**ANÁLISIS DE DATOS**

**TDSD**

|  |  |
| --- | --- |
|  | |
| ASIGNATURA: | ANÁLISIS DE DATOS |
| PROFESOR:  FECHA: | Ing. Lorena Chulde / Ing. Juan Pablo Zaldumbide  13 – 08 - 2024 |
| PERÍODO ACADÉMICO: | 2024-A |
|  | |

**PROYECTO FINAL – BIMESTRE 2**

|  |
| --- |
|  |
| herramientas de análisis de datos y su importancia | UdeCataluña |
|  |
| **Integrantes:**  **Karla Rodriguez**  **Isaac Quinapallo**  **Angel Villamil** |
|  |

**2024-A**

**Objetivo:**

Aplicar los conocimientos adquiridos sobre análisis de datos, extracción, limpieza, transformación, visualización de datos mediante la aplicación Power BI.

**Proyecto Final Análisis de Datos**

**Indicaciones:**

35% nota final

Fecha de entrega: 13 de agosto, defensa individual.

**Diseñar** una arquitectura en la cual confluyan al menos 10 fuentes de datos de al menos 5 de las siguientes temáticas. (Total 10 datasets)

1. Tráfico vehicular en las 5 principales ciudades del Ecuador.
2. Eventos deportivos a nivel mundial. (Juegos Olímpicos y Población)
3. Pulso político en 5 ciudades de Ecuador. ([Pulso político de 5 o más ciudades del Ecuador](https://github.com/JoseLuisColcha/Proyecto-Final-Data-Lake/tree/main/1.Pulso%20pol%C3%ADtico%20de%2020%20ciudades%20del%20Ecuador) y Provincias)
4. Actividades y hobbies. (Rating-games y Warzone)
5. Conciertos y eventos públicos.
6. Tema definido por el grupo. ([Enfrentamientos policiales en Estados Unidos](https://github.com/JoseLuisColcha/Proyecto-Final-Data-Lake/tree/main/4.Enfrentamientos%20policiales%20en%20Estados%20Unidos) y Estado y ciudad)
7. Restaurantes y sitios de esparcimiento.
8. Eventos o noticias mundiales. ([Covid-19](https://github.com/JoseLuisColcha/Proyecto-Final-Data-Lake/tree/main/5.%20Noticias%20y%20eventos%20mundiales) y sus respectivas muertes por países)

**Requisitos:**

* Se debe utilizar al menos dos bases de datos relacionales en su arquitectura (SQL Server, SQLite, MySQL, Mongo Atlas Oracle, PostgreSQL, etc.)
* Se debe utilizar al menos dos bases de datos NoSQL en su arquitectura (CouchDB, MongoDB, Riak, RavenDB, Redis)
* Se debe evidenciar el paso de datos de una base relacional a NoSQL y viceversa.
* Se debe evidenciar el paso de diferentes tipos de datos entre sí, por ejemplo, json a csv, csv a json, etc.
* El repositorio final debe ser SQL Server

Flecha

Descripción generada automáticamente con confianza baja

**Realizar:**

Crear dashboards explicativos de la información generada. Estos dashboards me permitirán tomar decisiones a problemáticas planteadas por cada uno de los grupos, se deben plantear al menos 15 casos de estudio.

Concluir sobre las visualizaciones y generar casos de estudio.

**Requisitos:**

Sumando los registros de los datasets escogidos, deben tener al menos un millón de registros.

Se debe trabajar con datos locales y en la nube.

**Entregables:**

# Informe (pdf, IEEE doble columna) (25%)

El informe debe contener al menos:

* Definición del caso de estudio.
* Objetivos General y específicos.
* Descripción del equipo de trabajo y actividades realizadas por cada uno.
* Cronograma de actividades incluído el diagrama de gantt (project)
* Recurso y herramientas utilizadas.
* Arquitectura de la solución.
* Extracción de datos.
* Análisis de información.
* Visualización de información.
* Resultados obtenidos.
* Conclusiones y Recomendaciones.
* Desafíos y problemas encontrados.
* Link de github del proyecto.

Se calificará aspectos tales como:

* Calidad del contenido.
* Cantidad de datos recopilados.
* Calidad del análisis.

**2. Videos explicativos (20%)**

* Un video explicativo en YouTube, el cual debe tener al menos 50 vistas, sobre el proceso completo de recopilación de datos. El video debe ser de 4 a minutos.
* Un segundo video explicativo en YouTube, el cual debe tener al menos 50 vistas, sobre las conclusiones de los dashboards. El video debe ser de 4 a minutos.

# Presentación del proyecto (ppt, pdf o prezi) (30%)

* Duración: mínimo de 20 minutos y debe contener los elementos técnicos y los mencionados en el informe.
* Preguntas: Cada asistente debe formular como mínimo una pregunta a un miembro del grupo expositor. Dicha pregunta será calificada tanto al que pregunta como al expositor.

**Nota:** la defensa se calificará de forma individual.

**3. URL del Repositorio de GitHub (25%)**

* Repositorio de GitHub con todos los script utilizados y explicando el proceso completo y las fuentes de datos.

<https://github.com/isaacquinapallo/ProyectoFinalAnalisisDeDatos.git>

**Nota:** Se tomará en cuenta la innovación en cualquiera de las etapas del proyecto, es un plus para la nota final.

**Caso de Estudio: Noticias y Eventos Mundiales**

**1. Definición del Caso de Estudio**

El caso de estudio se centra en el análisis global de eventos y noticias relacionados con el impacto del COVID-19 en varios aspectos, como la tasa de infección, la mortalidad, la vacunación y la polarización de opiniones. La finalidad es comprender mejor cómo se ha desarrollado la pandemia a nivel mundial y cómo ha afectado a diferentes países y poblaciones.

**2. Objetivos**

Objetivo General

Analizar y visualizar datos relacionados con el COVID-19 para identificar patrones y tendencias globales que permitan una mejor comprensión del impacto de la pandemia.

**3. Objetivos Específicos**

* Recolectar datos relevantes sobre el COVID-19 desde diferentes fuentes fiables como Kaggle y Statista.
* Transformar y limpiar los datos para asegurar su compatibilidad y consistencia para su análisis.
* Visualizar la información mediante herramientas de análisis de datos como Power BI, creando gráficos y mapas interactivos que faciliten la comprensión de los datos.
* Interpretar los resultados obtenidos para proporcionar insights significativos sobre la evolución de la pandemia y su impacto global.

**4. Actividades Realizadas**

*- Extracción de Datos:* Isaac Quinapallo se encargó de la extracción de datos relevantes sobre el COVID-19 de plataformas como Kaggle y Statista.

- Transformación de Datos: Se desarrollaron scripts para transformar y limpiar los datos obtenidos, asegurando su coherencia y formato adecuado.

- Visualización: Se integraron los datos en Power BI, creando gráficos y mapas interactivos que representan la evolución y el impacto del COVID-19.

- Análisis de Resultados: Interpretación de los datos visualizados para identificar patrones y proporcionar recomendaciones basadas en los resultados.

**5. Cronograma de Actividades**

Diagrama de Gantt (incluido en Project)

Se detalla el cronograma de actividades desde el 26 de julio hasta el 13 de agosto, donde se distribuyeron las tareas de manera equitativa entre los miembros del equipo.

(El diagrama de Gantt debería estar incluido como una imagen o gráfico adjunto al informe).

**6. Recursos y Herramientas Utilizadas**

- Fuentes de Datos: Kaggle, Statista.

- Herramientas de Transformación y Limpieza: Scripts personalizados en Python para conversión de formatos y limpieza de datos.

- Herramientas de Visualización: Power BI para la creación de gráficos y análisis visuales.

- Bases de Datos NoSQL: MongoDB para el almacenamiento y gestión de datos estructurados y no estructurados.

**7. Arquitectura de la Solución**

La arquitectura de la solución incluye la recolección de datos en formatos JSON, CSV y XLS, su transformación y limpieza, almacenamiento en bases de datos MongoDB, y posterior visualización en Power BI. La estructura permite un flujo de trabajo eficiente desde la obtención de datos hasta la presentación de resultados visuales.

**8. Extracción de Datos**

Los datos fueron extraídos de Kaggle y Statista, centrados en cinco conjuntos clave:

- FechasPersonasCuradas

- PaisesPersonasPolaridad

- PaisesPersonasEnfermas

- PaisesPersonasMuertesEnfermedad

- PaisesPersonasVacunadas

Estos datos fueron posteriormente transformados para su análisis.

**9. Análisis de Información**

El análisis se centró en identificar patrones y tendencias en la evolución del COVID-19, utilizando gráficos comparativos, mapas de calor y series temporales que muestran la propagación y el impacto del virus en diferentes regiones.

**10. Visualización de Información**

Se utilizaron las capacidades de Power BI para crear dashboards interactivos que presentan la información de manera clara y comprensible. Las visualizaciones incluyeron:

- Mapas de calor que muestran la propagación del virus.

- Gráficos de barras y líneas que comparan tasas de infección, mortalidad y vacunación entre países.

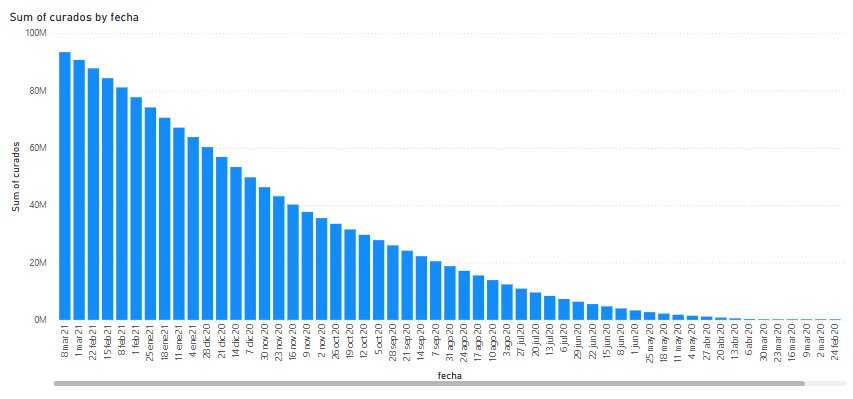
- Series temporales que rastrean la evolución de la pandemia.

**11. Resultados Obtenidos**

Los resultados revelaron patrones clave en la propagación del COVID-19, con diferencias notables entre países en términos de respuesta y manejo de la pandemia. Los gráficos mostraron correlaciones entre las tasas de vacunación y la reducción de casos y muertes.

**Tasa de Curación Por Fechas**

*Figura*



*Nota: En este grafico se muestra la tasa de vacunación por Pais de acuerdo a los índices generados por Who (World Health Organization)*

**12. Conclusiones y Recomendaciones**

- Conclusiones: La pandemia del COVID-19 ha tenido un impacto diverso en diferentes regiones, influenciado por factores como políticas de salud pública, acceso a vacunas y respuesta gubernamental.

- Recomendaciones: Se recomienda un enfoque coordinado a nivel global para futuras pandemias, con un énfasis en la equidad en el acceso a recursos médicos y vacunas.

**13. Desafíos y Problemas Encontrados**

Durante el proyecto, algunos desafíos incluyeron la heterogeneidad de los formatos de datos y la necesidad de limpiar y transformar grandes volúmenes de información. También se encontraron limitaciones en la disponibilidad de datos actualizados en ciertas regiones.

**14. Link de GitHub del Proyecto**

El repositorio del proyecto, que incluye todo el código, datasets, y documentaciones, está disponible en:

<https://github.com/isaacquinapallo/ProyectoFinalAnalisisDeDatos/tree/66bc726a55323f9da2c1be05f5b3a52ba398a679/5.-%20Noticias%20y%20Eventos%20Mundiales>

y en Youtube:

<https://youtu.be/bLGqD7c-uIU>