General Electric Aerospace

Reto Datos

Flavio Ruvalcaba Leija - A01367631

Oscar Eduardo Nieto Espitia - A01705090

Eduardo Gonzalez Luna - A01658281

Cristian Rogelio Espinoza Diaz - A01702752

Anatanael Jesus Miranda Faustino - A01769232

04/10/2023

1. Herramientas y tecnologías.

	Big Data		No Big Data	
Data Storage	**Apache Hadoop** **Descripción**: Hadoop es un framework de software open source para el almacenamiento y procesamiento de grandes volúmenes de datos sobre clusters de computadoras sencillas usando modelos de programación simples. **Pros**: Escalabilidad: Permite gestionar grandes cantidades de datos en múltiples nodos. Tolerancia a fallos: Ofrece alta disponibilidad y resiliencia en caso de fallos de hardware. Costo-efectividad: Utiliza hardware asequible y es de código abierto. **Contras**: Requiere conocimientos técnicos: Configurar y administrar Hadoop puede ser complejo. Rendimiento de procesamiento en tiempo real: No es la mejor opción para aplicaciones que requieren procesamiento de datos en tiempo real. Uso intensivo de recursos: Puede requerir recursos significativos en términos de almacenamiento y potencia de cálculo.	**Mongodb** **Descripción**: MongoDB es un sistema de gestión de bases de datos NoSQL de código abierto que almacena datos en formato JSON, lo que permite una escalabilidad y flexibilidad significativas. **Pros**: Flexibilidad: Almacena datos en formato JSON, lo que facilita la adaptación a diferentes tipos de datos. Escalabilidad horizontal: Puede escalar horizontalmente para manejar grandes volúmenes de datos y tráfico. Rendimiento: Ofrece un buen rendimiento para operaciones de lectura y escritura. **Contras**: Falta de soporte transaccional: MongoDB carece de transacciones ACID completas en algunas configuraciones. Consumo de recursos: Puede ser intensivo en términos de recursos de hardware. No es la mejor opción para relaciones complejas: No es la elección ideal para aplicaciones con relaciones de datos altamente complejas.	**Almacenamiento de archivos** **Descripción**: El almacenamiento de archivos implica guardar datos en archivos individuales en sistemas de archivos locales **Pros**: Simplicidad: Es un enfoque sencillo para almacenar datos, adecuado para proyectos pequeños y medianos. Portabilidad: Los archivos se pueden mover y copiar fácilmente entre sistemas y dispositivos. Costo: Por lo general, el almacenamiento de archivos es asequible y no requiere inversiones significativas. **Contras**: Escalabilidad limitada: No es la mejor opción para gestionar grandes volúmenes de datos o necesidades de escalabilidad. Búsqueda y consulta limitadas: La búsqueda y consulta de datos pueden ser menos eficientes en comparación con sistemas de bases de datos. Falta de control de acceso avanzado: La gestión de permisos y control de acceso puede ser limitada en comparación con sistemas de bases de datos más avanzados.	**Bases de datos relacionales** **Descripción**: Las bases de datos relacionales (RDBMS) son sistemas de gestión de bases de datos que almacenan y administran datos en tablas relacionadas, usando el modelo de datos relacional. **Pros**: Estructura y relaciones: Ofrecen una estructura tabular que permite definir relaciones entre datos de manera eficiente. Transacciones ACID: Proporcionan transacciones ACID (Atomicidad, Consistencia, Aislamiento, Durabilidad) que garantizan la integridad de los datos. Consultas SQL: Emplean el lenguaje SQL estándar para consultas, que es potente y ampliamente conocido. **Contras**: Menor flexibilidad con datos no estructurados: No son ideales para datos no estructurados o semiestructurados. Escalabilidad vertical limitada: La escalabilidad vertical (aumento de recursos en un solo servidor) puede ser costosa y tiene un límite. Rigidez en el esquema: Cambiar el esquema de una base de datos relacional existente puede ser complicado.
	Apache Spark **Descripción**: Apache Spark es un framework de procesamiento de datos de código abierto que proporciona un entorno de análisis y procesamiento distribuido para grandes	**Dask** **Descripción**:Dask es una biblioteca de Python para el procesamiento paralelo y distribuido de datos. Permite manejar datos que no caben en la memoria RAM de un solo equipo, lo	**Pandas y NumPy** **Descripción**: son bibliotecas de Python utilizadas para el análisis y manipulación de datos. **Pros**: Facilita la manipulación y	**R Studio** **Descripción**: R Studio es un entorno de desarrollo integrado (IDE) de código abierto diseñado específicamente para trabajar con el lenguaje de programación R. Facilita la

Data Analytics	volúmenes de datos. **Pros**: Velocidad: Spark es conocido por su procesamiento rápido de datos, gracias a su capacidad de procesamiento en memoria. Escalabilidad: Puede escalar horizontalmente para manejar grandes volúmenes de datos y es altamente escalable. Diversidad de fuentes de datos: Puede trabajar con una variedad de fuentes de datos, desde lotes hasta transmisiones y datos en tiempo real. **Contras**: Requiere conocimientos técnicos: Configurar y administrar Spark puede requerir conocimientos técnicos avanzados. Consumo de recursos: Puede requerir recursos significativos en términos de memoria y potencia de cálculo. Curva de aprendizaje: Puede haber una curva de aprendizaje empinada, especialmente para usuarios nuevos en el procesamiento distribuido.	que es especialmente útil para análisis de datos a gran escala. **Pros**: Escalabilidad: Dask es altamente escalable y puede manejar grandes conjuntos de datos que no caben en la memoria RAM de un solo equipo. Esto lo hace adecuado para análisis de datos a gran escala. Paralelismo y Distribución: Permite el procesamiento paralelo y distribuido de datos, lo que acelera significativamente las operaciones en grandes conjuntos de datos. Integración con Librerías Existentes: Se integra bien con otras bibliotecas de Python como NumPy, Pandas y Scikit-Learn, lo que facilita su adopción en proyectos existentes. **Contras**: Curva de Aprendizaje: Puede tener una curva de aprendizaje empinada para quienes no están familiarizados con el modelo de cálculo de Dask. Requiere Configuración: Configurar un clúster de Dask para un entorno distribuido puede ser complicado y requerir recursos adicionales. Complejidad: Aunque es potente, puede ser una herramienta demasiado compleja para proyectos pequeños y simples.	análisis de datos tabulares. Proporciona estructuras de datos eficientes para operaciones matemáticas y científicas. Permite realizar cálculos de manera rápida y eficiente en grandes conjuntos de datos. **Contras**: Puede no ser la mejor opción para datos muy grandes debido a la sobrecarga de memoria. Algunas operaciones pueden ser lentas en comparación con lenguajes compilados como C++. No proporciona estructuras de datos con nombres de columna, lo que puede hacer que la manipulación de datos tabulares sea menos intuitiva que con Pandas.	escritura, prueba y depuración de código R, así como la creación de informes y visualizaciones. **Pros**: Entorno especializado: Está diseñado exclusivamente para trabajar con R, lo que lo hace altamente especializado y optimizado para tareas relacionadas con análisis de datos y estadísticas. Interfaz de usuario amigable: R Studio ofrece una interfaz de usuario amigable y organizada que facilita la escritura y ejecución de código R. Soporte para visualización: Permite producir visualizaciones interactivas y gráficos de manera eficiente. **Contras**: Especialización en R: Aunque es excelente para R, puede ser menos versátil para otros lenguajes de programación, lo que limita su utilidad en proyectos que requieren múltiples lenguajes. Curva de aprendizaje: Para quienes no están familiarizados con R, puede haber una curva de aprendizaje al principio. Recursos de sistema: Puede ser exigente en términos de recursos del sistema, especialmente cuando se trabajan con grandes conjuntos de datos.
	Tableau **Descripción**: Tableau es una plataforma de visualización y análisis de datos que permite a las organizaciones transformar datos en información visualmente atractiva e interactiva. Ofrece	**Looker** **Descripción**: Looker es una plataforma de inteligencia de negocios que permite a las organizaciones explorar, analizar y visualizar datos de manera colaborativa. Se centra en la creación de paneles de control y la	**Matplotlib** Descripción: Matplotlib es una biblioteca de visualización de datos en Python que proporciona una amplia variedad de herramientas para crear gráficos y visualizaciones de datos de alta calidad. Es	**Seaborn** **Descripción**: Seaborn es una biblioteca de visualización de datos en Python que se basa en Matplotlib y proporciona una interfaz de alto nivel para suscitar gráficos estadísticos, atractivos y

Data Visualization	herramientas para crear paneles de control, informes y visualizaciones de datos. **Pros**: Facilidad de uso: Tableau tiene una interfaz intuitiva de arrastrar y soltar que facilita la creación de visualizaciones y paneles de control. Visualización interactiva: Permite a los usuarios explorar datos de manera interactiva y realizar análisis ad hoc. Conexiones a múltiples fuentes: Puede conectarse a una variedad de fuentes de datos, incluyendo bases de datos, hojas de cálculo y servicios web. **Contras**: Costoso: Tableau puede ser costoso, especialmente para implementaciones empresariales y de alto nivel. Requiere aprendizaje: Aunque es fácil de usar, los usuarios pueden requerir tiempo para aprender a utilizarlo completamente. Limitaciones en análisis avanzado: Aunque es excelente para visualización de datos, puede tener limitaciones en el análisis estadístico avanzado y la generación de informes altamente personalizados.	generación de informes para ayudar a las empresas a tomar decisiones basadas en datos. **Pros**: Colaboración: Facilita la colaboración entre equipos y usuarios al permitir la creación compartida de informes y paneles de control. Exploración de datos en tiempo real: Permite a los usuarios explorar datos en tiempo real y ejecutar análisis ad hoc sin requerir conocimientos técnicos avanzados. Integración de datos: Se puede integrar con una variedad de fuentes de datos, bases de datos y servicios en la nube. **Contras**: Costoso: Looker puede ser costoso, especialmente para implementaciones empresariales y personalizaciones avanzadas. Curva de aprendizaje: Aunque está diseñado para ser amigable para los usuarios finales, puede haber una curva de aprendizaje para los administradores y creadores de informes. Requerimientos de infraestructura: La implementación puede requerir recursos de infraestructura, como servidores y almacenamiento de datos.	ampliamente utilizada en la comunidad de ciencia de datos y análisis. **Pros**: Amplia gama de gráficos: Matplotlib ofrece una variedad de tipos de gráficos, desde gráficos de dispersión y barras hasta gráficos de pastel y de caja, lo que permite representar datos de diferentes maneras. Personalización: Es altamente personalizable, lo que permite ajustar los gráficos y las visualizaciones de acuerdo a tus necesidades. Integración con NumPy y Pandas: Se integra fácilmente con otras bibliotecas populares de Python como NumPy y Pandas, lo que facilita la visualización de datos. **Contras**: Curva de aprendizaje: Para usuarios nuevos en Python y Matplotlib, puede haber una curva de aprendizaje para dominar sus capacidades. Sintaxis detallada: La creación de gráficos más avanzados puede requerir una sintaxis más detallada y compleja. No es interactivo: Matplotlib es principalmente una biblioteca para crear visualizaciones estáticas, por lo que no es la mejor opción para visualizaciones interactivas en aplicaciones web.	efectivos. Está diseñada específicamente para crear visualizaciones que resalten relaciones estadísticas en los datos. **Pros**: Simplicidad: Seaborn simplifica la creación de gráficos estadísticos complejos con una sintaxis más simple y concisa en comparación con Matplotlib. Visualizaciones atractivas: Produce gráficos estadísticos, atractivos y efectivos que resaltan las relaciones en los datos. Personalización sencilla: Aunque es más simple que Matplotlib, Seaborn permite una personalización razonable de las visualizaciones. **Contras**: Menos versátil que Matplotlib: Aunque es excelente para visualizaciones estadísticas, Seaborn puede ser menos versátil que Matplotlib para gráficos no estadísticos y personalizaciones avanzadas. Sintaxis específica para estadísticas: La biblioteca está más enfocada en la visualización de relaciones estadísticas, por lo que no es la mejor opción para gráficos no estadísticos. No es tan popular como Matplotlib: Aunque Seaborn ha ganado popularidad, Matplotlib sigue siendo la biblioteca de visualización de datos más ampliamente utilizada en Python.

2. Posibles Arquitecturas.

	Arquitectura 1	Arquitectura 2
Data Storage	**Mongodb:** Es importante destacar que MongoDB es una base de datos NoSQL que es excelente para datos semiestructurados o no estructurados. Esto significa que es una elección sólida si tus datos son diversos y no se ajustan bien a una estructura de tabla tradicional. MongoDB puede ser escalable y flexible, lo que es beneficioso en un contexto de "big data".	**Almacenamiento de archivos:** Los datos están alojados en una unidad de Box de la empresa, para descargarlos en la computadora y actualizarlos (en caso de que sufran modificaciones por parte del socio formador). Necesitamos de una carpeta en el sistema de almacenamiento de Windows para utilizar Box Sync, una herramienta que facilita la actualización y descarga de los archivos de datos. Otra razón para el uso de almacenamiento de archivos
Data Analytics	**Apache Spark:** Apache Spark es una de las tecnologías líderes para análisis de datos a gran escala. Es especialmente adecuado para procesar grandes volúmenes de datos de manera eficiente y permite realizar análisis complejos en tiempo real o por lotes. Si estás tratando con grandes conjuntos de datos, Spark puede ayudar a procesarlos de manera eficiente y realizar análisis avanzados.	**Dask y NumPy:** En cuanto a las tecnologías de análisis de datos, Dask y NumPy son excelentes opciones para el análisis de datos en Python. Se puede realizar operaciones de análisis y transformación de datos.
Data Visualization	**Tableau:** Cuando se trata de la visualización de datos, Tableau es una de las herramientas más populares y eficaces en el mercado. Ofrece una amplia gama de capacidades de visualización y generación de informes, y es ampliamente utilizado en empresas de todo el mundo. Puedes integrar Tableau con las fuentes de datos, incluido MongoDB, para crear paneles de control y visualizaciones interactivas.	**Seaborn:** Debido a la simplificación de la sintaxis y a la visualización de gráficos más atractivos y dinámicos es la razón por la que se escogió esta tecnología por encima de Matplotlib.

3. Elección de la arquitectura.

La segunda elección de Dask, NumPy y Seaborn en esta arquitectura tiene sentido por varias razones:

Dask y NumPy:

> Dask y NumPy son herramientas altamente eficientes para el procesamiento de datos en Python. Dask permite el procesamiento paralelo y distribuido de datos, lo que es crucial cuando se trabaja con grandes conjuntos de datos. NumPy es ampliamente reconocido por su velocidad y eficiencia en el cálculo de arreglos numéricos. Esta combinación es ideal para el análisis de datos, especialmente cuando se requiere un alto rendimiento.

Seaborn:

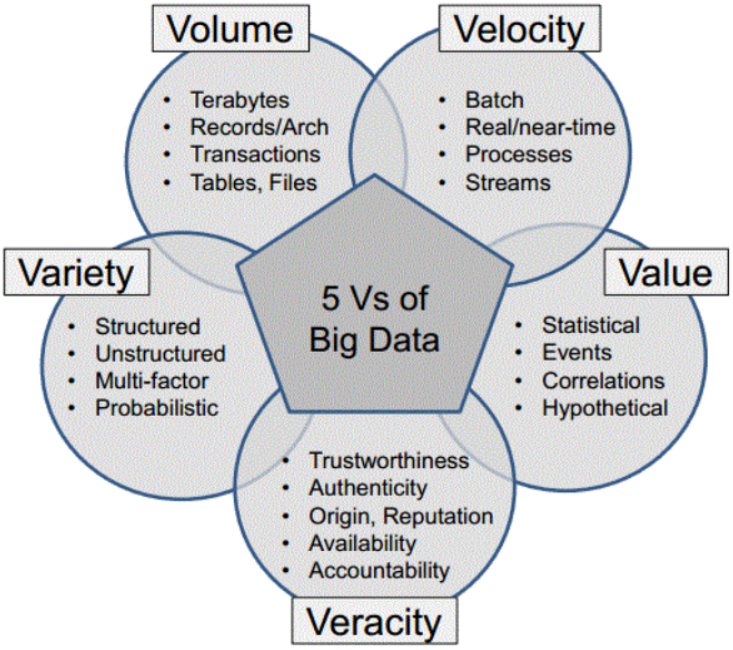
> Seaborn se destaca por su simplicidad y la creación de gráficos atractivos. La elección de Seaborn sobre Matplotlib se debe a su sintaxis más simple y a la generación de gráficos visualmente agradables. Esto es esencial para comunicar los resultados del análisis de datos de manera efectiva, especialmente en un entorno empresarial donde los resultados se deben presentar a diferentes partes interesadas.

Escalabilidad: Dask es altamente escalable, lo que significa que puede manejar grandes conjuntos de datos, lo que es una ventaja importante en un entorno empresarial donde los datos pueden ser voluminosos y crecer con el tiempo.

4. Big Data.

El término “Big Data” no se limita al empleo de tecnologías específicas como Apache Spark o bases de datos como MongoDB. Si bien es cierto que tecnologías como Spark y bases de datos NoSQL como MongoDB son comunes en el procesamiento y gestión de grandes volúmenes de datos, Big Data es un concepto más amplio y se refiere a la gestión y análisis de conjuntos de datos extremadamente grandes y complejos, independientemente de las tecnologías empleadas.

Es crucial destacar que Big Data también puede involucrar tecnologías y herramientas que no están específicamente diseñadas para Big Data, como pandas, NumPy y scikit-learn en el ecosistema de Python. Aunque estas bibliotecas son populares para el análisis de datos, son más adecuadas para conjuntos de datos que caben cómodamente en la memoria de una máquina y pueden no ser tan eficientes para tareas de Big Data.



Volumen: Tenemos 2,718 archivos CSV con un total de 40 columnas cada uno. El tamaño total de los datos es una consideración importante. Los datos no pueden caber cómodamente en la memoria de la máquina proporcionada por GE, entonces es posible que se necesite Big Data.

Velocidad: La velocidad se refiere a la rapidez con la que se generan y se deben analizar los datos. En el contexto de simulacros de avión, la velocidad es crucial para garantizar la seguridad y la eficiencia. La información debe procesarse en tiempo real o con la menor latencia posible para tomar decisiones oportunas. Los datos generados durante un simulacro de avión cambian constantemente y deben ser analizados instantáneamente para evaluar el rendimiento y prevenir accidentes. Sin embargo, nosotros no estaremos obteniendo datos a tiempo real, sino que serán otorgados una vez que han sido obtenidos del dispositivo FADEQ y trabajaremos con un set predefinido.

Variedad: La variedad se refiere a la diversidad de fuentes y tipos de datos. La diversidad de datos es innegable, lo que significa que se deben utilizar soluciones que puedan manejar datos estructurados y no estructurados, lo que suele ser una característica del Big Data. En simulacros de avión, los datos pueden variar desde enteros, flotantes y booleanos. En esta parte los datos son estructurados.

Valor: El valor de los datos es un criterio fundamental. ¿Qué valor pueden aportar los datos de estos simulacros de avión? Los datos pueden ser invaluables para mejorar la seguridad, la eficiencia y la capacitación en la aviación. Un análisis profundo de estos datos podría llevar a resolver la necesidad de nuestro socio formador. El valor potencial justifica la inversión en tecnologías de Big Data.

Veracidad: La veracidad se refiere a la confiabilidad de los datos. En el contexto de la aviación, la precisión de los datos es crucial. Un pequeño error podría tener consecuencias catastróficas. Sin embargo, estos datos han sido obtenidos mediante el dispositivo FADEC que obtiene datos en tiempo real de la simulación, comprobados y validados posteriormente por los expertos en General Electric, anonimizados y finalmente, entregados a nosotros.

Clasificación y Normalización: Implementaremos un proceso de clasificación y normalización de datos para asegurar la consistencia y facilitar el análisis.

Pipeline de Procesamiento de Datos: Estableceremos un pipeline de procesamiento que incluye ingestión, procesamiento, análisis y almacenamiento de datos en tiempo real.

Acceso Controlado: Garantizamos que el acceso a los datos está restringido y controlado para proteger la confidencialidad y la integridad.

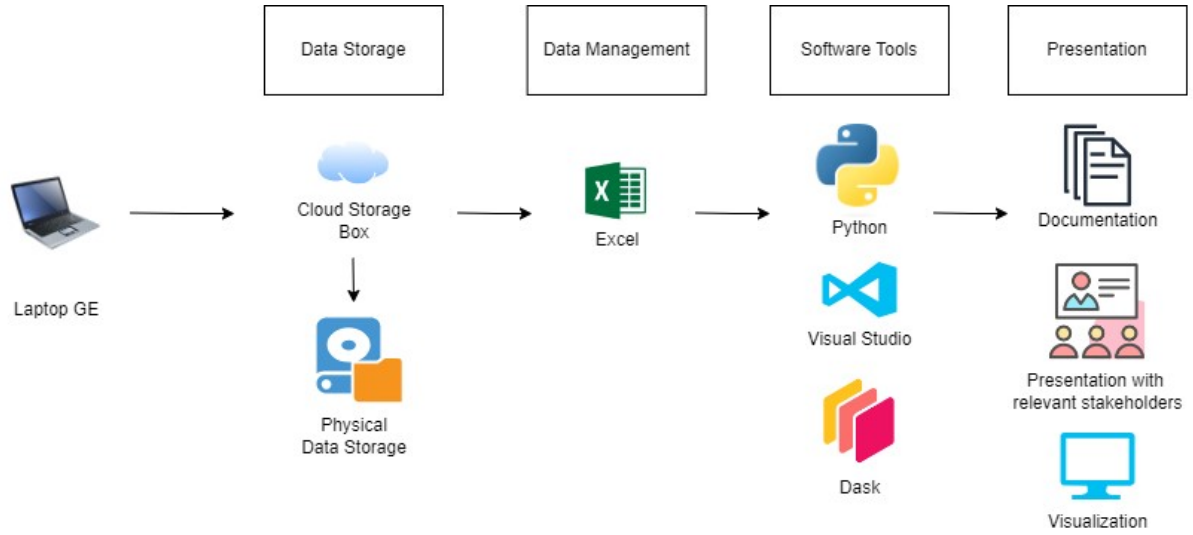
La necesidad de Big Data para el análisis de datos de simulacros de avión es evidente debido al volumen, velocidad, variedad, valor y veracidad de los datos. La adopción de tecnologías de Big Data y un modelo de almacenamiento es esencial para garantizar la seguridad, la eficiencia y el valor en la aviación. Este enfoque permitirá una mejor comprensión de los simulacros de avión, mejorando la seguridad y la eficiencia en la industria de la aviación.

5. Modelo de Almacenamiento y Manejo de Datos

Dada la necesidad de Big Data y los criterios de volumen, velocidad, variedad, valor y veracidad, se propone un modelo de almacenamiento y manejo de datos adaptado al reto:

Almacenamiento y acceso en Box	Los datos se almacenan en un Box con la seguridad de que es de la compañía General Electric, lo que proporciona una solución segura y accesible para almacenar y acceder a los archivos. A su vez, es donde el socio formador podrá actualizar los archivos csv en caso de anomalías o situaciones inusuales.
Descarga a la Computadora con Windows	Automatiza la descarga de datos desde Box a la computadora con Windows en intervalos regulares utilizando una herramienta de sincronización de Box conocida como Box Sync. Establecimos una ubicación específica en la computadora donde se guardarán los archivos descargados.
Gestión de Datos en la Computadora con Windows	Usamos un sistema de archivos organizado para mantener los datos descargados en la computadora.

Herramientas de Análisis	Emplearemos las tecnologías escogidas anteriormente para llevar a cabo análisis de datos en la computadora con Windows. Automatizar la carga y procesamiento de los datos desde la ubicación de descarga en la computadora.
Seguridad y Restricciones:	La computadora proporcionada es el único dispositivo autorizado para manejo y acceso a los datos. Debido a esto, tenemos acceso limitado a varias tecnologías. No se pueden enviar los datos de los archivos a plataformas en internet como Google Drive o Github, la única excepción es el Box Oficial de GE donde solo se puede acceder por el dispositivo previamente mencionado.



6. Separación de los datos en conjuntos para entrenamiento (train), prueba (test).

La estrategia definida para la separación de los datos en este contexto es utilizar la técnica de K-fold cross-validation es una técnica comúnmente utilizada en el campo del aprendizaje automático y la estadística para evaluar el rendimiento de un modelo predictivo o clasificador. Su objetivo es estimar qué tan bien se generaliza un modelo a datos no vistos.

Para esto se planea separar dividir los datos de la siguiente manera 80% para entrenamiento, 20% para prueba. Se planea usar diversas métricas para la evaluación de los modelos que se llegan a generar como la Accuracy, Precision, F1-Score y la matriz de confusión. Dichas métricas también pueden ser usadas para el tema de Cross Validation.

La técnica K-fold cross-validation elimina la necesidad de una división manual de datos en conjuntos de entrenamiento, validación y prueba. En su lugar, divide los datos en k pliegues,

entrenando el modelo k veces, con un pliegue como conjunto de validación en cada iteración. Esto proporciona una estimación más precisa del rendimiento del modelo, especialmente cuando los datos son limitados, reduciendo el riesgo de sesgos y resultados engañosos.

7. Links para scripts y dummies sobre la configuración y limpieza de datos

Con autorización e indicaciones tanto del socio formador como los profesores pudimos crear dummies con base en los datos de General Electric. Estos dummies nos permiten trabajar de manera asincrónica para crear código para analizar y crear modelos usando nombres de archivos, variables, entre otras cosas que son similares a como se encuentran dentro de la computadora de GE.

En la carpeta de Scripts, se encuentran los archivos .ipynb que hemos corrido dentro de la computadora de trabajo para el manejo y análisis de los datos.

Carpeta Dummies:

https://drive.google.com/drive/folders/1L_IEccCBnFDHSEK8a9EIwacIga-nI2Nx?usp=sharing

Carpeta Scripts:

https://drive.google.com/drive/folders/1ynf3kF6PxR37omAmRTMrEC-K92I2PaFt?usp=sharing

Referencias

Solís,Ismael.(2023). Módulo: Big Data. Google Drive https://drive.google.com/file/d/1Cl_RUMDonXri14x1J-x75CCF0mPszr66/view?usp=drive_link

DASK — DASK Documentation. (s. f.). https://docs.dask.org/en/stable/

Apache Hadoop. (2022). Apache Software Foundation. https://hadoop.apache.org/

MongoDB. (2022). MongoDB, Inc. https://www.mongodb.com/

Apache Spark. (2022). Apache Software Foundation. https://spark.apache.org/

McKinney, W. (2022). Pandas: Powerful data structures for data analysis.

https://pandas.pydata.org/

NumPy. (2022). Fundación de Software de Python. https://numpy.org

RStudio Team. (2022). RStudio: Integrated Development for R. RStudio, PBC.

https://www.rstudio.com/tags/python/

Tableau Software. (2022). Tableau. https://www.tableau.com/