
UNIVERSIDAD SAN FRANCISCO

# FUNDAMENTOS DE CIENCIA DE DATOS

## PROYECTO #2
## CRISP-DM, FEATURE ENGINEERING AND MODEL RESEARCH

**GABRIELA ZUMARRAGA – 345769**

**QUITO 6 DE ABRIL 2025**

# PROYECTO #2
# CRISP-DM, FEATURE ENGINEERING AND MODEL RESEARCH

**Tema:** Análisis del desempeño de las empresas en el S&P (Standard and Poor's) 500 utilizando datos históricos del mercado bursátil

## 1. FEATURE ENGINEERING

Una vez finalizado el EDA y Data Wrangling procedemos con el proceso de Feature Engineering. El objetivo es realizar la predicción de si sube o no una acción, por ende el problema a resolver es de clasificación. Se trabajará únicamente con el archivo sp500_stocks.

### a. Crear, transformar y seleccionar variables (features) que aporten valor a su modelo.

El archivo cuenta con las siguientes características (sus descripciones se encuentran en el EDA):

- Date Symbol
- Adj Close
- Close
- High
- Low
- Open
- Volume

En el data Wrangling se procedió con la creación de variables derivadas como:

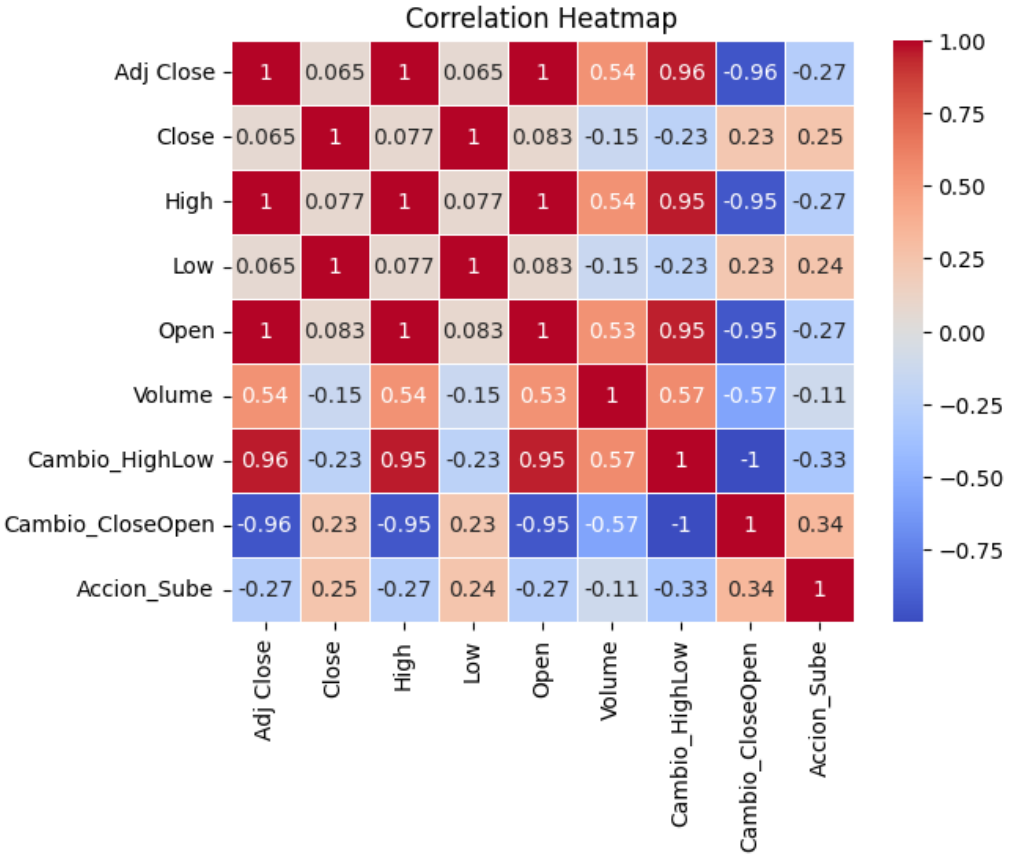
- Year: variable derivada de la fecha correspondiente al año.
- Month: variable derivada de la fecha correspondiente al mes.
- Day: variable derivada de la fecha correspondiente al día

Sin embargo, estas se eliminarán ya que se procederá más adelante con la creación de variables lag para manejar las características de fecha.

Adicional, se crearán variables temporales y derivadas que nos pueden servir para el análisis posterior, por lo que se crearon las siguientes.

- Cambio_HighLow: variable derivada de la diferencia entre el precio mas alto y el menor
- Cambio_CloseOpen: variable derivada de la diferencia entre el precio de apertura y de cierre.
- Accion_Sube: variable binaria para indicar si sube (1) o baja la acción (0).

Para seleccionar las variables con las que vamos a continuar el análisis, realizamos un gráfico de correlación.




Lo que se puede observar es que existe una alta correlación entre algunas variables (High, Low, Adj_Close, Volume, Cambio_HighLow, Year, Month, Day). Estas generan redundancia y no nos aportan al análisis que se quiere realizar, por lo que procedemos con su eliminación.

Lo siguiente es la creación de variables lag de 4 días anteriores para poder predecir el día siguiente. Se toma en cuenta el precio de apertura, precio de cierre, cambio_CloseOpen y Accion_Sube.

**b. Justificar por qué cada nueva variable (o transformación) podría mejorar el desempeño.**

La eliminación de las columnas Adj Close, High, y Low se justifica por redundancia o simplificación del análisis. Estas variables tienen una alta correlación por lo que podemos eliminarlas para evitar ruido en el análisis. Las columnas High y Low se reemplazan efectivamente con la nueva variable Cambio_HighLow. Esta tiene una correlación con la columna Cambio_CloseOpen por lo que también se la puede eliminar. Estas columnas de cambio capturan la volatilidad diaria de las acciones en un solo valor, simplificando el análisis sin perder información relevante.

Para poder realizar una predicción de si sube o baja una acción, se crearon variables lag para registrar los datos de los precios de apertura, cierre y cambios de los precios, de los 4 anteriores días. Igualmente se crea una variable binaria adicional que nos indica si ese día la acción subió (1) o bajo (0), dependiendo del cambio que haya tenido con respecto a las variables Open y Close en el periodo (instancia) analizada.

## 2. MODELING

**a. Probar diferentes modelos según su problema:**

La predicción de si una acción sube o baja es un problema de clasificación, ya que el objetivo es asignar una etiqueta discreta (por ejemplo, 1 si sube y 0 si baja). Por lo que se probaran algunos modelos para analizar que tan bien se ajustan al problema. Por ejemplo, se probarán: Regresión Logística, Árboles de Decisión, Random Forest, Ensembles de Boosting.

**b. Comparar resultados utilizando métricas adecuadas (ej. RMSE, MAE, R2 para regresión; accuracy, F1-score, AUC, etc., para clasificación).**

Adjunto los resultados obtenidos con cada modelo.

- **Regresión Logística**

```
La recuperacion con datos de validacion es de: 0.997673065735893
La precision con datos de validacion es de: 0.9894298724579152
El score del modelo con datos de validacion es de:  0.9966625638451536
```

- **Árboles de Decisión**

```
El total de acciones que se predice que suben es: 96264
El total de acciones de validaciones: (374539,)
El total de acciones de realemente suben es: 384305

La recuperacion con datos de validacion es de: 1.0
La precision con datos de validacion es de: 1.0
El f1-score con datos de validacion es de: 1.0
El score del modelo con datos de validacion es de: 1.0
```

- **Random Forest**

```
El total de acciones que se predice que suben es: 96264
El total de acciones que suben y se valida: (374539,)
El total de acciones que realemente suben es: 96264

La recuperacion con datos de validacion es de: 1.0
La precision con datos de validacion es de: 1.0
El f1-score con datos de validacion es de: 1.0
```

- **Ensembles de Boosting**

```
El total de acciones en validacion es: (374539,)
El total de acciones de realemente suben es: 96264
El total de acciones que se predice que suben es: 96264

La recuperacion con datos de validacion es de: 1.0
La precision con datos de validacion es de: 1.0
El f1-score con datos de validacion es de: 1.0
```

c. **Documentar sus hallazgos y escoger el modelo más prometedor, justificando su elección.**

Para probar los modelos, verificamos primero si las clases están balanceadas. En este caso hay un desbalance, el porcentaje de las acciones que suben es de 74.34% y las que bajan 25.66%. Para evitar afectar el rendimiento de los modelos de clasificación, y las métricas, se realizó un sobre muestreo las clases con una tasa de 2.

Los modelos de Random Forest, Arboles de decisión y Ensemble Boosting tienen métricas perfectas. Esto se debe volver a analizar , ya que puede estar cayendo en un overfitting.

Dado que ya tienes variables lag y derivadas que capturan relaciones temporales, el mejor modelo es Random Forest. Este modelo es robusto y maneja bien los datasets con características derivadas. Igualmente, es más robusto frente al sobreajuste y ofrece un buen balance entre rendimiento y complejidad.