UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA
INTELIGENCIA ARTIFICIAL

Proyecto inteligencia artificial

Entrega 2

Integrantes
Manuela Arteaga Arango
Rosa Puerta Campo

Documentos
1.152.710.365
1.002.389.631

Docente
Raúl Ramos Pollan

Medellín-Antioquia
2023

Descripción del proceso avanzado

1. **Visualización de dataset**

Se generó un código para automatizar el proceso de descarga y preparación de los datos sobre COVID-19 que vamos a utilizar en este estudio. En este código se realizan una serie de acciones para descargar, descomprimir y visualizar un conjunto de datos relacionados desde Kaggle. En primer lugar, se configura el archivo kaggle.json para poder acceder a Kaggle y descargar el conjunto de datos. Luego, se instalan algunas bibliotecas útiles, como numpy, matplotlib y pandas. A continuación, se crea un directorio .kaggle en el directorio de inicio y se copia el archivo kaggle.json allí, asegurando que tenga los permisos adecuados. Luego, se descarga un conjunto de datos llamado covid19-dataset desde Kaggle y se descomprime. Finalmente, se lee el archivo CSV resultante llamado 'owid-covid-data.csv' en un DataFrame llamado dataset y se muestra su contenido, el cual contiene 166326 filas y 67 columnas.

Nuestro dataset contiene información sobre los casos de covid 19 recopilados alrededor de algunos países del mundo. Algunas de las columnas que nos llamaron la atención fueron las siguientes:

- location: Contiene el país a donde está asociado el registro
- total_cases: Contiene el conteo de casos positivos totales hasta el momento
- new_cases: Contiene la información de nuevos casos de Covid reportados.
- total_deaths: Contiene la información de las muertes causadas por el covid hasta ese día.
- new_deaths: Contiene la información de las nuevas muertes reportadas en esa fecha.
- icu_patients: Contiene el conteo de pacientes en UCI hasta ese día.
- hosp_patients: Contiene el conteo de pacientes que han sido hospitalizados hasta el momento.
- people_vaccinated: Contiene el conteo de personas que han sido vacunadas hasta el momento.
- new_vaccinations: Contiene el número de nuevos vacunados ese día.
- date: Contiene la fecha del registro.

Sin embargo, las cinco primeras columnas son las de especial interés en nuestro estudio.

2. **Identificación de dataset**

En la segunda parte, el objetivo principal fue comprender cómo es la distribución de la información y cómo se comportan los datos.

Para ello se muestran las columnas disponibles en el conjunto de datos y se identifican los países involucrados en los registros. Luego, se cuenta el número de registros por país para determinar cuál país tiene la mayor cantidad de datos y se seleccionan los países que tiene la mayor cantidad de datos registrados en latino américa para su estudio.

Además, para entender mejor cómo se comportan los datos y cuánta información contiene, se llevaron a cabo mediciones estadísticas media, mediana, desviación estándar) para las columnas relacionadas con el total de muertes, total de casos y nuevos casos. Se generan histogramas para visualizar la distribución de estos datos.

Estas pruebas proporcionaron información valiosa sobre la naturaleza de los datos, como si siguen una distribución específica, si existen correlaciones significativas entre variables, entre otros aspectos. Este análisis inicialmente reveló que Argentina y México tenían el mayor número de registros, lo que sugiere que estos dos países podrían ser candidatos adecuados para un análisis más profundo en el proyecto.

3. Reconocimiento de datos

Para comprender la esencia de nuestro proyecto, que radica en la predicción precisa de la evolución de la pandemia del COVID-19, nos embarcamos en un proceso de reconocimiento de datos. Nuestra misión es desarrollar un modelo que pueda ofrecer predicciones sólidas en lo que respecta al número de casos nuevos de COVID-19 en diversas naciones o regiones. Con esta finalidad en mente, evaluamos las posibles variables objetivo a utilizar, entre las cuales se incluyen 'new_cases', 'new_cases_per_million' y 'new_cases_smoothed', todas ellas disponibles en el conjunto de datos.

Nuestra primera etapa de reconocimiento consistió en la identificación de datos faltantes. Para ello, creamos un código que nos permitiera calcular la cantidad de valores ausentes en cada columna del conjunto de datos. Durante este proceso, encontramos la presencia de columnas con una cantidad significativa de datos nulos, incluso llegando a registrar hasta 160,630 valores faltantes. Es importante destacar que el conjunto de datos completo consta de un total de 166,326 filas, lo que subraya la relevancia de abordar esta problemática de datos nulos.

Continuando con nuestro análisis, exploramos la distribución de los datos correspondientes a la variable objetivo seleccionada y observamos que esta distribución no sigue una forma gaussiana. Además, examinamos los tipos de datos representados en cada columna del conjunto de datos. Profundizamos en el análisis de las columnas numéricas, examinando estadísticas clave como la media, la desviación estándar y otros parámetros relevantes.

Además, exploramos la correlación entre las distintas columnas de manera gráfica, lo que nos proporcionó una visión más clara de las relaciones entre las variables, con el fin de comprender su influencia en el conjunto de datos y en nuestras futuras predicciones.

Tras esta fase inicial de reconocimiento de datos, continuamos con el proceso de limpieza y preparación de los datos. Abordamos las columnas que presentaban datos faltantes y consideramos diversas opciones para su tratamiento. Estas alternativas incluyeron la posibilidad de eliminar filas o columnas con datos faltantes, así como la opción de reemplazar los valores nulos. En este último caso, evaluamos la sustitución por ceros, la media de los otros datos de la columna o incluso la imputación de datos aleatorios dentro del rango proporcionado por las demás filas.

El próximo desafío en nuestro proyecto implica decidir el enfoque que adoptaremos para abordar las columnas y filas que presentan una notoria ausencia de datos. Para abordar esta cuestión, llevaremos a cabo pruebas de hipótesis. Actualmente, uno de los retos más destacados consiste en tomar decisiones informadas sobre qué columnas podemos eliminar sin que esto impacte negativamente en la precisión de nuestro modelo predictivo. Es importante considerar que existen columnas que no desempeñan un papel determinante en la generación de nuevos casos de COVID-19 y, además, cuentan con una alta cantidad de datos faltantes.