**UNIVERSIDAD EAFIT**

ÉNFASIS EN CIENCIA DE DATOS

PROYECTO INTEGRADOR 2023

Presenta:

**JUAN CAMILO FRANCO ROJAS**

Estudiante de Finanzas y Administración De Negocios

Medellín, 2023

1. **Introducción**

Se define como campo de la semántica el estudio relacionado con el significado de las palabras, se trata de un conjunto de palabras que comparte ciertos rasgos en su significado, por ejemplo, la palabra “tren” se encuentra dentro del campo semántico de la palabra transporte, del mismo modo que las lilas se encuentran dentro del campo semántico de las flores. (Blasco., 2022)

Por otro lado, la sintaxis estudia los principios de ordenamiento y combinación de las palabras y de los conjuntos que estas forman en una oración. (Concepto, 2023)

Las palabras no operan de manera individual y aislada, sino en una cadena o ristra, unas antes que otras, ya que su significado propio puede alterarse del lugar que ocupen en la oración. De hecho, la sintaxis del español establece un orden por defecto en el que se formulan las oraciones, que conocemos como SVP: Sujeto-Verbo-Predicado. (Concepto, 2023)

Permíteme señalar que todo lo anterior importa y mucho cuando se trata de clasificar sentimientos o bien se usen modelos básicos como un diccionario o muy robustos como las redes neuronales. El contexto importa, del mismo modo importa el vocabulario del área en cuestión, no es lo mismo usar un modelo entrenado para clasificaciones generales que uno especializado en un área o campo, como bien puede ser la medicina o la que atañe este trabajo las finanzas.

1. **Marco teórico o de referencia**

El análisis de lenguaje en el ámbito financiero representa una tarea de notable complejidad, que demanda un entendimiento profundo del sector. Este estudio se fundamenta en la investigación realizada por Dogu Tan Araci, plasmada en su tesis titulada "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models" (Araci, 2019)

Araci descubrió en su trabajo que los modelos de clasificación genéricos a menudo pecan de falta de contexto financiero. Esta característica puede introducir un sesgo, que se traduce, a la postre, en clasificaciones erróneas. Basándonos en su tesis, decidimos emplear el modelo FinBERT, que es una variante del ampliamente conocido modelo BERT (Araci, 2019).

1. **Desarrollo Metodológico**

**Fase 1:** Recolección de datos

Se recolectan los datos con el uso de la API de la SEC, lo anterior se realizará con el código de usuario de GitHub "alions7000”

**Fase 2:** Entendimiento de datos – Análisis de datos

No todo en un documento 10-K es esencial, se debe extraer la información estrictamente necesaria para el propósito del proyecto, esta información dentro de los 10-K está contenida en los temas 1, 1A, 7 y 7A

**Fase 3:** Definición de Objetivos desde la Perspectiva Analítica (descrito con mayor detalle en la Parte 2, "Enfoque Analítico")

En esta fase, transformamos los objetivos en una forma que nos permita identificar qué modelos o herramientas son los más relevantes para el problema en cuestión. Evaluamos múltiples opciones y tomamos una decisión, teniendo en cuenta las limitaciones de hardware. Aunque algunos modelos pueden ofrecer mejores resultados en teoría, no todos son viables. El que hacer bajo esta restricción se reglamenta en el apartado anterior de “Evaluar riesgos y contingencias”

**Fase 4:** Hallazgo

Desde esta fase, es posible establecer la implementación de los hallazgos logrados, desarrollar un plan de comunicación y presentar los resultados a los interesados. El informe resultante debe incorporar, pero no se limita a:

* Metodología Descriptiva: Proporcionamos un resumen del método analítico aplicado, dando detalles de los modelos que se han utilizado y las razones de su elección. Explicamos cualquier modificación realizada a estos modelos y los motivos de dichas modificaciones.
* Exposición de Resultados: se trasmiten los hallazgos del análisis. Para ello se debe hacer de una manera comprensible para los interesados, sin importar su nivel de conocimiento técnico.
* Interpretación de los Resultados: En esta parte, se aclara lo que los hallazgos representan en función de los objetivos del proyecto. Se debate cualquier implicación o conclusión que podamos extraer de los resultados.
* Recomendaciones: A partir de los hallazgos, se brindan sugerencias para los pasos a seguir. Estos podrían abarcar desde la implementación inmediata de los resultados hasta la necesidad de realizar más investigaciones o análisis.

**Fase 5:** Conclusiones:

Finalmente, concluimos con un resumen de los hallazgos y las recomendaciones, garantizando que los interesados tengan una comprensión nítida de lo que se ha logrado.

**Fase 6:** Retro alimentación

Con el propósito de realizar un análisis interno del proyecto, se plantean las siguientes preguntas: ¿Qué aspectos se pueden mejorar? ¿Qué se pudo haber hecho de mejor manera? ¿Qué elementos no se tomaron en cuenta? Estas cuestiones se formulan con el objetivo de identificar y anotar recomendaciones que contribuyan a la mejora del proyecto o como aprendizaje para nuevos proyectos.

**Presupuesto**

Se dispone de 200,000 pesos colombianos, que serán considerados como gastos de emergencia. Este proyecto es de naturaleza académica, por lo tanto, su perspectiva no es la de generar o gastar recursos monetarios. No obstante, este presupuesto se establece como solución a cualquier imprevisto.

**Bloomberg Terminal**

La Universidad EAFIT, en su sede campus principal en la ciudad de Medellín, Colombia, dispone de este recurso. El terminal cuenta con la información financiera de las empresas del índice bursátil S&P 500, entre muchos otros.

**API SEC**

La Comisión de Bolsa de Estados Unidos (SEC, por sus siglas en inglés) proporciona acceso a su base de datos EDGAR a través de su API. El usuario "alions7000" en GitHub nos permite descargar con su código la base de datos.

**Google Colab**

Google Colab es un servicio gratuito y pago en la nube con el cual se implementará y desarrollará todo el código pertinente al proyecto.

**Amazon Web Servicie (AWS)**

Es una plataforma que integra varios servicios desde la nube. Sus servicios son esenciales para este proyecto, destacándose servicios como almacenamiento, inteligencia artificial, aprendizaje automático, entre otros.

Para el desarrollo del proyecto, se tiene un laboratorio, en este se tiene un presupuesto de 100 dólares americanos.

Como oportunidad de mejora nuestro proyecto se puede trabajar desde Amazon SageMarker **<En construcción>** El SDK de Python de SageMaker JumpStart Industry tiene licencia bajo la licencia Apache 2.0

**Computadora Acer Nitro 5**

Para el proyecto se cuenta con una computadora con las siguientes especificaciones:

* Sistema operativo Windows 11 en su última versión.
* Tarjeta gráfica Nvidia GeForce GTX 1650.
* Procesador Inter Core I5 de novena generación
* Disco duro SSD de 200 GB.
* Disco duro HDD de 1 TB.
* Memoria RAM 12 Gb

**Con Google Colab Pro**

Se tiene acceso a las siguientes características.

* Memoria RAM 80 Gb
* GPU A100, V100 y V4
* TPU

**Evaluación de riesgos y contingencias**

Entre los riesgos que se pueden anticipar, existen dos que resultan especialmente preocupantes. A continuación, se detallan:

**Riesgo A**: Gastar la totalidad del dinero disponible en AWS.

**Mitigación** **(A)**: Con un presupuesto de 200 mil pesos, se gastarían entre 10 y 20 dólares, equivalentes a 50 mil y 100 mil pesos, siendo este último el límite. En tal caso, si la solución no se presenta una vez efectuado el gasto, me limitaré a recomendar AWS y su infraestructura como una mejora para el proyecto.

**Riesgo B:** Los requisitos computacionales superan las capacidades disponibles para el proyecto.

**Mitigación B:** Ante esta situación, se dispone de la alternativa de Google Colab Pro, que para el año 2023 tiene un costo mensual de 43,164.87 pesos. De no ser viable esta opción, se consideraría la alternativa de reducir la cantidad de datos, con el objetivo de hacer el proyecto más manejable en términos computacionales.

1. **Entendimiento del problema, pregunta de negocio o hipótesis**

El objetivo consiste en clasificar los documentos 10-K de acuerdo con dos categorías: "Mal sentimiento" y "Buen sentimiento". Se analizan los 10-K desde enero uno del 2014 hasta junio 6 de 2023. Para brindar más detalle acerca de este objetivo, es esencial llevar a cabo un repaso histórico de un periodo puntual, brindando la posibilidad de respondes a la pregunta implícita de cómo y por qué ese objetivo.

El 2022 fue un periodo álgido, al igual que 2020 o 2014 pero, me intereso más el 2022 no solo por la acumulación de lo visto en 2020 (Pandemia COVID-19) sino también por los propios hechos sociales, políticos y económicos durante ese periodo.

**Recuento histórico del 2022**

A lo largo del 2022 se produjeron eventos económicos, sociales y políticos de gran relevancia. A continuación, se detallan algunos de estos sucesos:

* **4 de enero**, la empresa inmobiliaria china Evergrande suspendió la cotización de sus acciones en la bolsa, generando un riesgo inminente de crisis financiera mundial debido al efecto dominó.[1]
* **22 de enero**, Ucrania recibió armamento por parte de los Estados Unidos.[2]
* **23 de enero**, los Estados Unidos evacuaron a los diplomáticos no esenciales de su embajada en Ucrania.[3]
* **24 de ener**o se produjo un golpe de estado en Burkina Faso.[4]
* **4 de febrero** comenzaron los Juegos Olímpicos de invierno en Pekín, China, en medio de tensiones bélicas en Europa. [5]
* **7 de febrero**, el presidente de Francia, Emmanuel Macron, se reunió con el presidente ruso Vladimir Putin en Moscú con el objetivo de buscar una solución a la crisis en Ucrania. [6]
* **10 de febrero**, Rusia y Bielorrusia realizaron ejercicios militares cerca de la frontera ucraniana. [7]
* **3 de febrero** se llevaron a cabo elecciones presidenciales en Alemania para el periodo 2022-2027, en las cuales el presidente en funciones, Frank-Walter Steinmeier, resultó reelegido. [8]
* **14 de febrero**, las bolsas mundiales se desplomaron por temor a una inminente invasión de Rusia a Ucrania. [9]
* **15 de febrero**, medios de comunicación ucranianos sufrieron un ciberataque masivo. [10]
* **16 de febrero**, Rusia garantizó el retiro de sus tropas de la zona del Donbás, sin embargo, los recursos de inteligencia de la OTAN alertaron que la declaración era una distracción y que, por el contrario, las tropas estaban en aumento. [11]
* **21 de febrero**, Rusia reconoció la independencia de la República de Donetsk y la República de Lugansk, y comenzó a militarizar la región del Donbás, acciones que fueron repudiadas en las Naciones Unidas. [12]

**22 de febrero**, Alemania suspendió el flujo del gas ruso Nord Stream. Al mismo tiempo, los Estados Unidos y el Reino Unido impusieron sanciones a individuos rusos y al propio Kremlin. [13]

* **23 de febrero**, tras las sanciones del Reino Unido y Estados Unidos, la Unión Europea se sumó y aplicó sanciones económicas a Rusia. [14]
* **24 de febrero**, Rusia invadió Ucrania. [15]

Ese mismo día, se produjeron enormes caídas en Wall Street. Las bolsas internacionales colapsaron, la bolsa rusa se desplomó y el precio de las materias primas como la soja, el trigo, el petróleo y el maíz subió considerablemente. [16]

También, las tropas rusas tomaron control de la Central nuclear de Chernóbil. [17]

* **26 de febrero**, la Unión Europea, Estados Unidos y sus aliados acordaron remover a los bancos rusos del sistema financiero SWIFT e imponer restricciones al Banco de Rusia, provocando una devaluación dramática en el rublo ruso. [18]
* **8 de marzo**, empresas multinacionales como McDonald's, Adidas, Coca-Cola, Shell y Starbucks suspendieron sus operaciones en Rusia. [19]
* **16 de marzo**, se reportó un aumento de agresiones de soldados rusos contra ciudadanos ucranianos. El Teatro de Mariúpol, que se utilizaba como refugio civil, fue destruido por varios misiles rusos, matando a aproximadamente 300 refugiados. [20]
* **8 de abril**, los precios de los alimentos aumentaron al nivel más alto desde que la ONU comenzó a medirlo, con productos básicos como el trigo incrementándose un 20% a raíz de la guerra. [21]
* **4 de mayo**, la Reserva Federal de Estados Unidos realizó el mayor incremento de tipos de interés desde el año 2000, con el objetivo de combatir la tasa de inflación. [22]
* **15 de junio**, la Reserva Federal de Estados Unidos aumentó las tasas de interés en 75 puntos básicos, equivalentes al 0.75%, el mayor aumento desde 1994. [23]
* **7 de julio**, el primer ministro británico Boris Johnson renunció a su cargo como líder del Partido Conservador. [24]
* **8 de julio**, el ex primer ministro japonés Shinzo Abe fue asesinado de un disparo en la espalda mientras daba un discurso electoral en la ciudad de Nara. [25]
* **13 de julio**, el euro y el dólar alcanzaron un valor de paridad histórico por primera vez en 20 años. Esto se ocasiono por la incertidumbre de la continuidad del suministro de gas ruso y la decisión del Banco Central Europeo de mantener la tasa de interés, dejando a Europa ante una posible recesión económica. [26]
* **4 de agosto**, China llevó a cabo su ejercicio militar más grande en la historia en torno a Taiwán, en respuesta a la visita de la congresista Nancy Pelosi. [27]
* **17 de agosto**, la tasa de inflación del Reino Unido llegó a un nuevo máximo del 10.1%, siendo el costo de los alimentos el mayor contribuyente al incremento. [28]
* **8 de septiembre**, la Reina Isabel II de Reino Unido falleció a los 96 años, tras 70 años de reinado. Su hijo Carlos se convirtió en el monarca. [29]
* **20 de octubre**, la nueva primera ministra Liz Truss renunció a su cargo, luego de 45 días al frente del gobierno. [30]
* **27 de octubre**, Elon Musk concretó la compra de Twitter, Inc. por 44 mil millones de dólares. [31]
* **15 de noviembre**, la humanidad alcanzó los 8 mil millones de habitantes, según la ONU. [32]
* **20 de noviembre**, se inauguró la edición XXII del Mundial de Fútbol en Catar. [33]
* **18 de diciembre**, el Mundial de Catar 2022 finalizó con Argentina como campeona. [34]
* **29 de diciembre**, falleció el jugador de fútbol Pelé a los 82 años. [35]

(Wikipedia, s.f.) . Eventos del año 2022. Recuperado el [6/11/2023] de [2022 - Wikipedia, la enciclopedia libre](https://es.wikipedia.org/wiki/2022) [2022 ]

Cada acontecimiento social, político y económico influye directamente en la toma de decisiones de inversión. De la misma forma que los inversores están supeditados por los hechos mencionados, las empresas también lo están. Factores como el gasto, las ventas o los ingresos pueden cambiar, afectando el estado de ánimo o sentimiento de estas empresas. Estas últimas, a menudo, sintetizan de manera efectiva todos los hechos relevantes durante un periodo específico. Por ello, el objetivo es clasificar los documentos 10-K presentados a la Comisión de Bolsa y Valores de Estados Unidos, más conocida por sus siglas en inglés, SEC. Para ello los sentimientos se clasificarán en dos categorías: "Buen sentimiento" y "Mal sentimiento".

Se establecen las hipótesis de correlación

**Hipótesis Nula (H0):**

* No hay correlación entre el sentimiento positivo y el precio del S&P 500.
* No hay correlación entre el sentimiento negativo y el precio del S&P 500.

**Hipótesis Alternativa (Ha):**

* Existe una correlación entre el sentimiento positivo y el precio del S&P 500.
* Existe una correlación entre el sentimiento negativo y el precio del S&P 500.

Los inversores toman decisiones acordes a los periodos de mercado, que se caracterizan por alzas (Bulls, Buen sentimiento) y bajas (Bears, Mal sentimiento). Dependiendo de cómo es el mercado bien sea Bulls o Bear la estrategia una inversión cambia, si cambia el contexto o entorno, como inversor debo de adaptarme a la coyuntura.

Mercados Alcistas (Bulls)

Los mercados alcistas, o mercados "**Bulls**", están asociados con expansiones económicas y la confianza del inversionista. En nuestro contexto, esto último se traduce como "buen sentimiento". Por definición, los mercados alcistas poseen las siguientes características:

* **Aumento de la confianza en los inversores**: Los inversores, convencidos de que los precios continuarán aumentando, depositan su confianza en el mercado a través de compras, lo que en sí mismo contribuye a un aumento de los precios. (John Rambow, 2023)
* **Las empresas invierten en su futuro:** Impulsadas por el incremento del consumo, las empresas invierten en sí mismas y se enfocan en su expansión. (John Rambow, 2023). Esto lleva a:
* **Disminución de las tasas de desempleo:** A medida que las empresas se expanden, contratan a más empleados, generando bajas tasas de desempleo. Al mismo tiempo, los salarios medios aumentan, (John Rambow, 2023) permitiendo que:
* **El dinero sea más fácil de gastar**: El aumento de los salarios da lugar a la percepción de que es más fácil gastar dinero, debido a la facilidad para obtener más ingresos. (John Rambow, 2023)

Mercado Bajista (Bear)

Los inversores se vuelven pesimista en un mercado bajista en nuestro contexto se definiría como mal sentimiento. En sí, un mercado bajista (Bear) no necesariamente es malo, incluso puede ser mejor que un periodo alcista. Es en este mercado que los precios de compra están bajos presentando una oportunidad de entrada para una inversión, aunque con un riesgo mayor de no obtener retornos

* Este periodo suele ir acompañado de un crecimiento económico bajo (con un PIB de crecimiento reducido), una disminución en los precios de las acciones, una liquidez reducida y una alta tasa de desempleo. (Campbell, 2023)

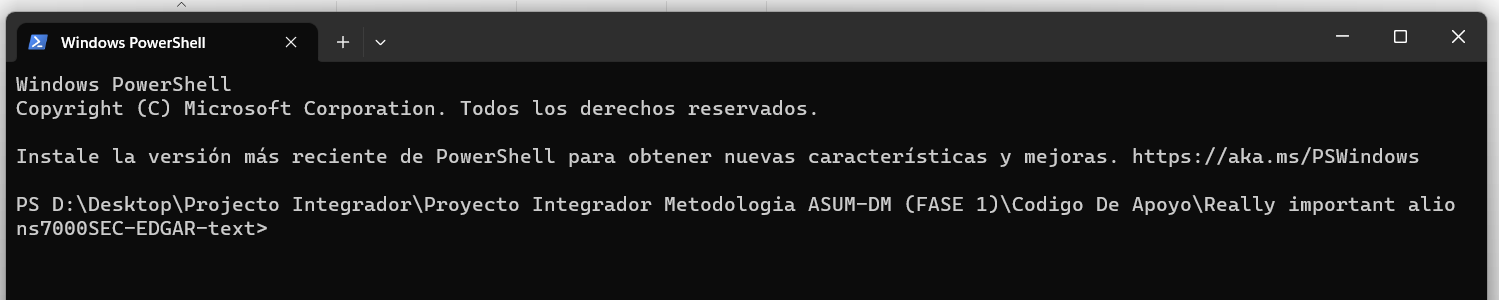
1. **Análisis Exploratorio de Datos**

**Fase 1:** Recolección de datos y **Fase 2:** Entendimiento de datos – Análisis de datos

Para la recopilación de datos, se utilizó el código del usuario "alions7000" en GitHub. Este código, propuesto por el usuario, incluye elementos que resultan muy beneficiosos para el proyecto.

Es conocido que cuando se trabaja con datos no estructurados (10-K) con el objetivo de clasificación, existen diferencias en cómo se deben procesar estos datos. En algunos casos, se eliminan números, tablas, etiquetas HTML, etc. Podrías preguntarte por qué estamos hablando del procesamiento de datos en la fase de recolección. Esto se debe a que el código empleado para la recolección de datos permite evitar varios procesos normalmente incluidos en el procesamiento. Procesos como la eliminación de etiquetas HTML o números se llevaron a cabo durante la descarga de la base de datos.

Para extraer los datos, se inicializa con el Windows PowerShell

Posteriormente: se agrega el siguiente comando:

//python SEC-EDGAR-text --storage=/ruta/a/mi\_ubicación\_almacenamiento --start=20150101 --end=99991231 --filings=10-K --multiprocessing\_cores=0 --traffic\_limit\_pause\_ms=500//

* /ruta/a/mi\_ubicación\_almacenamiento: representa la ubicación donde se guardarán los archivos descargados.

* --start=20150101 y --end=99991231: se utilizan para definir las fechas de inicio y fin del periodo de interés. En nuestro caso, hemos especificado desde el 1 de enero de 2015 hasta una fecha en el futuro distante para obtener todos los archivos disponibles.
* --filings=10-K: nos permite seleccionar el tipo de archivos que nos interesa. En este caso, son los informes 10-K. Sin embargo, también se podrían seleccionar otros tipos como 8-K y 10-Q.
* --multiprocessing\_cores=0: permite definir la cantidad de núcleos de la máquina que se van a utilizar para el proceso de descarga. Este valor puede ser modificado de acuerdo a las características de cada máquina.
* --traffic\_limit\_pause\_ms=500: se utiliza para evitar saturar el servidor de la SEC con demasiadas solicitudes en un corto periodo de tiempo. En este caso, se ha dejado en su valor predeterminado.

Durante el proceso de descarga solo se extrae, los siguientes elementos:

**Item 1,** Business.

**Item 1A,** Risk factors,

**Item 7,** Management’s Discussion and Analysis of Financial Condition and Results of Operations.

**Item 7A,** Quantitative and Qualitative Disclosures About Market Risk

Las razones del porque solo se extraen están soportadas por el trabajo o tesis de Sander Blomme y Julie Dedeyne “Predicting the effect of 10-K, 10-Q and 8-K company reports on abnormal stock returns using FinBERT NLP methods” (Dedeyne, 2019-2020).

Dentro de los distintos temas que puede abarcar un documento financiero de la característica de un 10-K. estos temas resultan ser los de mayor interés para el análisis, algunos abarcan las áreas sensibles a los que una empresa está expuesta, riesgos cambiarios o de precio; otros señalan los litigios, las preocupaciones u oportunidades se plasman bien dentro de estas áreas.

**Item 1,** Business.

En este rubro se describen las actividades de la empresa, sus productos, servicios, filiales y el mercado su mercado. También se comentan acerca de los acontecimientos especiales, la competencia, los costos operativos especiales y los posibles problemas laborales. (Dedeyne, 2019-2020)

**Item 1A,** Risk factors

Como su nombre lo indica se analizan los riesgos globales, factores de exposición a los que se enfrenta la empresa, tales como: riesgos del sector, riesgos geográficos. A qui so lo que se da una descripción de lo que enfrenta mas no como los afronta. (Dedeyne, 2019-2020)

**Item 7,** Management’s Discussion and Analysis of Financial Condition and Results of Operations.

La gerencia o dirección describe como está el balance de los últimos ejercicios, se tratan los temas financieros como la liquidez, recursos de capital, cambios en los precios. También entran en juego como fueron los acuerdos extra contables y obligaciones, estimaciones de hipótesis sobre lo que la compañía hizo y finalmente se comenta acerca de los riesgo, tendencias e incertidumbre que enfrenta la empresa. (Dedeyne, 2019-2020)

**Item 7A,** Quantitative and Qualitative Disclosures About Market Risk

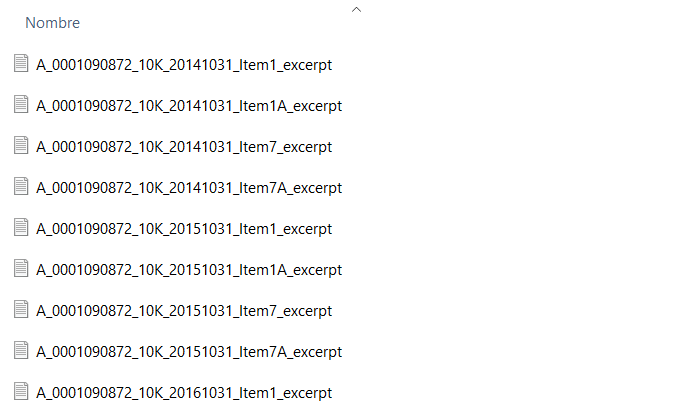
Se encuentra la exposición de la empresa a algunos riesgos como el tipo de interés, el riesgo de cambio, el riesgo del precio de las acciones y finalmente se da una descripción de como se gestiona estos riesgos. (Dedeyne, 2019-2020)

Con lo anterior, estamos acortando pasos que deberíamos hacer en el procesamiento de datos, como la extracción de número, retirar la Etiquetas, remover tablas, filtra palabras, etc. Para finalmente quedar con los elementos que definen el sentimiento de un documento.

A close-up of a document

Description automatically generated with low confidence

**Figura 1, tomada de**: (United States Securities And Exchange Commission)

Básicamente desde la extracción pasamos de tener documentos de 80, 300 e incluso 800 páginas a solo documentos en formato texto de unas cuantas páginas.

**Finalmente, de esta manera se tiene una base de datos de 19,057**

Una vez los datos se han descargado se reciben en formato TXT.

**Conclusiones**

**Fuentes de datos y naturaleza:** Los datos provienen de la API de la SEC, esto con ayuda del código del usuario “alions7000”, son datos semiestructurados y transaccionales dada su característica de informe financiero.

**Ingesta de datos:** Como son informes anuales su ingesta es Bach. Se descargan y guardan para su posterior almacenamiento.

**Almacenamiento:** Data-Lake en S3 de AWS, para almacenar grandes volúmenes (con la API se puede descargar hasta un millón de documentos). En cuyo caso S3 de AWS, seria ideal para ese propósito.

**Ambiente de procesamiento:** Dado el volumen que implica procesar una base de datos de 19 mil se uso toda la capacidad de Google Colab pro, pero un entorno de procesamiento avanzado como apache Spark seria ideal sobre todo si se busca el millón de documentos.

1. **Selección de modelos (supervisados y no supervisados), Ingeniería de Características, Entrenamiento, Evaluación**

**Modelo FinBERT para Clasificación de Texto**

FinBERT, un modelo de aprendizaje mediante transferencia puede ser pre-entrenado en un corpus sin necesidad de datos etiquetados. Exploraremos a fondo cómo funciona y por qué es el modelo ideal para nuestro propósito (Araci, 2019).

El análisis de sentimientos se basa en extraer emociones u opiniones del lenguaje escrito. A este fin, existen dos grandes grupos de métodos: los métodos de aprendizaje automático, caracterizados por el recuento de palabras, y los métodos de aprendizaje profundo, cuya característica principal es representar el texto mediante incrustaciones. Los métodos de aprendizaje automático tienen dificultades para representar información semántica y, según Dogu Tan Araci, los métodos de aprendizaje profundo se consideran excesivamente "hambrientos de datos". (Araci, 2019).

El análisis de textos financieros difiere del análisis de textos generales, ya que lo esencial es captar el contexto de la expresión. Loughran y McDonald (2011), citados en (Araci, 2019) crearon un diccionario de términos financieros con valores positivos e inciertos, entre otros, que mide el tono del documento mediante el conteo de palabras. Otro ejemplo, de Pagolu et al. (2016), citado en (Araci, 2019) introdujo n-grams de tuits en algoritmos de aprendizaje automático para detectar el sentimiento de la entidad financiera mencionada.

**Para Entender FinBERT hay que entender BERT**

* Método Transformers de BERT (Bi-directional Encoder Representations of Transformers)

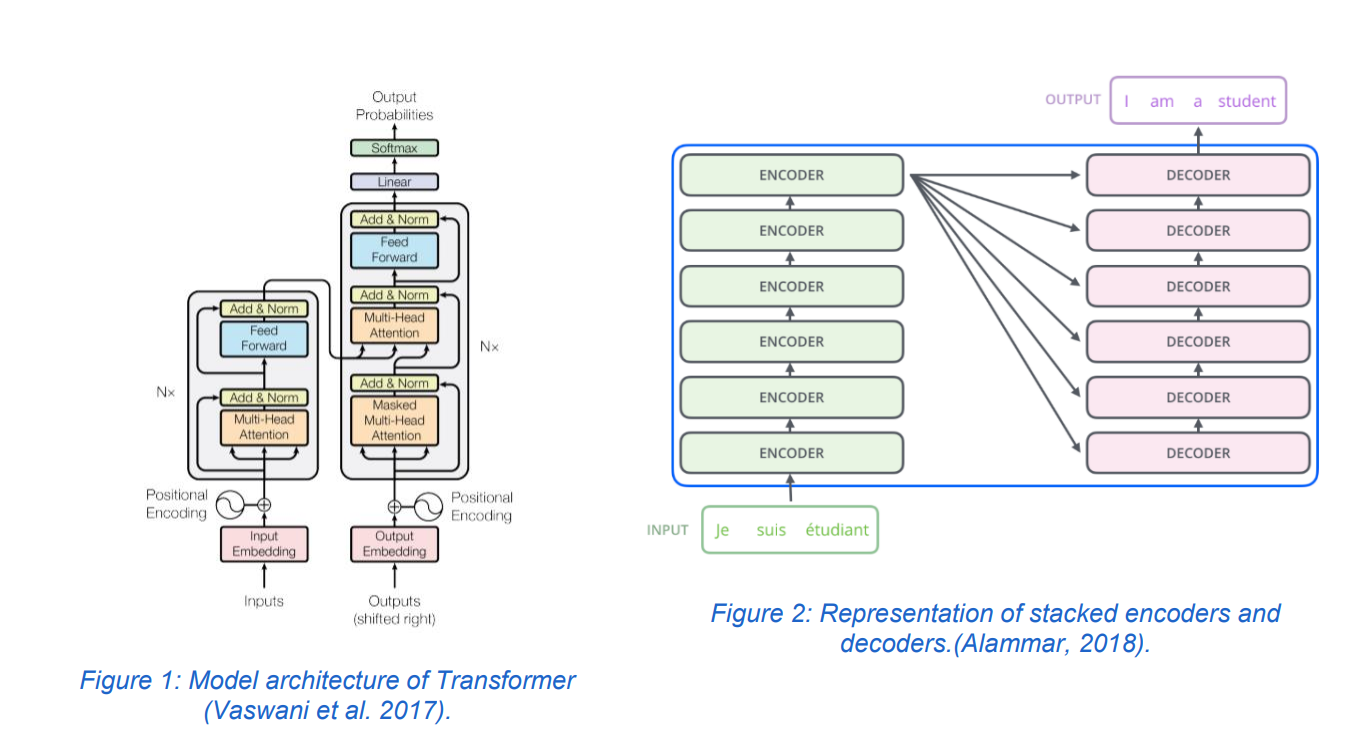
El modelo de Transformer se sustenta en una arquitectura de codificador-decodificador, introducida por Vaswani et al. (2017), tal y como se cita en Dedeyne (2019-2020). La esencia de este modelo radica en aprender a partir de un vector de contexto que suministra información global acerca de todas las entradas. Este vector de contexto también destaca la información más relevante, a la que se debe prestar mayor atención. (Dedeyne, 2019-2020)

The Transformer is an attention-based architecture for modeling sequential information, that is an alternative to recurrent neural networks [29]. It was proposed as a sequence-tosequence model, therefore including encoder and decoder mechanisms. Here, we will focus only on the encoder part (though decoder is quite similar) (Araci, 2019) [El Transformer es una arquitectura basada en la atención para modelar información secuencial, que es una alternativa a las redes neuronales recurrentes [29]. Se propuso como un modelo secuencia-secuencia, por lo que incluye mecanismos codificadores y decodificadores. Aquí nos centraremos sólo en la parte del codificador (aunque el decodificador bastante similar)] traducción hecha en: (Deepl, 2023)

El codificador del Transformer se compone de varias capas idénticas. Cada una de estas capas cuenta con una capa de autoatención de múltiples cabezas y una red completamente conectada. La capa de autoatención aprende tres mapeos de incrustaciones, definidos como clave, consulta y valor. (Araci, 2019)

Se calcula una puntuación de similitud para cada token utilizando sus respectivas claves y consultas. Estas puntuaciones se emplean posteriormente para ponderar los vectores de valor y obtener así la nueva representación del token. (Araci, 2019)

Con la autoatención de múltiples cabezas, estas capas se concatenan, lo que permite evaluar la secuencia desde diferentes "perspectivas". A continuación, los vectores resultantes pasan por redes completamente conectadas con parámetros compartidos (Araci, 2019)

En las Figuras 1 y 2 se presenta la arquitectura del modelo Transformer, incluyendo el mecanismo de codificador-decodificador. En la imagen de la izquierda, se ilustra la arquitectura multicabeza del mecanismo de codificador-decodificador, lo que implica que los componentes están apilados completamente.

**Figura 2, Tomado de:** (Dedeyne, 2019-2020)

* BERT se acomoda de manera flexible para ejecutar diversas tareas del Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés) (Dedeyne, 2019-2020)

El mecanismo de autoatención de BERT, una función heredada del modelo Transformer, facilita la adaptación del modelo para diversas tareas simplemente modificando las entradas apropiadas. Para cada tarea específica, las entradas y salidas se introducen en BERT y todos los parámetros se recalibran conjuntamente, dando lugar a una afinación de extremo a extremo (Dedeyne, 2019-2020)

Durante el preentrenamiento, BERT recibe pares de oraciones (A y B) y aprende a predecir si la oración B es una continuación lógica de la oración A. Esto se conoce como la tarea de "predicción de la siguiente oración" y es uno de los dos objetivos de preentrenamiento de BERT, siendo el otro la "enmascaramiento de lenguaje”. (Dedeyne, 2019-2020)

Una vez pre-entrenado, BERT se puede ajustar para una variedad de tareas de procesamiento de lenguaje natural (NLP) simplemente cambiando la forma en que se presentan los datos de entrada. (Dedeyne, 2019-2020)

A picture containing text, screenshot, diagram, parallel

Description automatically generatedA nivel de salida, los tokens se ingresan en una capa de salida para tareas a nivel de token, mientras que la representación CLS se dirige hacia la capa de salida para la clasificación. El token CLS es un token especial de clasificación, y el último estado oculto asociado a este token se utiliza como una representación agregada de la secuencia para las tareas de clasificación (Dedeyne, 2019-2020)

**Figura 3, Tomada de**: (Dedeyne, 2019-2020)

**Figura 3**: En la **subfigura (a)** se presenta una tarea de clasificación de pares de oraciones. Aquí se toman dos oraciones como entrada, las cuales están separadas por el token SEP. **E** representa las incrustaciones de entrada, mientras que **T**, la capa verde, es la última capa oculta del BERT preentrenado y representa la representación contextual. Para las tareas de clasificación, se da la representación CLS del BERT preentrenado a la capa de salida. (Dedeyne, 2019-2020)

Por otro lado, en la **subfigura (c)** se muestra el problema de la respuesta a preguntas. En este caso, se introduce una pregunta y un párrafo separados por el token de separación. La representación del token se introduce en una capa de salida para tareas a nivel de token. (Dedeyne, 2019-2020)

**Entrenamiento de FinBERT**

Entre los grandes desafíos al realizar clasificaciones de sentimientos en textos financieros se encuentra la falta de datos etiquetados. Esto se traduce en la dificultad de entrenar una red neuronal profunda especializada en análisis de sentimientos. El lenguaje financiero es muy específico, cuenta con un vocabulario preciso, lo que hace que los modelos de clasificación general resulten ineficientes para captar el contexto. Por otro lado, el uso de léxicos especializados, como el diccionario de sentimiento financiero de Loughran-McDonald (2011) citado en (Dedeyne, 2019-2020) se limita a un recuento de palabras, descuidando en consecuencia la semántica del texto. Como se señaló al principio, FinBERT es un modelo de lenguaje preentrenado. (Dedeyne, 2019-2020)

Araci (2019), citado por (Dedeyne, 2019-2020), propone el uso de métodos de aprendizaje por transferencia en NLP. Esta estrategia utiliza modelos de lenguaje ya entrenados sobre corpus sin etiquetas. Estos modelos, a su vez, pueden ser pre-entrenados adicionalmente en corpus que pertenecen a un dominio específico, como es el caso del ámbito financiero, antes de ajustarlos para tareas particulares. De esta forma, se creó el modelo FinBERT, que se pre-entrenó en un corpus financiero para luego afinarse para tareas específicas de clasificación financiera.

En relación con el preentrenamiento adicional del modelo de lenguaje BERT, (Araci, 2019) citado en (Dedeyne, 2019-2020). Propone dos enfoques:

* Pre-entrenar BERT en un corpus financiero de gran tamaño.
* Pre-entrenar BERT únicamente con las oraciones del conjunto de datos de entrenamiento para la clasificación.

Además, para evitar el problema conocido como "olvido catastrófico" durante el afinamiento, (Araci, 2019)mencionado también en (Dedeyne, 2019-2020) aplica tres métodos:

* **Tasa de aprendizaje** con forma triangular inclinada: este método implica que la tasa de aprendizaje incrementa linealmente hasta llegar a un punto específico, después de lo cual comienza a disminuir también de manera lineal. esta estrategia permite acercarse rápidamente a un conjunto de parámetros óptimos. A medida que avanza el entrenamiento, estos parámetros son progresivamente refinados para una optimización adicional. (Dedeyne, 2019-2020)
* **Descongelamiento gradual**: en este caso, el entrenamiento inicia con todas las capas en estado congelado, exceptuando la capa de clasificación. A medida que avanza el entrenamiento, las capas se van descongelando de manera gradual, comenzando por

las superiores. De esta manera, las características de nivel inferior, que contienen información de lenguaje más general, son las que menos se ajustan. (Dedeyne, 2019-2020)

* A picture containing text, pattern, diagram, line

  Description automatically generated**Ajuste fino discriminativo**: este método permite ajustar cada capa con una tasa de aprendizaje distinta, asignando tasas más bajas a las capas inferiores. La premisa subyacente a esta técnica es que las capas inferiores contienen información profunda del lenguaje, mientras que las capas superiores incluyen información relevante para la tarea de clasificación actual. De esta forma, dado que cada capa capta información diferente, cada una se ajusta de una manera diferente. (Dedeyne, 2019-2020)

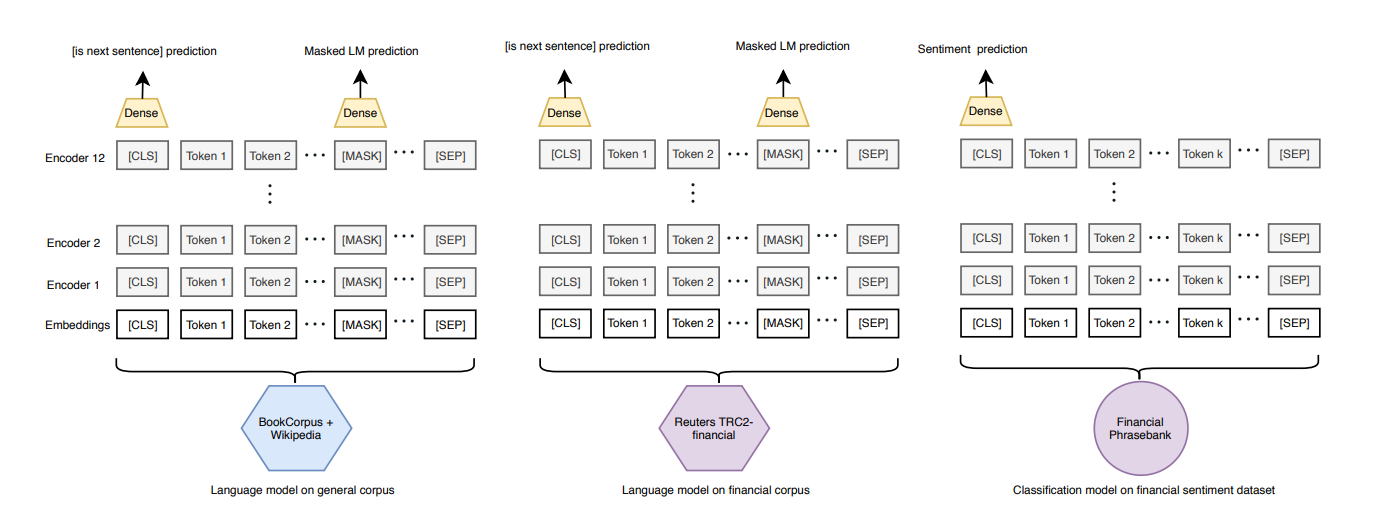
**Figura 4:** ajuste discriminativo véase las figuras b y c**, tomado de**: (Dedeyne, 2019-2020)

**Datasets**

El modelo BERT se pre-entrena adicionalmente con un corpus financiero denominado TRC2-financial. Este corpus es una selección de TRC2 de Reuters, un conjunto compuesto por 1.8 millones de artículos noticiosos publicados entre 2008 y 2010. Se seleccionan términos clave financieros para que el corpus sea más relevante y manejable en función de la capacidad de procesamiento disponible. El TRC2-financial resultante contiene 46,143 (Araci, 2019)documentos, sumando más de 29 millones de palabras y aproximadamente 400,000 oraciones.

Del conjunto de datos Financial PhraseBank. Este consiste en un análisis de sentimientos de 4,845 frases en inglés seleccionadas al azar de noticias financieras extraídas de la base de datos LexisNexis. Las frases fueron etiquetadas por 16 expertos en finanzas y negocios, según su interpretación de cómo la información en cada frase podría influir en el valor de las acciones de la empresa mencionada. Además, el conjunto de datos proporciona información

sobre el grado de concordancia entre los anotadores. Los investigadores destinaron el 20% de todas las frases para el conjunto de prueba y el 20% del resto para el conjunto de validación, dejando 3101 ejemplos para el conjunto de entrenamiento. También aplicaron la técnica de validación cruzada de 10 iteraciones en algunos de sus experimentos. (Araci, 2019)



**Figura 5, tomada de**: (Araci, 2019)

**Dataset** de DeSola, Hanna, & Nonis (2019), como se citó en (Dedeyne, 2019-2020)

Para preentrenar el modelo de lenguaje BERT en el dominio financiero, se emplearon dos enfoques.

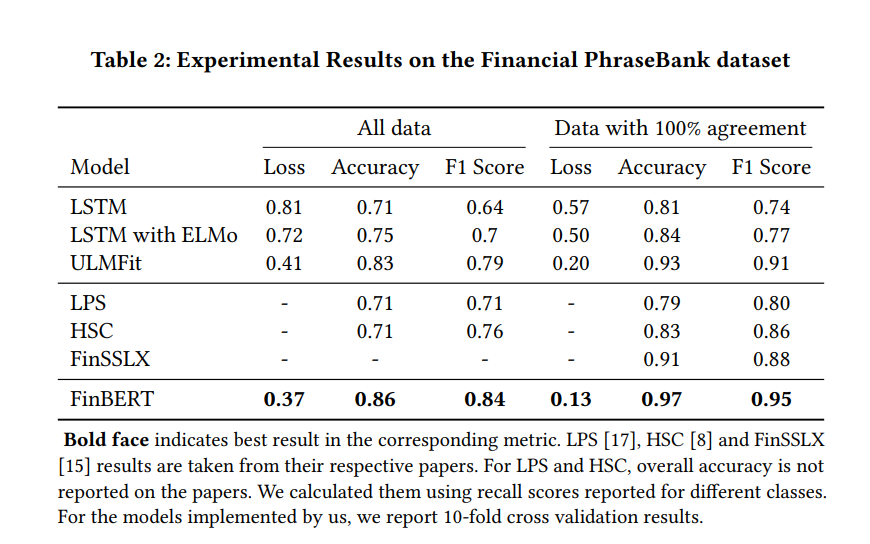
En el primero, se entrenó un modelo FinBERT desde cero usando la arquitectura original de BERT. Este modelo, aunque capta el contexto específico de los documentos financieros, no considera la información del modelo BERT previamente entrenado con artículos de Wikipedia y corpus de libros. (Dedeyne, 2019-2020)

Se entrenaron dos variantes del modelo BERT, inspiradas en el modelo original, en dos conjuntos de datos distintos.

* **El primer modelo** fue entrenado en un corpus denominado Prime, compuesto por presentaciones SEC 10-K del periodo 2017-2019. El entrenamiento fue realizado durante 500,000 pasos, con una longitud máxima de secuencia de 128, un tamaño de lote de 96 y un período de calentamiento de 50,000 pasos. (Dedeyne, 2019-2020)
* **El segundo modelo**, por su parte, se entrenó en otro corpus llamado Pre2K, que incluye presentaciones SEC 10-K de los años 1998 a 1999. Este modelo fue entrenado durante 250,000 pasos, con una longitud máxima de secuencia de 128 tokens, tamaño de lote de 96 presentaciones y un período de calentamiento de 50,000 pasos. (Dedeyne, 2019-2020)

Los modelos generaron incrustaciones de palabras que mostraron cambios notables. Un ejemplo es la palabra "cloud", que en el modelo previo al año 2000 estaba asociada con "hydra", pero en el modelo más reciente se asocia con "software". Además, términos como "business" y "subscriptions" estaban presentes en el modelo de lenguaje de 2019, pero no en aquel entrenado con los archivos 10-K de los años 1998-1999. (Dedeyne, 2019-2020).

**Evaluación de los modelos de clasificación**

Para evaluar los modelos de clasificación, se recurre a tres métricas: la precisión, la pérdida de entropía cruzada y el promedio macro F1. La pérdida de entropía cruzada se ajusta con la raíz cuadrada de la tasa inversa de frecuencia. El promedio macro F1, por otro lado, calcula las puntuaciones F1 para todas las clases y luego toma su promedio. Dada la presencia de un desequilibrio de etiquetas en los datos de Financial PhraseBank, estas medidas ofrecen una evaluación eficaz del rendimiento de la clasificación. Para la evaluación del modelo de regresión, se utilizan el error cuadrado medio y R2 (Araci, 2019)

**Figura 6, tomada de:** (Araci, 2019)

En todas las técnicas de medidas FinBERT es el de mejor puntaje.

**Conclusión**

K-Mean

El algoritmo K-Means es una herramienta sencilla pero eficaz. Se pretende aplicarlo a los resultados obtenidos del modelo FinBERT, es decir, se usarán sobre las clasificaciones previamente realizadas. El objetivo es observar los patrones de agrupamiento resultantes. De esta forma, es posible identificar correlaciones interesantes. Específicamente, quizás las secciones de los documentos 10-K se agrupan en clústeres definidos por el tipo de sentimiento que les fue asignado durante la etapa de clasificación previa.

El valor óptimo para K, que representa los puntos centrales o centroides, se procede con un proceso iterativo. Cada uno de los puntos de datos se asigna al centroide más cercano. De esta forma, los puntos de datos cercanos a un centroide en particular conforman un clúster. (JavaTpoint, s.f.)



**Figura 7, tomada de**: (JavaTpoint, s.f.)

El funcionamiento del algoritmo K-Mean se explica en los siguientes pasos

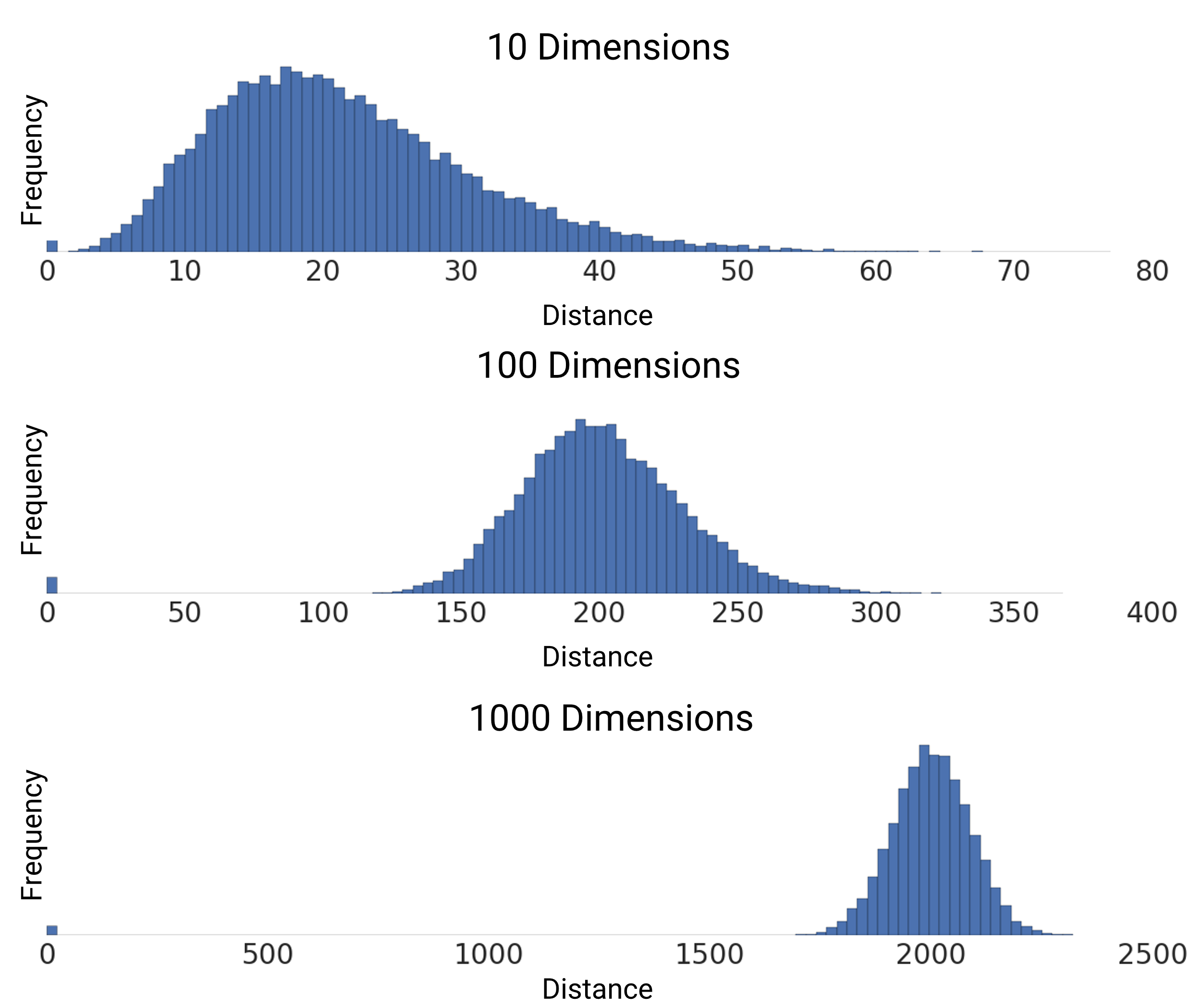
* **Paso 1**: Decidir la cantidad de clústeres, representada por el número K. (JavaTpoint, s.f.)
* **Paso 2**: Escoger aleatoriamente K puntos o centroides. (Estos pueden ser otros puntos dentro del conjunto de datos). (JavaTpoint, s.f.)
* **Paso 3**: Cada punto de datos se asigna al centroide más cercano, lo que resulta en la formación de los K clústeres previamente definidos. (JavaTpoint, s.f.)
* **Paso 4**: Se re calculan los centroides (JavaTpoint, s.f.)
* **Paso 5**: Repetir el tercer paso, es decir, reasignar cada punto de datos al nuevo centroide más cercano de cada clúster. (JavaTpoint, s.f.)
* **Paso 6**: Si se da alguna reasignación, retornar al paso 4, de lo contrario, proceder a finalizar. (JavaTpoint, s.f.)
* **Paso 7**: El modelo ya está listo. (JavaTpoint, s.f.)

El entrenamiento se detiene hasta alcanzar el número máximo de interacciones.

**Preocupaciones para tener en cuenta**

Un algoritmo como K-Mean es susceptible a la maldición de la alta dimensionalidad

El término "maldición de la alta dimensionalidad" alude a las dificultades que aparecen al manejar datos de alta dimensión, es decir, conjuntos de datos con una vasta cantidad de atributos o características. Un problema es que, en espacios de alta dimensión, medidas como la distancia euclidiana, comúnmente empleada en k-mean, deja de ser efectivas. (Machine Learning Google, 2022)

**Figura 8, tomada de:**  (Machine Learning Google, 2022)

En espacios de alta dimensión, los puntos tienden a parecer equidistantes entre sí, complicando la distinción entre los puntos considerados "cercanos" y "lejanos" Esto evidente en la **figura 8**: conforme se incrementa el número de dimensiones, la desviación estándar de las distancias se reduce, y la distancia promedio entre los ejemplos parece converger hacia un valor fijo. En consecuencia, en un espacio de alta dimensión, el algoritmo k-means puede enfrentar dificultades para segmentar los datos en grupos de manera eficaz. (Machine Learning Google, 2022)

**PCA**

La técnica de reducción de dimensionalidad conocida como Método de Análisis de Componentes Principales (PCA), se emplea para simplificar conjuntos de datos grandes. Su objetivo es reducir el número de variables manteniendo al mismo tiempo la mayoría de la información significativa. (Jaadi, 2023)

**PCA paso a paso**

**Estandarización**

La estandarización de variables continuas es un paso esencial, cuyo propósito es igualar el rango de las variables, permitiendo que todas ellas aporten de manera equitativa al análisis. (Jaadi, 2023)

La estandarización es particularmente crucial antes del Análisis de Componentes Principales (PCA) debido a la alta sensibilidad de este método a las varianzas de las variables iniciales. Es decir, si existe una gran disparidad en los rangos de las variables iniciales, las variables con rangos más amplios tendrán un impacto desmedido en comparación con las de rangos más reducidos. (Jaadi, 2023)

Matemáticamente la estandarización se expresa de la siguiente manera:

A black text on a white background

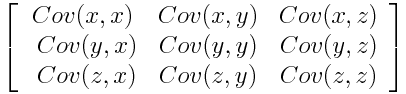
Description automatically generated with low confidence

**Figura 9, tomada de**: (Jaadi, 2023)

**Matriz de correlación**

A veces, existe una fuerte correlación entre las variables, lo que indica que portan información redundante. Para identificar estas correlaciones, se recurre a la matriz de covarianza. (Jaadi, 2023)

La covarianza de una variable es su varianza, Cov(a,a)=Var(a), se tiene que la diagonal principal de la matriz de covarianza; representa las varianzas de cada una de las variables iniciales, ya que la covarianza es una operación Cov(a,b)=Cov(b,a), los valores en la matriz de covarianza se consideran simétricos alrededor de la diagonal principal. Por lo tanto los elementos en la sección triangular superior son idénticos a los de la triangular inferior (Jaadi, 2023)



**Figura 10, tomada de:** (Jaadi, 2023)

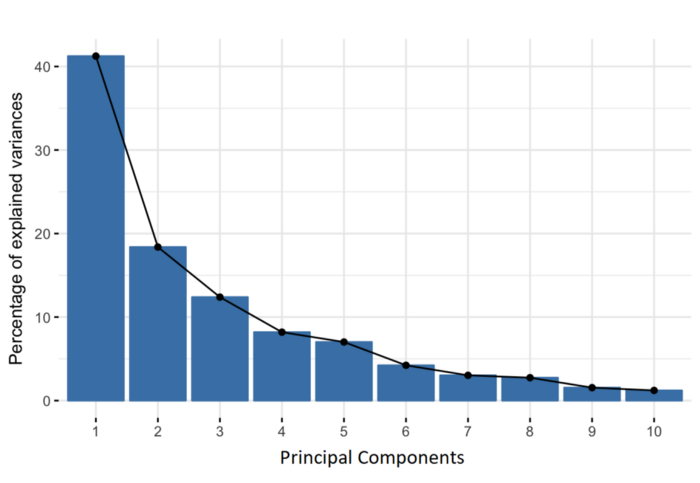
. (Jaadi, 2023)

Lo que realmente importa en la matriz de covarianza es el signo:

* **Si es positivo entonces:** las dos variables aumentan o disminuyen juntas (correlacionadas) (Jaadi, 2023)
* **Si es negativo entonces**: uno aumenta cuando el otro disminuye (inversamente correlacionado) (Jaadi, 2023)

**Se Calculan los valores y vectores propios de la matriz covarianza para identifica los componentes principales.**

Los componentes principales son variables recién formadas que se generan como combinaciones lineales de las variables iniciales. Estas mezclas se realizan de manera que las nuevas variables, es decir, los componentes principales, se mantienen no correlacionadas y la mayoría de la información contenida en las variables originales se concentra o condensa en los primeros componentes. En consecuencia, la estrategia es que, si disponemos de datos de 10 dimensiones que nos proporcionan 10 componentes principales, el análisis de componentes principales (PCA) intenta encapsular la mayor información posible en el primer componente, luego la máxima cantidad restante de información en el segundo, y así sucesivamente. Al final, se obtiene algo similar a lo representado en el diagrama (figura 10) mostrado a continuación. (Jaadi, 2023)



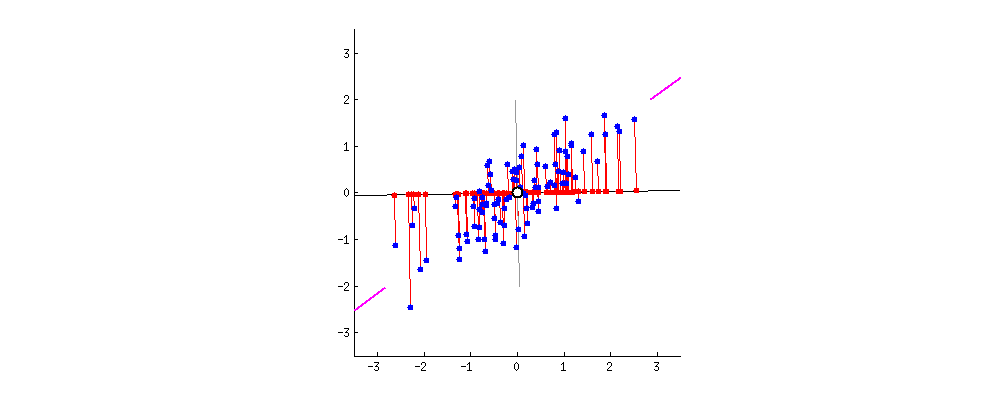
**Figura 11, tomada de**: (Jaadi, 2023)

Ordenar la información en componentes principales permite la reducción de la dimensionalidad manteniendo la máxima cantidad de información, al descartar componentes de menor importancia y emplear los restantes como nuevas variables. (Jaadi, 2023)

Es importante señalar que la interpretación de los componentes principales, al ser combinaciones lineales de las variables originales, es limitada y no tienen un significado concreto en sí mismos. (Jaadi, 2023)

Desde lo geométrica, los componentes principales se interpretan como las direcciones en las que se observa la mayor variación en los datos, es decir, las líneas que encapsulan la mayor cantidad de información intrínseca en el conjunto de datos. En este marco, la varianza se asocia directamente con la cantidad de información: una línea que presenta una alta varianza indica que los puntos de datos están más dispersos a lo largo de ella, lo que implica que esa línea contiene una gran cantidad de información. En resumen, los componentes principales pueden verse como nuevos ejes que proporcionan la mejor vista para examinar y analizar los datos, de manera que las diferencias entre las observaciones sean más evidentes. (Jaadi, 2023)

* **Valores y vectores propios de la matriz covarianza**

Se generan los componentes principales de manera tal que el primer componente principal se encargue de capturar la máxima variabilidad en el conjunto de datos. Ilustrativamente, en un gráfico de dispersión de los datos, esta línea principal sería aquella que, pasando por el origen, genera la mayor dispersión en las proyecciones de los puntos. Expresado en términos matemáticos, dicha línea optimiza la varianza, que se traduce en el promedio de las distancias al cuadrado de los puntos proyectados respecto al origen.

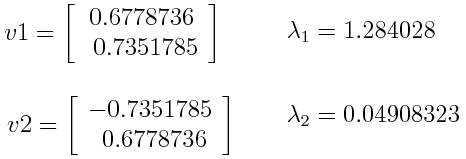
**Figura 12, tomada de**:  (Jaadi, 2023)

El cálculo del segundo componente principal se realiza de forma similar al primero, pero con la restricción de que debe ser ortogonal (es decir, no correlacionado) al primer componente y que maximice la siguiente varianza más alta. Esta metodología continúa hasta obtener un total de 'p' componentes, donde 'p' corresponde a la cantidad de variables iniciales. (Jaadi, 2023)

Es crucial entender que los vectores y los valores propios se generan en pares, y su cantidad total coincide con la dimensionalidad del conjunto de datos. Dentro de la matriz de covarianza, los vectores propios representan las direcciones (o los ejes) donde se encuentra la mayor variabilidad (o sea, los componentes principales), mientras que los valores propios indican la cantidad de variabilidad que cada componente puede explicar. (Jaadi, 2023)

Finalmente, al organizar los vectores propios en base a sus correspondientes valores propios, de mayor a menor, se obtienen los componentes principales ordenados según su relevancia. (Jaadi, 2023)

Imaginemos que poseemos un conjunto de datos de dos dimensiones con las variables 'x' e 'y'. Los vectores y los valores propios de la matriz de covarianza serían los siguientes:



**Figura 13, tomada de**: (Jaadi, 2023)

Si ordenamos los valores propios de mayor a menor, encontramos que λ1 es mayor que λ2. Esto nos indica que el vector propio correspondiente al primer componente principal (PC1) es v1 y que v2 se corresponde con el segundo componente principal (PC2). (Jaadi, 2023)

Una vez identificados los componentes principales, podemos calcular el porcentaje de varianza (o información) representada por cada componente. Para hacer esto, dividimos el valor propio de cada componente entre la suma de todos los valores propios. Aplicando esta fórmula a nuestro ejemplo, vemos que PC1 y PC2 representan respectivamente el 96% y el 4% de la varianza de los datos. (Jaadi, 2023)

**Vector De Característica**.

En la etapa anterior, identificamos los componentes principales ordenándolos por la importancia de sus valores propios correspondientes. Ahora, debemos decidir cuántos de estos componentes retenemos y cuántos descartamos, en particular los de menor relevancia (con valores propios más bajos). La matriz que se compone de los vectores propios que hemos decidido mantener se llama matriz de vectores característicos. (Jaadi, 2023)

Esta matriz de vectores característicos es nuestro primer paso hacia la reducción de la dimensionalidad. Si decidimos mantener solo p de n vectores propios (o componentes), la dimensionalidad de nuestro conjunto de datos final se reducirá a p. (Jaadi, 2023)

Siguiendo con el ejemplo del paso previo, tenemos la opción de formar un vector característico utilizando los dos vectores propios, v1 y v2: (Jaadi, 2023)

Análisis de componentes principales eigen vectores

**Figura 14, tomada de**: (Jaadi, 2023)

Alternativamente, podemos descartar v2, el vector propio de menor importancia, y formar un vector característico solamente con v1. (Jaadi, 2023)

Análisis de componentes principales eigen vectors 2

**Figura 15, tomada de**: (Jaadi, 2023)

Al eliminar el vector propio v2, se reduce la dimensionalidad en una dimensión, ocasionando una cierta pérdida de información en el conjunto de datos finales. No obstante, dado que v2 sólo representa el 4% de la información total, la pérdida es relativamente mínima, manteniendo aún el 96% de la información a través de v1. (Jaadi, 2023)

A partir de lo anterior se retienen todos los componentes o se eliminan los menos significativos depende de las necesidades específicas. Si el objetivo principal es describir los datos con nuevas variables no correlacionadas, también conocidas como componentes principales, sin necesidad de reducir la dimensionalidad, no sería necesario descartar los componentes menos significativos. (Jaadi, 2023)

**Recalibras los datos a lo largo de los ejes principales**

Para los pasos anteriores, salvo la estandarización, no se han realizado cambios en los datos. Aunque se han seleccionado los componentes principales y se ha creado el vector de características, los datos siguen estando alineados con los ejes originales, es decir, en función de las variables iniciales. (Jaadi, 2023)El objetivo de esta etapa final es emplear el vector de características, derivado de los vectores propios de la matriz de covarianza, para recalibrar los datos desde los ejes originales hacia los ejes representados por los componentes principales. Este cambio de orientación se realiza mediante la multiplicación de la transpuesta del conjunto de datos original por la transpuesta del vector de características, de ahí el término Análisis de Componentes Principales.

Vector de características de análisis de componentes principales

**Figura 16, tomada de**: (Jaadi, 2023)

**Coeficiente de Correlación De Pearson**.

Es la forma mas comuna de medir la correlación lineal. Es un número de -1 a 1 que mide la fuerza y dirección entre dos variables, se considera un estadístico descriptivo, específicamente describe la fuerza y dirección entre dos variables continuas. (Turney, 2022)

* Entre 0 y 1 es una correlación positiva
  + Mas cercana a cero es una correlación positiva débil
    - * Mas cercano a uno su correlación es muy fuerte
        + Cuando una variable cambia la otra también lo hace, en la misma dirección.

Entre -1 y 0 es una correlación negativa

Cuando una variable cambia la otra lo hace en la dirección opuesta.



**Figura 17, tomada de**: (Turney, 2022)

1. **Análisis y Conclusiones**

**Conclusiones**

**El modelo FinBERT** ha demostrado ser el más adecuado no sólo por su rendimiento superior en comparación con otros modelos, sino también por su capacidad para capturar la semántica a través de la arquitectura Transformers de BERT. Además, ha sido entrenado con corpus financieros y su preentrenamiento ha sido diseñado para evitar problemas como el olvido catastrófico. En consecuencia, se descartan los siguientes modelos:

* La bolsa de palabras (BoW) ignora la información de contexto esencial en el texto, se requiere entender el contexto financiero (Genc, 2020)
* Naive Bayes, se requiere de una enorme cantidad de datos para su entrenamiento más precisamente datos etiquetados, recurso del cual carezco, por último, no garantiza la semántica.

**Objetivo:** Se usara el modelo FinBERT para la clasificación de los sentimientos.

**Aunque algoritmos como K-mean** son simples en su diseño, han demostrado ser efectivos y prácticos en su implementación. Sin embargo, su eficacia puede verse comprometida en contextos de alta dimensionalidad. Aquí es donde el PCA puede ser beneficioso, al ser aplicado antes de K-mean.

**Objetivo:** Se utilizará PCA antes de K-mean con el fin de evitar problemas asociados con la alta dimensionalidad**.** Todo ello enmarcado dentro del objetivo propuesto para K-mean, mencionado inicialmente.

**La correlación de Pearson** se considera como estadístico inferencial, por lo que puedo ser usado para para probar hipótesis estadísticas, específicamente se puede corroborar si existe una relación significativa entre las variables.

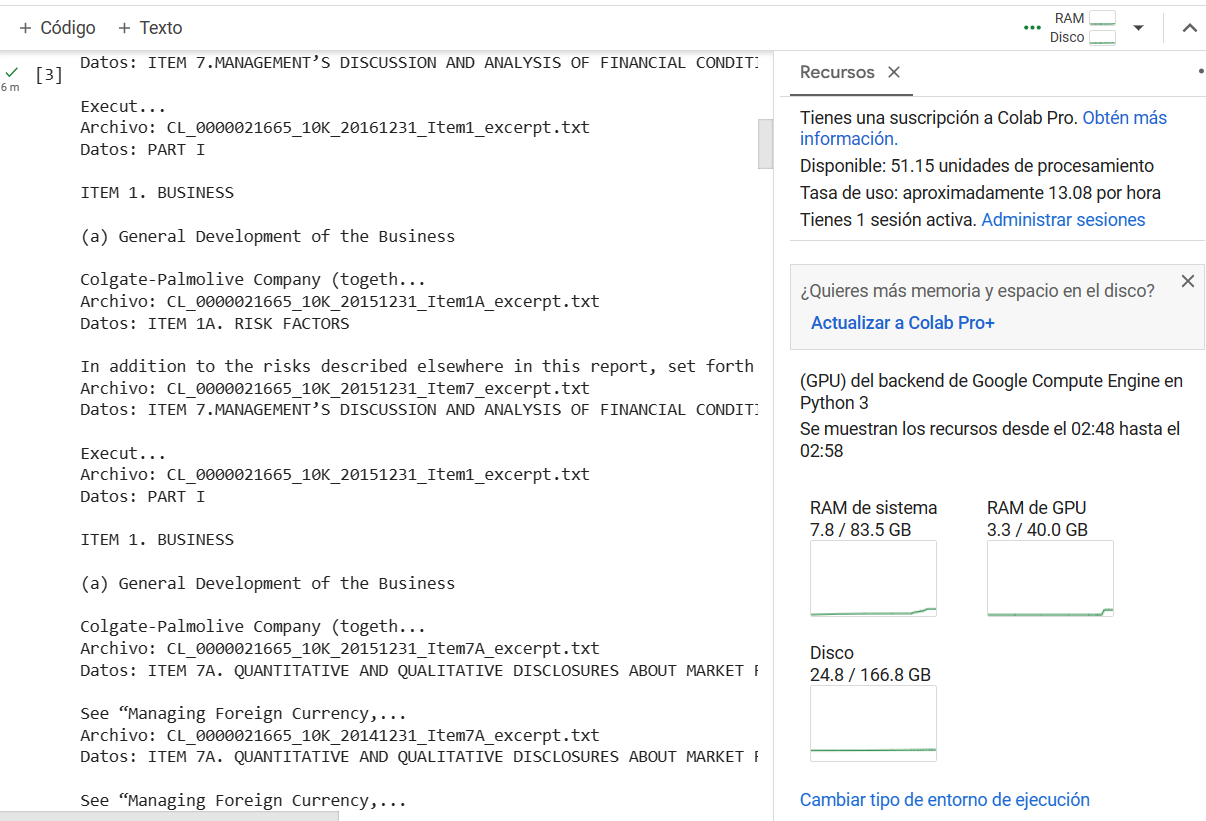
**Objetivo:** Se usará la correlación de Pearson para probar la hipótesis nula.

1. **Desarrollo del proyecto**

Se leen los datos

A picture containing text, screenshot, font, document

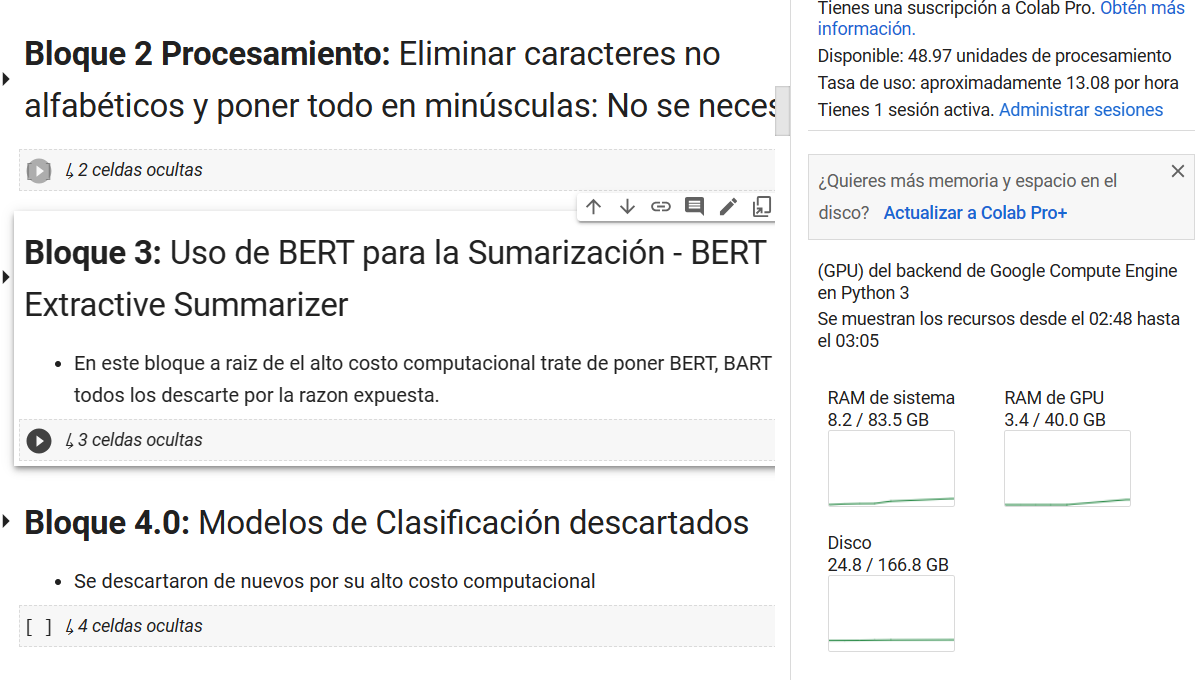
Description automatically generated



Como se está usando un modelo pre-entrenado, como FinBERT, no se requiere de las etiquetas. Además, este posee su propio tokenizador, no se necesita la remoción de Stop-Words, ni convertir palabras en minúsculas. Además, como lo mencioné antes, desde la propia descarga de datos se descartaron todos los elementos innecesarios para la clasificación como lo son: las tablas, etiquetas HTML, URL y demás ítems que no aportarán nada a la clasificación de sentimientos, así que los datos se leen y pasan directamente a FinBERT.

FinBERT es un modelo que tiene una limitante considerable: la longitud del vector de entrada es de 512 tokens. Esta limitante representa un gran reto; a pesar de la reducción de nuestros documentos 10-K a unas cuantas páginas, estos siguen siendo lo suficientemente robustos para ser más grandes que 512 tokens. Para superar este obstáculo, se intentó ejecutar modelos como el T5 "Pipeline" en el Bloque 3. La intención con este modelo era resumir los textos para poder cumplir con el requisito de entrada para FinBERT. Este modelo básicamente extrae las palabras más importantes y genera nuevo texto. La idea era tomar esos resúmenes e incorporarlos a FinBERT. Lo que no se tuvo en cuenta es que es un modelo computacionalmente muy demandante; con una base de datos de 195 documentos, este se ejecutó durante media hora, tiempo en el cual nunca terminó. Eso para 195 documentos cuando nuestra base de datos es de 19 mil. Se descartó también el uso de BART (Bidirectional and Auto-Regressive Transformers) y BERT, ambos por la misma razón: un coste computacional muy alto.

Luego de los fallidos intentos con los modelos, se optó por hacer particiones a los documentos, esto de a 512 tokens. Consiste en que una vez el documento ingresa a FinBERT, se divide en x partes y cada parte se clasifica; finalmente, se realiza un promedio de las clasificaciones para asignar el sentimiento a la totalidad del documento. De nuevo no funcionó, la razón es el coste computacional. FinBERT, al pertenecer o tener su origen en el modelo BERT, también exige demasiado. Al realizar este proceso básicamente es multiplicar la base de datos por x cantidad. Asumamos que cada documento se particiona por 4, si este número es constante para cada documento sería equivalente a tener una base de datos de 19 mil por 4.



Finalmente, la técnica de menos agrado pero que si funciono fue el truncado se optó que cada documento o más bien cada vector tendrá una longitud 512 tokens lo demás se pierde, esta es la razón de desagrado con esta opción. Uno de los componentes principales de la clasificación con base a los sentimientos es que un sentimiento es positivo o negativo siempre y cuando haya un respeto a la sintaxis y semántica, el desechar o descartar las de más partes de los documentos se distorsiona todo.

Los documentos ya clasificados presentan sus resultados a partir de las probabilidades asignadas por el modelo. Se realiza un descripción de los resultados, con los siguientes hallazgos.



* **Recuento**: Hay un total de 19017 registros.
* **Media**: En promedio, los documentos tienen un sentimiento positivo del 11.2%, un sentimiento neutral del 19.7% y un sentimiento negativo del 69.1%. Esto indica que en general, los documentos tienden a tener un tono más negativo que positivo o neutral.
* **Desviación estándar**: Hay una gran variabilidad en los sentimientos, especialmente en los sentimientos positivos y negativos. Esto indica que, aunque en promedio los documentos pueden ser más negativos, hay muchos documentos que se desvían significativamente de la media.
* **Mínimo y Máximo**: Los valores mínimos y máximos indican que los puntajes de sentimiento varían en un rango bastante amplio, desde tan bajo como ~0.006 hasta tan alto como ~0.96.
* **Percentiles 25%, 50% y 75%**: Los cuartiles dan una idea de la distribución de los datos. Por ejemplo, el 50% por definición teórica es la mediana indica que la mitad de los documentos tienen un sentimiento positivo de al menos 3.8%, un sentimiento neutral de al menos 3.6% y un sentimiento negativo de al menos 85.4%.
* **Año**: Los documentos se distribuyen entre los años 2013 y 2023. La media de los años es 2017.53, es decir que la mayoría de los documentos están en el rango de 2017-2018. Sin embargo, la desviación estándar de 2.83 años sugiere que hay una dispersión considerable en la distribución temporal de los documentos.

Se aplica PCA, sobre los resultados del modelo de clasificación, en este caso sobre tres variables “Sentimiento Positivo” “Sentimiento Neutro” y “Sentimiento Negativo”

Se calcula los componentes principales de la matriz de diseño

A close-up of numbers

Description automatically generated with low confidence

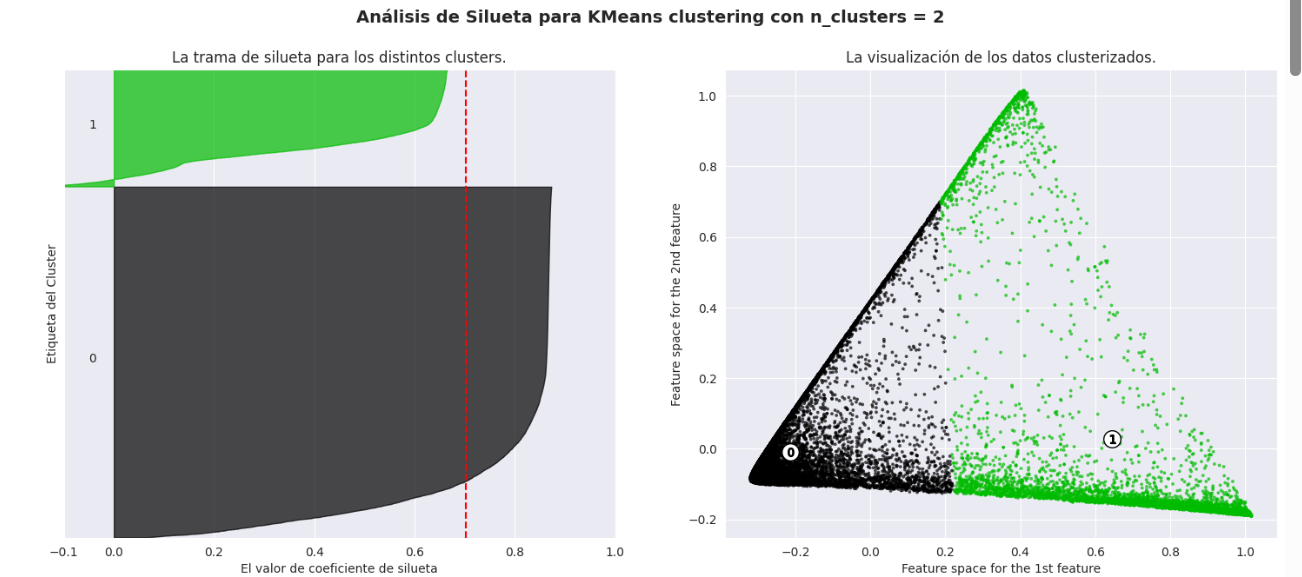
* El primer componente principal es 0.05037696
* El segundo componente de la variable original es 0.680571
* El tercer componente de la variable original es 0.73094803
  + El segundo componente está compuesto por 0.8149409 de la primera variable.
  + -0.4510981 de la segunda variable.
  + -0.36384267 de la tercera variable
    - La correlación indicada por el signo es la dirección entre la variable principal y el componente original.
    - Un coeficiente positivo indica correlación positiva a medida que la variable original aumenta el componente principal también lo hace
    - Una correlación negativa, a medida que la variable principal aumenta el componente principal disminuye.

variabilidad explica la primera componente principal



* El primer componente captura 77.29% de la varianza de los datos, es decir que recoge una gran información de los datos originales
* El segundo componente captura el 22.7% de la varianza de los datos.

Después de aplicar PCA, aplicamos K-mean



A picture containing text, screenshot

Description automatically generatedA picture containing text, screenshot, design

Description automatically generated

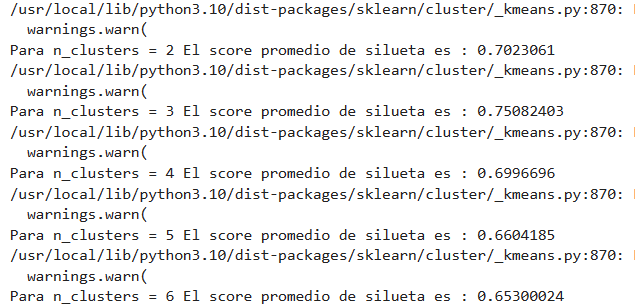
A picture containing text, screenshot, diagram, plot

Description automatically generated

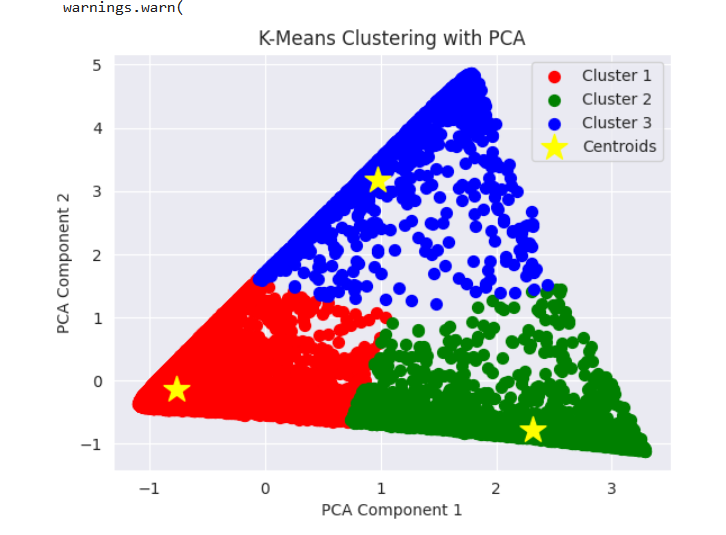
A picture containing text, screenshot, diagram, plot

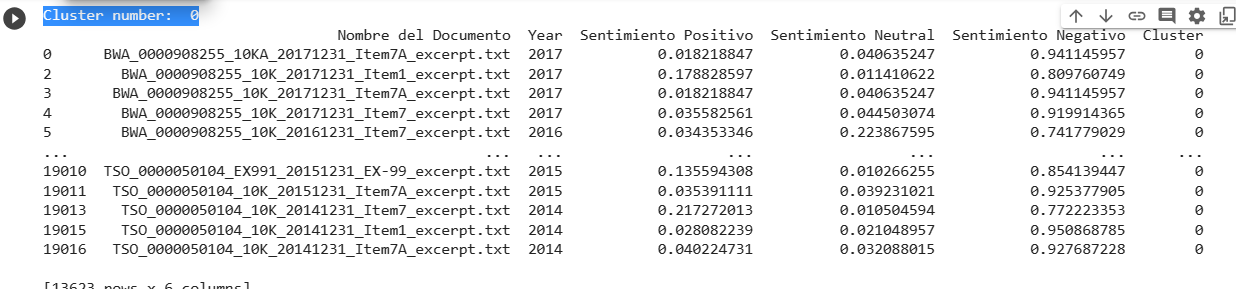
Description automatically generated

El “Silhouette score” es la medida para evaluar que tan efectivo fue la segmentación en algoritmos como K-mean, su rango es desde -1 hasta 1, siento en este ultimo la mejor segmentación cero, indiferente y -1 una mala segmemtacón. (scikit-learn, s.f.). Los puntajes se presentan a continuación:



Se determina bajo el Score o puntaje de Silhouette que K=3, es el K optimo.



Dentro de los clouster, K-mean agrupo las variables principalmente por su sentimiento:

A picture containing text, screenshot, font

Description automatically generated

A picture containing text, font, document, screenshot

Description automatically generated

**Conclusiones y Rechazo o no de la hipótesis nula.**

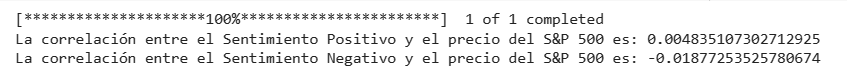
**Hipótesis Nula (H0):**

* No hay correlación entre el sentimiento positivo y el precio del S&P 500.
* No hay correlación entre el sentimiento negativo y el precio del S&P 500.

**Hipótesis Alternativa (Ha):**

* Existe una correlación entre el sentimiento positivo y el precio del S&P 500.
* Existe una correlación entre el sentimiento negativo y el precio del S&P 500.

Con la correlación de Pearson, se busco como esta correlacionado el precio y los sentimientos, en siguiente imagen, se observan los resultados.



Los valores están muy cerca a cero, por lo que no hay una relación lineal muy fuerte, basado en lo anterior se **rechaza la hipótesis nula**.

**Retroalimentación**

Uno de los elementos que considero clave, es que muchos de los modelos planteados en términos de computo son bastante exigentes, lo vimos con pipeline T5, tratando de resumir y capturar texto, también cuando se trato de hacer particiones en los textos para que fueran de 512 token, decisión que de nuevo aumento el costo computacional. Son elementos que me llevan a preguntarme, que tanto cambiaria el resultado final de poder llevarse a cabo.

Elementos que seguiré retomando, no solo por lo aprendido en relación con estos modelos sino también por la posibilidad de aumentar la base de datos a un millón. Esto plantea un enorme reto al tiempo que nace una motivación para continuar, no todo esta dicho con estos modelos, pueden ofrecer mas y desde luego ese más es proporcional a la exigencia de hardware y software, pero he allí la gracia del asunto, es ver que tan creativo se es ante un reto así.

**Conclusiones**

El modelo FinBERT es reconocido por su excelente capacidad para clasificar. *“Las palabras no operan de manera individual y aislada, sino en una cadena o ristra, unas antes que otras, ya que su significado propio puede alterarse dependiendo del lugar que ocupen en la oración”* (Concepto, 2023). FinBERT, al igual que BERT, son modelos que han demostrado que el contexto y la forma en que se dice una expresión determinan el sentimiento. Se pasó de contabilizar palabras con diccionarios, como el Harvard IV-4, a tener modelos precisos que permiten no solo captar la sintaxis, sino que también desde su concepción se especializan en un campo; en el caso de FinBERT, en las finanzas. Este es un sector en el que las palabras pueden tener un doble sentido, y los diccionarios de antaño se limitan a contabilizar, dejando de lado el contexto.

# Bibliography

Araci, D. T. (2019, 06 25). *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models.* Thesis , university of amsterdam, faculty of science, Amsterdam, The Netherlands. Retrieved 06 09, 2023, from https://arxiv.org/abs/1908.10063

Blasco., R. (2022, 8 1). Retrieved 6 10, 2023, from https://www.unprofesor.com/lengua-espanola/que-es-un-campo-semantico-y-ejemplos-5478.html

Bloomberg. (n.d.). Retrieved 5 28, 2023, from https://www.bloomberg.com/profile/company/101581Z:US#xj4y7vzkg

Campbell, T. (2023, 6 2). Bear market vs bull market: When should you invest? (L. G. Tarpley, Ed.) *INSIDER*. Retrieved 6 2, 2021, from https://www.businessinsider.com/personal-finance/bear-market-vs-bull-market#bear-market-vs.-bull-market

Concepto. (2023, 6 10). *Concepto*. Retrieved from https://concepto.de/sintaxis/

Dedeyne, S. B. (2019-2020). *Predicting the effect of 10-K, 10-Q and 8-K company reports on abnormal stock returns using FinBERT NLP methods.* Universiteit Gent, Faculteit economie. Retrieved from file:///D:/Desktop/Projecto%20Integrador/Proyecto%20Integrador%20Metodologia%20ASUM-DM%20(FASE%201)/Bibliografia%20De%20Soporte/Predicting%20the%20effect%20of%2010-K,%2010-Q%20and%208-K%20company%20reports%20on%20abnormal%20stock%20returns%20using%20FinBE

Deepl. (2023, 06 9). Retrieved from https://www.deepl.com/translator#en/es/The%20Transformer%20is%20an%20attention-based%20architecture%20for%20modeling%20sequential%20information%2C%20that%20is%20an%20alternative%0Ato%20recurrent%20neural%20networks%20%5B29%5D.%20It%20was%20proposed%20as%2

Genc, Z. (2020, 7 31). Retrieved from https://medium.com/prosus-ai-tech-blog/finbert-financial-sentiment-analysis-with-bert-b277a3607101

Jaadi, Z. (2023, 3 29). *A Step-by-Step Explanation of Principal Component Analysis (PCA)*. Retrieved 6 10, 2023, from https://builtin.com/data-science/step-step-explanation-principal-component-analysis

JavaTpoint. (n.d.). *JavaTpoint*. Retrieved 6 10, 2023, from https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

John Rambow, P. K. (2023, 6 2). What is a bull market? What it means, how it works, and when to invest. (L. G. Tarpley, Ed.) *INSIDER*. Retrieved 6 2, 2023, from https://www.businessinsider.com/personal-finance/what-is-a-bull-market

Machine Learning Google. (2022, 11 18). *Google*. Retrieved 6 10, 2023, from https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages?hl=es-419

scikit-learn. (n.d.). *electing the number of clusters with silhouette analysis on KMeans clustering*. Retrieved 6 11, 2023, from https://scikit-learn.org/stable/auto\_examples/cluster/plot\_kmeans\_silhouette\_analysis.html

Turney, S. (2022, 05 13). Retrieved 6 11, 2023, from https://www.scribbr.com/statistics/pearson-correlation-coefficient/

United States Securities And Exchange Commission. (n.d.). *ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF.* United States Securities And Exchange Commission, Washington, D.C. Retrieved from file:///D:/Desktop/Projecto%20Integrador/Proyecto%20Integrador%20Metodologia%20ASUM-DM%20(FASE%201)/Bibliografia%20De%20Soporte/form10-k%20How%20to%20use%20it.pdf

Wikipedia. (n.d.). *Wikipedia*. Retrieved 6 11, 2023, from Wikipedia: https://es.wikipedia.org/wiki/2022