

2023 - 2024

# 

**Máster en Big Data y Data Science**

**05MBID – Estadística Avanzada**

**Nombre: Brahian Andrey Giraldo Alzate**

**Fecha: 08/30/2023**

**Curso 2022 – Ed. Abril**

**ACTIVIDAD GUIADA 1**

Contenido

[1](#_Toc137419356)

[1 Introducción, motivación y objetivo (Entendimiento del dominio) 3](#_Toc137419357)

[1.1 Introducción 3](#_Toc137419358)

[1.2 Motivación 3](#_Toc137419359)

[1.3 Objetivo 3](#_Toc137419360)

[2 Fuente de datos y selección de datos 5](#_Toc137419361)

[2.1 Fuente de datos 5](#_Toc137419362)

[2.2 Selección de los datos 6](#_Toc137419363)

[3 Preparación, limpieza y transformación 9](#_Toc137419364)

[3.1 Preparación 9](#_Toc137419365)

[3.2 Limpieza 11](#_Toc137419366)

[3.3 Transformación 11](#_Toc137419367)

[4 Data mining y método de evaluación. 13](#_Toc137419368)

[4.1 Análisis de sentimiento: 13](#_Toc137419369)

[4.2 Modelos de pronósticos: 13](#_Toc137419370)

[4.3 Método de evaluación: 15](#_Toc137419371)

[5 CONCLUSIONES 15](#_Toc137419372)

[6 ANEXOS 16](#_Toc137419373)

[6.1 Anexo 1: visualización de los campos existentes en la base de datos descargada desde el datalake de twitter. 16](#_Toc137419374)

# Introducción, motivación y objetivo (Entendimiento del dominio)

La aplicación de técnicas de análisis de regresiones juega un papel fundamental en la búsqueda de patrones y tendencias en los datos financieros. Se realizará un análisis exhaustivo de datos financieros, utilizando herramientas estadísticas para modelar relaciones y construir estrategias de trading automatizadas.

El propósito de este trabajo es explorar cómo las técnicas de regresión pueden ser empleadas en el ámbito del trading algorítmico.

El análisis de los datos se llevará a cabo utilizando el lenguaje de programación R, una herramienta versátil que permite realizar cálculos estadísticos precisos y modelado avanzado. Se proporcionará un script con instrucciones paso a paso para reproducir cada fase del análisis.

En el ámbito del trading, las regresiones pueden revelar patrones ocultos en los datos históricos, lo que puede llevar a la generación de señales de compra y venta más informadas. Además, investigaremos cómo las regresiones logísticas pueden ser útiles en la clasificación de eventos financieros y la predicción de movimientos del mercado.

## Motivación

Automatizar herramientas de análisis utilizadas en el mercado financiero, como el análisis fundamental y técnico, que se enfoca en el análisis de los precios en la bolsa de valores y sus volúmenes de venta para predecir su comportamiento e inferir probabilidades a favor del inversor.

## Objetivo

A aquellos objetivos de inversión validados previamente por un proceso NLP se les aplicará un análisis de pronóstico con regresiones. Se aplicarán diversos modelos de regresión (se puede aplicar regresión logística) y otros pronósticos al histórico de la base de datos.

# Fuente de datos y selección de datos

## Fuente de datos

Se hizo un pre-procesamiento y limpieza de los datos en la asignatura de minería de datos en de la cual se puede destacar lo siguiente:

Se unieron 7 bases de datos (6 de yahoo finance, y 1 de twitter).

Inicialmente, entre todos los conjuntos de datos se tienen 10.574 filas. Entre las cuales, 2574 filas pertenecen a las bases de datos de las 6 empresas con 7 columnas cada empresa, y 8000 filas en la base de datos de twitter con 99 columnas cada fila.

La unificación de las bases de datos de las 6 empresas tiene como resultado 625 filas que incluye las columnas Open, High, Low y Close de cada empresa entre el año 2021 y el año 2023. Estas bases de datos se mezclaron usando cómo indice la fecha, y se generó como resultado una base de datos de 625 filas y 24 columnas.

Después de unir la base de datos de twitter y la de las empresas, al final del preprocesamiento y calidad de los datos se obtuvo una tabla .csv con 5838 registros y 48 columnas o atributos.

## Selección de datos (variables)

Hay 5 variables independientes relacionadas al precio, y son el precio de cierre del día para cada una de las 5 empresas.

Hay 5 variables independientes relacionadas a la demanda y son el volumen de ventas del día para cada una de las 5 empresas.

A pesar de que la base de datos tiene 48 atributos, dichas variables independientes son suficientes para desarrollar las regresiones deseadas.

Las regresiones servirán para añadir un criterio de decisión de compra o venta al logaritmo de trading que se busca desarrollar. Esa regresión se validará con un proceso NLP (en estudios posteriores a este), para validar si los pronósticos de regresión coinciden con las recomendaciones de trading ubicadas en la base de datos a la que se aplicó minería de datos.

**¿Qué datos son discretos y cuáles continuos?**

Son 48 campos diferentes, de los cuales no se utilizarán todos, por lo cual se clasifican a continuación los más útiles que son tipo float (valores continuos), int (valores enteros), datetime (fecha).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Date (datetime)** | **Close\_AAPL (float)** | | **Volume\_AAPL (int)** | | **Close\_COIN (float)** | | **Volume\_COIN (int)** | |
| 5/6/2022 | 157,279999 | | 116124600 | | 103,739998 | | 9026500 | |
|  |  | |  | |  | |  | |
| **Close\_GEHC (float)** | | **Volume\_GEHC (int)** | | **Close\_RUN (float)** | | **Volume\_RUN (int)** | | **Close\_TMUS (float)** | | **Volume\_TMUS (int)** |
| 76,900002 | | 1751400 | | 23,41 | | 9343700 | | 126,800003 | | 4701900 |

Los anteriores atributos son necesarios como variables independientes que se usarán para calcular los valores de regresión.

Los demás atributos de la base de datos de minería que no se explican acá, se pueden ver en el anexo 1. Estos atributos no tienen mucho valor para las regresiones, pero si sirven para un proceso de validación de tendencias NLP posterior a las regresiones.

A continuación, con se describen los atributos útiles para la regresión que se pueden encontrar en la base de datos pre-procesada:

Date: Esta columna representa la fecha en la que se registraron los precios en yahoo finance. Permite alinear la fecha de inversión con la fecha de sugerencia de tendencia alcista en twitter para el posterior proceso de validación NLP.

Close: representa el precio de cierre de las acciones de la empresa al final del día de negociación. Hay uno precio de cierre para cada una de las empresas (5 empresas).

Volume: indica el volumen de acciones negociadas durante el día. Representa la cantidad total de acciones compradas y vendidas para cada una de las empresas (5 empresas).

Todos los datos se usan para desarrollar un diagrama lineal, cuyo ejemplo de visualización sería el siguiente:

# Preparación, limpieza y transformación

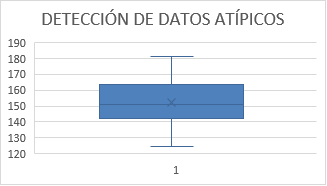
## Preparación

**¿Cuáles parecen ser features importantes? ¿Cuáles podemos descartar? ¿Por qué crees que tienes que descartar estas características?**

Los campos señalados anteriormente son los features importantes, porque son los que en principio generan valor agregado al caso de estudio. El resto de campos que no coinciden con los campos resaltados y que se ubican en el anexo 1 se pueden descartar, porque no generan valor agregado al trabajo actual. Esto permite simplificar y mejorar el análisis, aumentar la precisión de los modelos y facilitar la interpretación de los resultados.

¿**Cuáles son los Outliers? (unos pocos datos aislados que difieren drásticamente del resto y “contaminan” ó desvían las distribuciones) – Podemos eliminarlos? ¿es importante conservarlos? – son errores de carga o son reales?**

En el siguiente gráfico, con respecto al rango máximo y mínimo de los datos, podemos observar que de momento no hay ningún punto, valor de las acciones o dato fuera de los cuartiles esperados.



Se hizo el análisis 3-sigma, para descartar valores atípicos entre todos los datos, pero el análisis indica que no hay ningún dato a una distancia de 3 desviaciones de la media. En caso de encontrar un dato atípico se reemplazaría por el promedio entre el valor de la acción el día anterior y el día siguiente.

No se han observado valores faltantes en los campos o atributos elegidos para cumplir con el objetivo de investigación.

**Integración de datos:** se deben hacer un join entre los datos que coincidan con las fechas mencionadas. Aquellas fechas de los valores de las acciones registrados en yahoo finance que no tengan una mención en las cuentas de twitter que indiquen intención de inversión, se tomarán como objetivo de inversión no válido según el criterio de análisis de tendencias alcistas en el mercado.

**¿Hay correlación entre features (características)?**

La idea del caso de estudio con respecto a la combinación de la base de datos de twitter con la de yahoo finance, es precisamente, determinar en qué casos existe esta correlación. Por lo cual se pueden analizar casos determinados y ver si cuando se afirma que es recomendable invertir en una acción, tiene coherencia con las tendencias que presenta el valor de las acciones en yahoo finance.

**¿Estamos ante un problema dependiente del tiempo? Es decir, un TimeSeries**

Si, el valor de las acciones que registra yahoo finance, depende del tiempo y de otros factores, pero por lo regular, el valor siempre depende del día, e inclusive de la hora en la que se definen los precios. Si se realiza un análisis más profundo, se puede identificar que los análisis pueden ser más volátiles, buscando así variaciones en periodos de tiempo en minutos o segundos inclusive.

**¿Tenemos posible sesgo de datos? (por ejemplo, perjudicar a clases minoritarias por no incluirlas y que el modelo de ML discrimine)**

## Limpieza

Se eliminarán de la colección, aquellos datos que no correspondan a los datos campos mencionados anteriormente.

Si se llegase a encontrar un dato atípico en el valor de las acciones, se reemplazaría por el promedio entre el valor de la acción el día anterior y el día después.

Aquellos valores de las acciones que no tengan mención en una fecha específica correspondiente en twitter, obtendrán un nuevo campo que indique que la inversión en ese periodo no debería ser aconsejable según el criterio de análisis de tendencias alcistas o bajistas. El hecho de que una acción determinada no sea recomendada en una fecha determinada, no indica que sea mala opción de inversión, simplemente indica que

## Transformación

**¿Puedo transformar alguna variable para generar información adicional?**

Las fechas de la colección de mongodb están en formato texto, por lo cual se deberán pasar a formato fecha para operar con los tiempos en caso de ser necesario.

Los valores de la base de datos de yahoo finance, se pueden transformar para generar información adicional. Para este caso de estudio, se utilizará el precio del campo Adj close, puesto que es el precio destinado a calcular retornos históricos con mayor precisión, puesto que ha recibido ajustes correspondientes a diferentes eventos que no se tienen en cuenta en los demás campos, como dividendos, desdoblamiento de acciones, etc.

Este valor se puede transformar a una variable predictiva a través de una granvariedad de modelos de pronóstico. Cada modelo de pronóstico puede arrojar un resultado distinto, por lo cual a partir de la columna Adj close se pueden derivar numerosas variables objetivo para hacer forecasting.

Se pueden llevar a cabo procesos de binning o discretización para identificar patrones no lineales o para simplificar el análisis en ciertos casos, por ejemplo, con patrones mariposa, que son muy conocidos en el trading.

**¿Conozco conocimiento experto para incorporarlo en el dataset? ¿Cómo lo puedo modelar?**

En el caso de los modelos de predicción con respecto a los datos históricos de la columna Adj close, tengo el conocimiento para aplicar diferentes modelos de predicción. Entre ellos el análisis de promedio móvil es una opción de análisis muy frecuente en el trading, y el patrón de mariposa es de alto interés para mi, aprender como desarrollarlo.

En el caso del promedio móvil se deben realizar los siguientes pasos para desarrollar el modelo:

* **Selección de un período de tiempo:** Se debe determinar el período de tiempo que se utilizará para el análisis. Por ejemplo, se puede elegir utilizar datos diarios, semanales o mensuales, dependiendo de las necesidades y el horizonte de pronóstico. Para este caso de estudio aplicaríamos un periodo diario.
* **Cálculo del promedio móvil:** Se aplica el método del promedio móvil para suavizar los datos y identificar tendencias. El promedio móvil se calcula sumando los valores de un número determinado de períodos y dividiendo el resultado entre ese mismo número. Por ejemplo, para un promedio móvil de 5 días, se sumarían los valores de los últimos 5 días y se dividirían entre 5.
* **Representación gráfica del promedio móvil:** Se debe graficar el promedio móvil junto con los datos de precios de las acciones. Esto ayudará a visualizar las tendencias y patrones en los datos.
  + Visualmente, la gráfica tendría un aspecto parecido al siguiente:
    - Para el caso de predicción del valor de las acciones, en vez se usaría el valor de las acciones, y se añadiría una tercera gráfica correspondiente al cálculo de los valores futuros que permitirán al trader tomar decisiones de acuerdo a esta estrategia. La gráfica roja, indica los valores del promedio móvil, y la gráfica azul lo valores reales.
* **Interpretación de los resultados:** Se analiza la relación entre los valores de los precios de las acciones y el promedio móvil. Si el precio de las acciones cruza por encima del promedio móvil, puede ser una señal alcista, mientras que si cruza por debajo, puede ser una señal bajista.
* **Realización de pronósticos:** Se utiliza el promedio móvil para hacer pronósticos futuros del valor de las acciones. Se puede extrapolar la tendencia identificada y estimar posibles valores futuros.

**¿Siguen alguna distribución?**

Si, La distribución de los precios de las acciones puede variar según la empresa y el período de tiempo considerado. Sin embargo, es común observar una distribución asimétrica con colas pesadas, lo que refleja la naturaleza volátil y dinámica del mercado de valores.

# Data mining y método de evaluación.

## Análisis de sentimiento:

El análisis de sentimiento de un texto es una técnica utilizada para determinar la actitud emocional expresada en un texto, generalmente en términos de positividad, negatividad o neutralidad. Para realizar dicho análisis se definieron los siguientes pasos:

1. Preprocesamiento de texto: Limpiar y procesar el texto para eliminar ruido y elementos irrelevantes, como signos de puntuación, números, stopwords (palabras comunes sin significado específico) y convertir el texto en minúsculas.
2. Tokenización: Dividir el texto en unidades más pequeñas llamadas tokens, como palabras individuales o frases cortas.
3. Construcción de un diccionario de sentimientos: Crear un diccionario que asocie palabras o frases con un valor de sentimiento predefinido, como positivo, negativo o neutral. Esto se conoce como un lexico de sentimientos.
4. Asignación de puntuaciones de sentimiento: Asignar puntuaciones de sentimiento a cada token del texto en función de las palabras o frases presentes en el diccionario de sentimientos. Estas puntuaciones pueden ser valores numéricos, binarios o categóricos.
5. Agregación de puntuaciones: Calcular una puntuación agregada para todo el texto sumando o promediando las puntuaciones de sentimiento de los tokens individuales. Esto proporciona una medida general del sentimiento expresado en el texto.
6. Análisis e interpretación: Interpretar la puntuación agregada para determinar si el texto tiene un sentimiento positivo, negativo o neutral. También se pueden realizar análisis adicionales, como identificar las palabras o frases más influyentes en el sentimiento expresado.

Cálculos matemáticos asociados:

Puntuación de sentimiento promedio: Se calcula sumando todas las puntuaciones de sentimiento de los tokens y dividiendo entre el número total de tokens.

Porcentaje de sentimiento positivo/negativo: Se calcula dividiendo el número de tokens con sentimiento positivo/negativo por el número total de tokens y multiplicando por 100.

Frecuencia de palabras positivas/negativas: Se calcula contando el número de palabras con sentimiento positivo/negativo presentes en el texto.

Puntuación de sentimiento ponderada: Se puede asignar un peso a las puntuaciones de sentimiento de acuerdo con la importancia relativa de las palabras en el texto. La puntuación ponderada se calcula multiplicando cada puntuación de sentimiento por su peso correspondiente y sumando los productos resultantes.

## Modelos de pronósticos:

Se define el modelo de promedio móvil como predictor que permitirá validar la tendencia alcista o bajista del mercado mencionada en las cuentas de twitter. Adicionalmente, para asegurar que se cumplen los requisitos del modelo seleccionado se escoge un modelo propuesto en el pdf guía de la actividad.

Los modelos de pronósticos suelen hacer parte de un sistema que permite el análisis de los datos, en este caso, se ve de la siguiente forma:

Diagrama, Escala de tiempo

Descripción generada automáticamente

De lo cual se puede observar que de las fuentes de datos se extraen los datos necesarios. De yahoo finance los valores de las acciones de distintas entidades u activos financieros y de twitter appi, se extraen los valores de los campos que permiten validar la credibilidad del tweet y analizar el texto con un análisis de sentimiento para determinar si existe intención de compra o inversión en la acción a analizar.

Podría ser interesante agregar un análisis de demanda de las acciones basado en el volumen de los datos proporcionado por yahoo finance, ya que esto nos permitirá predecir a partir de la demanda de acciones cuando subirán y cuando bajarán los precios. Con esto se lograría la integración de dos enfoques distintos de pronóstico al mismo tiempo en las fechas correspondientes al análisis.

**Promedio móvil**

El uso del promedio móvil en el entorno de los algoritmos en Big Data puede ayudar a identificar patrones, realizar pronósticos y tomar decisiones basadas en el comportamiento histórico de la serie temporal analizada.

*=*

**La regresión lineal**

Es un algoritmo de aprendizaje supervisado que se utiliza en Machine Learning y en estadística. En su versión más sencilla, lo que haremos es “dibujar una recta” que nos indicará la tendencia de un conjunto de datos continuos (si fueran discretos, utilizaríamos Regresión Logística).

El modelo matemático para este caso sería el siguiente:

*t =*

Los parámetros a y b se estiman mediante el método de mínimos cuadrados. Consiste en encontrar los valores para los parámetros a y b, tal que la suma de los errores al cuadrado sea mínima.

Para el cálculo de los errores del pronóstico se suelen utilizar los siguientes indicadores:

ERROR PROM. (Error Promedio): Es la diferencia promedio entre el pronóstico y el valor real de la demanda. Se calcula sumando todos los errores individuales y dividiéndolos por el número total de observaciones. Proporciona una medida general de la dirección y magnitud del sesgo en los pronósticos.

MAD (Desviación Absoluta Media): Es la media de los valores absolutos de los errores entre el pronóstico y el valor real de la demanda. Calcula la magnitud promedio de los errores, sin tener en cuenta su dirección. Cuanto menor sea el valor de MAD, mejor será la precisión del pronóstico.

ECM (Error Cuadrático Medio): Es la raíz cuadrada de la media de los errores al cuadrado entre el pronóstico y el valor real de la demanda. El ECM tiene en cuenta tanto la dirección como la magnitud de los errores, y penaliza los errores grandes. Cuanto menor sea el valor de ECM, mejor será la precisión del pronóstico.

MAPE (Error Porcentual Absoluto Medio): Es el promedio de los errores porcentuales absolutos entre el pronóstico y el valor real de la demanda. Se calcula como el promedio de los valores absolutos de los errores porcentuales, dividido por el número total de observaciones, y se expresa como un porcentaje. El MAPE permite evaluar la precisión del pronóstico en términos de porcentaje y es útil para comparar la precisión entre diferentes series temporales.

MAPE' (Error Porcentual Absoluto Medio Ajustado): Es similar al MAPE, pero ajusta el cálculo para evitar divisiones por cero cuando los valores reales son cercanos a cero. El MAPE' es especialmente útil cuando se trabaja con series temporales que tienen valores cercanos a cero o valores negativos.

## Método de evaluación:

Para evaluar si el modelo del análisis de sentimiento obtuvo resultados confiables, se seleccionan los datos correspondientes a las fechas elegidas en la mención del tweet caso de estudio, y se valida que la tendencia mencionada corresponda con la tendencia de los datos mostrada en el análisis de promedio móvil o de regresión lineal.

# CONCLUSIONES

Revisando las posibilidades con respecto al aporte del big data en la vida diaria de los traders, se puede observar la posibilidad de automatizar muchas tareas diarias que permite a los traders enfocarse en aquello que realmente les genera ganancias sin dejar de lado el hecho que siempre deben tomar decisiones informadas con respecto al mercado, el análisis técnico enfocado al crecimiento de las empresas.

Es muy importante asegurar la calidad de los datos, para obtener resultados confiables.

La combinación de varias fuentes de datos puede generar un valor agregado inesperado, puesto que incluye en los análisis, perspectivas diferentes sobre el mismo problema.

Las redes sociales a pesar de tener información cuestionable en algunos casos, puede ser útil y confiable si se valida la calidad de la información generada a través de análisis de credibilidad del texto, por ejemplo.

# ANEXOS

## Anexo 1: visualización de los campos existentes en la base de datos descargada desde el datalake de twitter.

Los siguientes campos corresponden a la base de datos generada con procesamiento de datos en la clase de minería de datos:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **\_id** | **user.followers\_count** | **user.friends\_count** | **user.favourites\_count** | **retweet\_count** | **favorite\_count** | **created\_at** |
| 645a831bf55dd5c337018609 | 234157 | 771 | 4211 | 10 | 54 | Fri May 06 10:09:07 +0000 2022 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **text** | **user.name** | **user.screen\_name** | **user.description** | **user.url** |
| This is an important part of the Archegos / GSX / $GOTU story. As GAX squeezed higher, desks told us there was a la… https://t.co/bPBYLz2cN0 | MuddyWatersResearch | muddywatersre | Activist short seller, skeptic, First Amendment advocate, foot soldier in the Global War to Defend Truth | https://t.co/bZSmfKyo8x |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Date (datetime)** | **Open\_AAPL (float)** | **High\_AAPL (float)** | **Low\_AAPL ((float)** | **Close\_AAPL (float)** | **Adj Close\_AAPL (float)** | **Volume\_AAPL (int)** | **Open\_COIN (float)** | **High\_COIN (float)** | **Low\_COIN (float)** |
| 5/6/2022 | 156,009995 | 159,440002 | 154,179993 | 157,279999 | 156,34642 | 116124600 | 112,5 | 112,5 | 100,25 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Close\_COIN (float)** | **Adj Close\_COIN (float)** | **Volume\_COIN (int)** | **Open\_GEHC (float)** | **High\_GEHC (float)** | **Low\_GEHC (float)** | **Close\_GEHC (float)** | **Adj Close\_GEHC (float)** | **Volume\_GEHC (int)** |
| 103,739998 | 103,739998 | 9026500 | 76,860001 | 77,540001 | 76,264999 | 76,900002 | 76,870926 | 1751400 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **High\_RUN (float)** | **Low\_RUN (float)** | **Close\_RUN (float)** | **Adj Close\_RUN (float)** | **Volume\_RUN (int)** | **Open\_TMUS (float)** | **High\_TMUS (float)** | **Low\_TMUS (float)** | **Close\_TMUS (float)** |
| 25,200001 | 22,66 | 23,41 | 23,41 | 9343700 | 128,259995 | 129,490005 | 125,089996 | 126,800003 |