los datos recolectados, analizando las correlaciones entre variables, identificando anomalías y patrones relevantes.
- Data Wrangling: procesar, limpiar y normalizar los datos asegurando su preparación adecuada para ser utilizados en los modelos de predicción.
- Modelado: evaluar diferentes modelos predictivos de clasificación para determinar el más adecuado en la predicción del movimiento diario del precio de una acción (subida o bajada). Esta tarea se llevará a cabo mediante el análisis y la comparación de las métricas

de desempeño de cada modelo, seleccionando finalmente el que ofrezca los mejores resultados para abordar el problema planteado.

### e. Limitaciones

- Fiabilidad de los datos: el dataset fue sacado de Kaggle y según la fuente fue actualizado hace 3 meses. Si bien los datos no están al día, se puede encontrar posibles inconsistencias si la data no fue bien actualizada como se menciona. Esto podría influir en los modelos predictivos.
- Datos faltantes: se debe analizar qué tipo de estrategia se debe abordar en caso de que haya muchos datos faltantes ya que pueden comprometer el análisis. Si el porcentaje es pequeño se puede considerar la imputación de valores según corresponda.

## 2. DATA UNDERSATANDING

### a. Fuente de datos

Los datos que se utilizaron provienen de Kaggle y contiene precios históricos (2010-2024) de acciones para todas las empresas que actualmente se encuentran en el S&P 500.

- Enlace: S&P 500 Stocks (daily updated)

El dataset contienen 3 archivos que se encuentran dentro de la carpeta data/clean. Estos archivos contienen la siguiente estructura.

- sp500_stocks: Este es el archivo principal con el que se realizara el análisis para el modelo predictivo. Los campos que contiene este archivo permitirán ver los patrones históricos de las empresas y explorar relaciones significativas que permitan predecir el comportamiento de los precios en el mercado.

| Campo | Tipo de dato | Descripción |
|---|---|---|
| Date | float | Fecha del registro en formato yy-mm-dd |
| Open | float | Precio de la acción al inicio del mercado (esta es información del NYSE, por lo tanto, todo en USD) |
| High | float | Precio más alto alcanzado del periodo |
| Low | float | Precio más bajo alcanzado del periodo |
| Close | float | Precio de cierre del mercado. |
| Volume | int | Cantidad de acciones negociadas. |
| Adj Close | float | Similar al precio de cierre del mercado, pero incluye acciones de la empresa como dividendos y divisiones de acciones. |
| Symbol | float | Símbolo de la empresa |

- sp500_companies: Este archivo contiene datos descriptivos de las empresas.

| Campo | Tipo de dato | Descripción |
|---|---|---|
| Exchange | Object(string) | Intercambio donde se negocian sus acciones. |
| Symbol | Object(string) | Símbolo de la acción. |
| Shortname | Object(string) | Nombre corto de la empresa. |
| Longname | Object(string) | Nombre completo de la empresa. |
| Sector | Object(string) | Sector en el que opera la empresa. |
| Industry | Object(string) | Industria, dentro de un sector, en la que opera la empresa. |
| Currentprice | Float | Precio actual de la acción. |
| Marketcap | int | Capitalización de mercado actual. |
| Ebitda | Float | Ganancias antes de intereses, impuestos, depreciación y amortización. |
| Revenuegrowth | Float | Crecimiento de ingresos. |
| City | Object(string) | Ciudad matriz de la empresa. |
| State | Object(string) | Estado matriz de la empresa. |
| Country | Object(string) | País matriz de la empresa. |
| Fulltimeemployees | int | Número de empleados a tiempo completo. |
| Longbusinesssummary | Object(string) | Descripción general de la compañía. |
| Weight | float | Porcentaje de participación en el índice S&P 500 (según su capitalización de mercado). |

- sp500_index: este archivo contiene datos directamente relacionados con el índice S&P 500 con respecto al tiempo.

| Campo | Tipo de dato | Descripción |
|---|---|---|
| Date | Datetime | Fecha del registro. |
| S&P 500 | float | Valor del índice S&P 500. |

## b. Descripción y calidad de los datos

Como se mencionó, los datos relevantes para el análisis se encuentran en el archivo 'sp500_stocks', por lo que este será descrito en detalle en esta sección. Es importante mencionar también se realizó el proceso de análisis y limpieza para los otros dos archivos. El detalle de estos análisis está documentado en los notebooks EDA y Data Wrangling (en una sección para cada archivo), como parte del trabajo realizado previamente.

Para el archivo 'sp500_stocks' tenemos el siguiente detalle obtenido en el análisis inicial (EDA):

```
RangeIndex: 1891536 entries, 0 to 1891535
Data columns (total 8 columns):
 #   Column     Non-Null Count    Dtype
---  ------     --------------    -----
 0   Date       1891536 non-null  object
 1   Symbol     1891536 non-null  object
 2   Adj Close  617831 non-null   float64
 3   Close      617831 non-null   float64
 4   High       617831 non-null   float64
 5   Low        617831 non-null   float64
 6   Open       617831 non-null   float64
 7   Volume     617831 non-null   float64
dtypes: float64(6), object(2)
```

| | Date | Symbol | Adj Close | Close | High | Low | Open | Volume |
|---|---|---|---|---|---|---|---|---|
| 0 | 2010-01-04 | MMM | NaN | NaN | NaN | NaN | NaN | NaN |
| 1 | 2010-01-05 | MMM | NaN | NaN | NaN | NaN | NaN | NaN |
| 2 | 2010-01-06 | MMM | NaN | NaN | NaN | NaN | NaN | NaN |
| 3 | 2010-01-07 | MMM | NaN | NaN | NaN | NaN | NaN | NaN |
| 4 | 2010-01-08 | MMM | NaN | NaN | NaN | NaN | NaN | NaN |

- Cantidad de instancias: 1891536

- Existen datos desde el 2010-01-04 hasta 2024-12-20

- Tipos de datos:
  - Campo fecha es de tipo ‘float’, se lo pude conviertir a ‘datetime’ y posteriormente se lo puede manejar como índice para entrenar el modelo.
  - Campo volumen se puede tratar como tipo entero ya que es la cantidad total de acciones.
  - El resto de ‘features’ son numéricos y corresponden a los precios de las acciones, por lo que se los puede tratar como tipo ‘float’.

- Valores nulos: El archivo contiene 1891536 instancias. Podemos observar que en los siguientes campos existen valores nulos.

| Variable | Valores no nulos |
|---|---|
| Adj Close | 617831 |
| Close | 617831 |
| High | 617831 |
| Low | 617831 |
| Open | 617831 |
| Volume | 617831 |

  El porcentaje de valores nulos existentes en estas columnas es considerable (67,36%), por lo que pueden impactar el análisis. Se podrían tratar, por ejemplo, mediante transformaciones al imputar valores. Por ejemplo: para la variable *‘adj close’* se puede reemplazar los valores nulos con el precio de cierre ya que están estrechamente correlacionados, para *‘close’, ‘open’, ‘high’, ‘low’,* se pueden reemplazar con datos o la media de días cercanos.

- Estadísticas descriptivas: se obtuvo un detalle de las estadísticas de cada columna numérica.

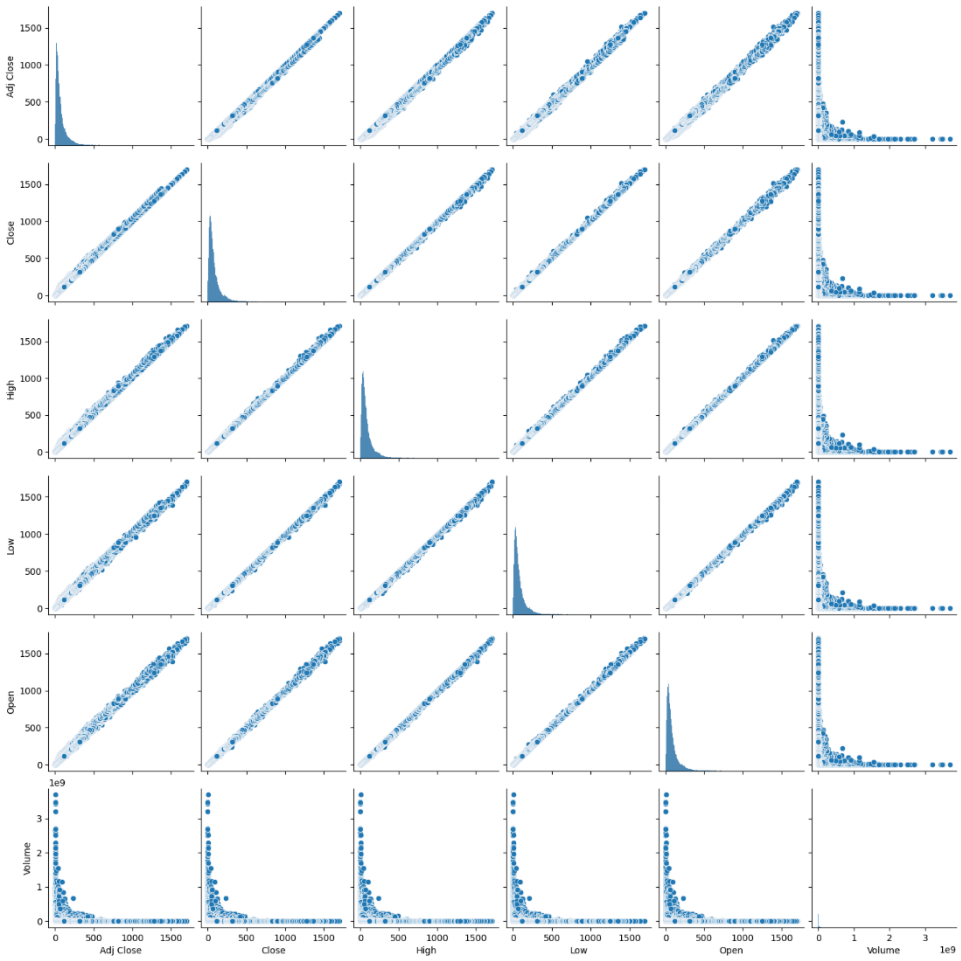
| | Adj Close | Close | High | Low | Open | \ |
|---|---|---|---|---|---|---|
| count | 617,831.00 | 617,831.00 | 617,831.00 | 617,831.00 | 617,831.00 | |
| mean | 79.67 | 87.47 | 88.42 | 86.48 | 87.46 | |
| std | 102.74 | 104.52 | 105.68 | 103.30 | 104.52 | |
| min | 0.20 | 0.22 | 0.23 | 0.22 | 0.22 | |
| 25% | 26.57 | 32.70 | 33.06 | 32.30 | 32.69 | |
| 50% | 49.82 | 59.14 | 59.72 | 58.50 | 59.12 | |
| 75% | 94.83 | 105.02 | 106.13 | 103.89 | 105.00 | |
| max | 1,702.53 | 1,702.53 | 1,714.75 | 1,696.90 | 1,706.40 | |

| | Volume |
|---|---|
| count | 617,831.00 |
| mean | 9,347,124.55 |
| std | 47,716,693.57 |
| min | 0.00 |
| 25% | 1,144,000.00 |
| 50% | 2,453,400.00 |
| 75% | 5,657,850.00 |
| max | 3,692,928,000.00 |

Según la imagen anterior, se puede observar:

- Adj close: El valor máximo (1,702.53) es significativamente mayor que el percentil 75 (94.83), lo que indica outliers positivos.
- Close: el valor máximo (1,702.53) es un outlier positivo es significativamente mayor que el percentil 75 (105.02).
- High: El valor máximo (1,714.75) es un outlier positivo en comparación con el percentil 75 (106.13).
- Low: El valor máximo (1,696.90) es un outlier positivo. Este valor es mucho mayor que la media y este lejos del percentil 75.
- Open: La media es mucho significativamente más baja que el máximo, desviación estándar y que el percentil 75.

Igualmente se obtuvo el detalle de la relación por pares entre las variables numéricas.




Según esta imagen, se puede observar que las variables estan altamebte correlacionadas, indicando que tienen un comportamiento parecido. Las variables *open*, *adj close*, *close*, *high*, y *low* tienen una alta correlación ya que son indicadores de los precios de las acciones. El volumen, no parece tener una relación con las variables de los precios; esto se puede dar ya que depende de la actividad del mercado más que de los precios.

## 3. DATA PREPARATION

Esta etapa se la realizo con el script 02_DataWrangling.ipynb y 03_FeatureEngineering.ipynb

## a. Limpieza y transformación

A continuación, detallo lo que se encontró en el dataset y el proceso a seguir para cada punto.

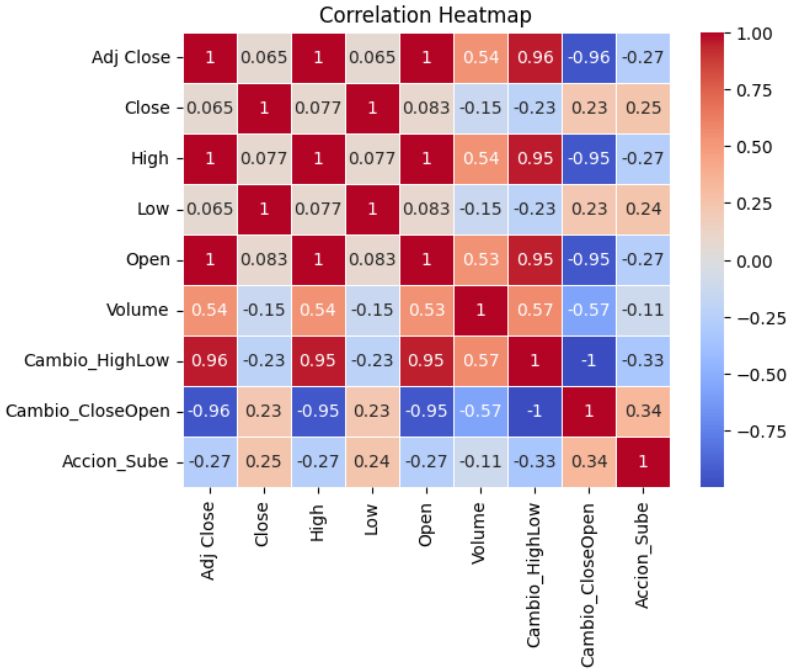
- Duplicados: no se encontraron.

- Valores Nulos: 67.33%

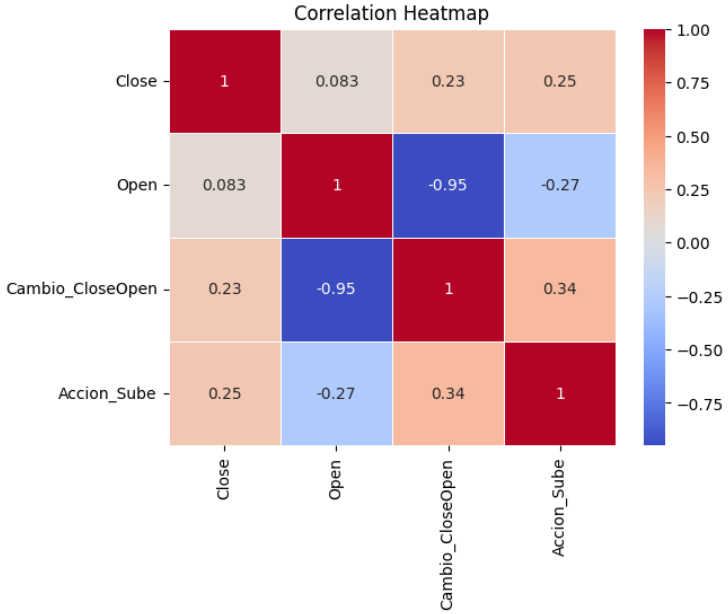
  En este caso el porcentaje de valores nulos considerablemente alto (67,33%) y dada la importancia de las columnas (*Adj Close*, *Close*, *High*, *Low*, *Open*, *Volume*), se optó por imputar valores. Se debe tomar en cuenta que la desviación estándar es muy alta por lo que rellenar con la media puede afectar el análisis, por lo que se rellenó con valores de días cercanos.

- Transformación: Se ajustaron 2 variables.
  *'Date'* a tipo Datetime y posterior a índice para poder entrenar el modelo y analizar la tendencia de días anteriores.
  *'Volume'* a tipo entero ya que es la cantidad total de acciones.

- Columnas eliminadas: Se analizo el porcentaje de correlación que existe entre las columnas.




Dada la alta correlación se optó por eliminar las columnas que generan ruido en el análisis (*Adj Close*, *Low*, *High*, *Volume*) y las que son irrelevantes para el análisis (*Symbol*, *Cambio_HighLow*). Finalmente quedando con las variables (*Close*, *Open*, *Cambio_CloseOpen*, *Accion_Sube*):



- Variables Derivadas: A partir de las variables originales presentes en el dataset (*open* y *close*), se generaron dos variables derivadas con el objetivo de proporcionar información relevante para predecir el movimiento de los precios de las acciones. La primera, *CambioOpenClose*, representa la diferencia entre el precio de apertura (*open*) y el precio de cierre (*close*), lo que permite observar cambios diarios en el valor de la acción. La segunda, una variable binaria denominada *Accion_Sube*, indica si la acción presentó una subida o una bajada según el valor de *Cambio_OpenClose*; si este es positivo, la acción sube, y si es negativo, la acción baja.

- Variables Lag: se crearon variables lag generan para capturar información histórica del comportamiento de los precios de las acciones y medir su impacto en la predicción de movimientos futuros. En este caso, se consideró el precio de apertura (*open*), precio de cierre (*close*), la diferencia entre ambos (*CambioOpenClose*) y la variable derivada binaria '*Accion_Sube'*. Para realizar diferentes pruebas sobre los modelos y determinar si la inclusión de datos históricos recientes o más antiguos afecta la capacidad predictiva del modelo, se trabajó con variables lag 5, 10 y 20 días.

## 4. MODELING

### a. Selección de modelos

La predicción sobre si una acción sube o baja representa un problema de clasificación, cuyo objetivo es asignar una etiqueta discreta: 1 si la acción sube, y 0 si baja. Para abordar este problema, se evaluarán diferentes algoritmos de clasificación con el fin de analizar su ajuste y desempeño en esta tarea específica. Entre los modelos considerados se incluyen:

- Regresión Logística
- Árboles de Decisión

- Random Forest
- Ensambladores de Boosting

Estos algoritmos han sido seleccionados debido a su capacidad para manejar problemas de clasificación y sus distintos enfoques en términos de interpretabilidad, robustez y eficiencia.

**b. Entrenamiento y validación**

El proceso de entrenamiento y validación de los modelos individualmente se realizó con una división del dataset (sin modificaciones) en proporciones del 80% para entrenamiento y 20% para prueba (train/test split). Sin embargo, los resultados obtenidos a través de esta metodología mostraron métricas de desempeño muy similares y con valores bajos, alcanzando una precisión aproximada de 0.52. Debido a esto, se optó por adoptar una estrategia alternativa enfocada en minimizar los falsos positivos y mejorar la calidad de las predicciones.

Para ello, se realizó una búsqueda exhaustiva de hiperparámetros en los cuatro modelos seleccionados, explorando distintas configuraciones mediante iteraciones sobre una lista definida en el script 04_Modeling.ipynb en la sección de Model Testing. Adicionalmente, se probaron las iteraciones con variables lag de 5, 10 y 20 días con el objetivo de identificar patrones temporales que pudieran mejorar la capacidad predictiva del modelo. El criterio de éxito definido para este enfoque fue alcanzar una precisión mínima del 70%.

La misma estrategia se aplicado al dataset sobre muestreado con el objetivo de analizar si había alguna diferencia significativa en las métricas de desempeño al balancear las clases. No obstante, se priorizó el análisis con el dataset original sin muestreo. El interés principal del análisis radica en obtener resultados que sean precisos y representativos, basados en datos reales y sin modificaciones. De esta manera, se garantiza que las predicciones del modelo reflejen de manera fiel el comportamiento observado en los datos originales, sin introducir sesgos asociados al proceso de muestreo.

Los resultados detallados de la ejecución del código y los valores obtenidos para las métricas de evaluación se encuentran documentados en la carpeta reportes/resultados.

### c. Métricas de evaluación

Para determinar cuál es el mejor modelo predictivo, se utilizaron las métricas de precisión, recall y F1-score. Se priorizó el enfoque en la precisión, dado que el objetivo principal es minimizar los falsos positivos y garantizar que, de los casos predichos como positivos (acciones que se predice que suben), la mayor proporción posible sea realmente positiva (es decir, que las acciones efectivamente suban). Esto es fundamental en el contexto de este proyecto, ya que un falso positivo podría llevar a decisiones de inversión erróneas, como la compra de acciones que no subirán, resultando en pérdidas financieras y costo de oportunidad al no aprovechar opciones más rentables. Las métricas recall y F1-score complementan el análisis ya que proporcionan una visión equilibrada de la sensibilidad y el desempeño general del modelo, pero la precisión se consideró clave para alinearse con el objetivo financiero del proyecto.

## 5. EVALUATION

### a. Análisis de desempeño

Como se menciono anteriormente todos los resultados de la ejecución del código se encuentra en la carpeta de resultados. En esta sección, adjunto los resultados del mejor modelo encontrado en cada ejecución con cada variable lag y su interpretación. Cabe mencionar que dado que los resultados con el dataset balanceado no tenia mayor diferencia, nos enfocaremos solo en el análisis de los resultados obtenidos con el dataset original.

- 5 días:

```
Resultados sin sobremuestreo - Lag 5 dias

Mejor modelo: RandomForestClassifier
Hiperparámetros: {'n_estimators': 300}
Exactitud: 0.8238450316442854
Recall: 0.581275766266356
Precision: 0.5350223239603634
F1: 0.5571907970683777
```

- 10 días:

```
Resultados sin sobremuestreo - Lag 10 dias

Mejor modelo: LogisticRegression
Hiperparámetros: {'max_iter': 500}
Exactitud: 0.8218496620900011
Recall: 0.5067385142780323
Precision: 0.5346184757942511
F1: 0.5203052838023592
```

- 20 días:

```
Resultados sin sobremuestreo - Lag 20 dias

Mejor modelo: RandomForestClassifier
Hiperparámetros: {'n_estimators': 100}
Exactitud: 0.8280535447669538
Recall: 0.5783936993759873
Precision: 0.5463569501031801
F1: 0.5619190665883019
```

los modelos con mejor desempeño en términos de precisión fueron la Regresión Logística y el Random Forest Classifier, alcanzando valores de 0.53 (Regresión Logística) y entre 0.53 y 0.54 (Random Forest). Estos resultados sugieren que, aunque ambos modelos presentan un rendimiento similar, ninguno logra alcanzar un nivel de precisión que sea satisfactorio

para su implementación en un contexto de negocios. También se observó que, al incluir un mayor número de variables lag (como 5, 10 y 20 días), hubo una ligera mejora en la precisión de los modelos, aunque no fue significativa ni suficiente para cumplir con las expectativas planteadas inicialmente.
Estas limitaciones resaltan que, a pesar de haber realizado una búsqueda exhaustiva de hiperparámetros y probar múltiples configuraciones, el objetivo de alcanzar una precisión del 70% no se logró. Esto evidencia la necesidad de considerar ajustes adicionales, como el uso de otros enfoques de modelado, análisis de tendencias a más largo plazo o un refinamiento en la selección de características, para abordar mejor el problema planteado.

**b. Factores de riesgo y éxito**

El análisis indica que el modelo con la mejor combinación de parámetros alcanza una precisión del 54%; logrando predecir correctamente el 54% de las acciones que suben, de las cuales solo el 57% son aciertos. Estas métricas presentan riesgos considerables, como decisiones de inversión erróneas debido a falsos positivos, lo que puede generar pérdidas de capital significativas. Además, existe un costo de oportunidad asociado a la falta de precisión del modelo, ya que impide identificar adecuadamente acciones con potencial real de subida.

Se observó que incorporar más variables lag mejora ligeramente la precisión del modelo; sin embargo, no se probaron valores de lag más amplios debido a las limitaciones computacionales. Un análisis con valores lag más altos podría ser una opción para mejorar el rendimiento del modelo, siempre y cuando se disponga de la infraestructura adecuada.

Otra limitación identificada fue el alto porcentaje de valores nulos en el dataset original, lo cual pudo haber afectado el análisis y el rendimiento del modelo. Para mitigar este riesgo, se sugiere evaluar la posibilidad de utilizar un dataset más confiable, con menor cantidad de valores nulos, y

repetir el análisis para determinar su impacto en la calidad de las predicciones.

## 6. PLAN DE IMPLEMENTACION

La implementación del proyecto contempla varios aspectos clave. A continuación, detallo los posibles pasos a seguir en caso de realizar el paso a producción del modelo propuesto.
En primer lugar, se debe contenerizar el modelo utilizando Docker, para garantizar su portabilidad y escalabilidad. Posteriormente, se debe considerar el diseño de una arquitectura de microservicios para organizar y optimizar las funcionalidades del sistema. Esta arquitectura debe incluir la creación de una API que facilite la interacción con el modelo. Finalmente, el entorno de producción se debe establecer en la nube, evaluando plataformas como AWS, Azure o Google Cloud, considerando sus características técnicas y sus costos de infraestructura.
Cabe mencionar que antes de realizar estos pasos, se debe analizar la complejidad, el tiempo y costos de desarrollo requerido para cada etapa.

## 7. CONCLUSIONES Y PROXIMOS PASOS

Los resultados obtenidos reflejan que el modelo predictivo tiene una precisión limitada, alcanzando un 54% de precisión. Aunque este logra identificar correctamente un 54% de las acciones que suben, solo el 57% de estas predicciones son acertadas, lo que significa que aproximadamente solo un 30% de las inversiones serían certeras. Esto implica que, si se basaran decisiones de inversión exclusivamente en estas predicciones, se incurriría en pérdidas de capital significativas debido a compras equivocadas y se enfrentaría el costo de oportunidad al no invertir en acciones con potencial de subida real.

Dada esta situación, el modelo en su forma actual no es útil para tomar decisiones de inversión confiables. Sin embargo, si se lograra mejorar su precisión, se podrían obtener los siguientes beneficios:

- **Incremento en el rendimiento del portafolio:** La capacidad de identificar con mayor certeza las acciones que suben optimizaría las estrategias de inversión.
- **Reducción de pérdidas:** Al minimizar decisiones incorrectas basadas en predicciones no precisas.
- **Maximización de oportunidades:** Aprovechando mejor las tendencias positivas en el mercado para tomar decisiones de inversión o venta.

Para mejorar los resultados, se sugiere:

- **Ajuste de los modelos actuales:** Afinar los parámetros y considerar otros algoritmos que puedan adaptarse mejor al problema de clasificación.
- **Incorporar tendencias de más largo plazo:** Analizar movimientos más amplios en lugar de limitarse a predicciones diarias.
- **Gestión de variables lag:** el incremento en la dimensionalidad podría requerir mayor capacidad de procesamiento, por lo que sería importante considerar infraestructura adecuada para manejar estos datos y ejecutar modelos más complejos.
- **Uso de datasets más confiables:** Garantizar que los datos utilizados sean de alta calidad y con menor proporción de valores nulos, evitando imputaciones que puedan afectar negativamente las predicciones.

## 8. APENDICES

a. Repositorio: https://github.com/gabyzumarraga/DS-SP500

b. Documentación técnica: El proyecto se realizó en 4 partes, los notebooks utilizados están ordenados por tema en la carpeta ‘notebooks/’ dentro de las fuentes del proyecto.
   - 01_EDA.ipynb
   - 02_DataWrangling.ipynb
   - 03_FeatureEngineering.ipynb
   - 04_Modeling.ipynb

Los datos fueron cargados desde el archivo sp500_stocks, utilizando la librería ‘pandas’ en cada uno de los archivos utilizados.

Análisis Exploratorio de Datos (EDA): Se realizaron visualizaciones y estadísticas descriptivas utilizando matplotlib y seaborn para comprender patrones iniciales y detectar anomalías.

Preparación de los Datos: Se ejecutaron procesos de limpieza (tratamiento de valores nulos y outliers) y transformaciones (escalado y normalización).

Entrenamiento de Modelos: Se probaron diferentes algoritmos de clasificación, como árboles de decisión, regresión logística y random forest, para predecir la dirección de los precios de las acciones.

Evaluación de Modelos: Se compararon las métricas de desempeño de cada modelo (precisión, recall y F1-score) para seleccionar el más eficiente.

Generación de Resultados: Se generaron informes por cada etapa del proyecto y se documentaron los resultados obtenidos con los modelos probados con distintos parámetros, detallando los rendimientos de los modelos.

c. Referencias y fuentes:

- Versión Python: 3.12.4
- Librerías:
    - pandas==1.5.3
    - numpy==1.23.5
    - matplotlib==3.6.2
    - seaborn==0.12.2
    - scikit-learn==1.1.3
    - jupyter==1.0.0