UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA
DEPARTAMENTO DE BIOINGENIERÍA
Inteligencia Artificial

Proyecto
Entrega 2

Nombres
Nataly Rodriguez Ateourtua
Juan Camilo Cataño Zuleta
Daniel Enrique López Yepes

Docente
Raul Ramos Pollan

Documentos
1.001.505.918
1.001.738.289
1.000.407.016

Medellín - Antioquia
2022

Descripción del progreso alcanzado:

Debido a problemas con el set de datos anteriormente hubo necesidad de cambiarlo a último momento. El problema surgió en que mucha de la información que contenía no era de mucha ayuda, mucha información censurada y demasiados datos inválidos/vacíos, por lo cual al hacer una pequeña limpieza con el método .dropna() el dataset quedaba completamente vacío.

Nuestro dataset inicial era este:

```
[14] #Importamos pandas y vemos la información
     import pandas as pd
     data = pd.read_excel("/content/dataset.xlsx", engine="openpyxl")
     data
```

	Patient ID	Patient age quantile	SARS-Cov-2 exam result	Patient addmited to regular ward (1=yes, 0=no)	Patient addmited to semi-intensive unit (1=yes, 0=no)	Patient addmited to intensive care unit (1=yes, 0=no)	Hematocrit	Hemoglobin	Platelets	Mean platelet volume	...	Hb saturation (arterial blood gases)	pCO2 (arterial blood gas analysis)	Base excess (arterial blood gas analysis)	pH (arterial blood gas analysis)	Total CO2 (arterial blood gas analysis)	HCO3 (arterial blood gas analysis)	(arte blood analy
0	44477f75e8169d2	13	negative	0	0	0	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	
1	126e9dd13932f68	17	negative	0	0	0	0.236515	-0.022340	-0.517413	0.010677	...	NaN	NaN	NaN	NaN	NaN	NaN	
2	a46b4402a0e5696	8	negative	0	0	0	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	
3	f7d619a94f97c45	5	negative	0	0	0	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	
4	d9e41465789c2b5	15	negative	0	0	0	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
5639	ae66feb9e4dc3a0	3	positive	0	0	0	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	
5640	517c2834024f3ea	17	negative	0	0	0	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	
5641	5c57d6037fe266d	4	negative	0	0	0	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	
5642	c20c44766f28291	10	negative	0	0	0	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	
5643	2697fdccbfeb7f7	19	positive	0	0	0	0.694287	0.541564	-0.906829	-0.325903	...	NaN	NaN	NaN	NaN	NaN	NaN	

5644 rows × 111 columns

Al realizar el .dropna() el resultado fue el siguiente:

```
[16] #Como hay mucha información faltante se va a hacer un vaciado de NaN, pues es información que no nos sirve de mucho
     filter=data.dropna()
     filter
```

	Patient ID	Patient age quantile	SARS-Cov-2 exam result	Patient addmited to regular ward (1=yes, 0=no)	Patient addmited to semi-intensive unit (1=yes, 0=no)	Patient addmited to intensive care unit (1=yes, 0=no)	Hematocrit	Hemoglobin	Platelets	Mean platelet volume	...	Hb saturation (arterial blood gases)	pCO2 (arterial blood gas analysis)	Base excess (arterial blood gas analysis)	pH (arterial blood gas analysis)	Total CO2 (arterial blood gas analysis)	HCO3 (arterial blood gas analysis)	pO2 (arterial blood gas analysis)	Arteiral Fio2

0 rows × 111 columns

El nuevo dataset a utilizar sigue por la línea del covid 19, en este caso es un dataset con información que vincula a todo el mundo. El link del kaggle es el siguiente: https://www.kaggle.com/datasets/georgesaavedra/covid19-dataset

Para esta entrega nuestro enfoque fue básicamente comprender qué es lo que teníamos con esta información, reestructurar el proyecto que ya teníamos pero con este nuevo set de datos, aplicar pruebas de hipótesis para determinar la distribución de la información y establecer la calidad del dataset.

Para esto en primera instancia se visualiza el set de datos y se estudia algunas medidas estadísticas, como es el caso de la media, la mediana y la desviación estándar. Posteriormente estudiamos los datos por medio de un histograma y un diagrama de cajas y bigotes, el cual indica las medidas estadísticas para las columnas correspondientes a un total de casos, nuevos casos y total de muertos. Con el fin de conocer cómo es la distribución de información del arreglo, se considera necesario aplicar algunas pruebas de hipótesis y otras mediciones para conocer cómo se comportan los datos, cuánta cantidad de información tenemos, entre otros valores.

En el dataset, se recopiló datos de todos los países del mundo sobre múltiples indicadores, desde el comienzo del brote de COVID-19, por este motivo se decide contar el número de registros por país, para determinar cuál país estudiar, evidenciando que Argentina y México presentaban el mayor número de registros.
Para Argentina se podía evidenciar que el número de casos está cayendo rápidamente mientras que la curva de nuevas muertes en algunos casos parece haberse estabilizado. Para corroborar esto, se visualizan las curvas obtenidas de enero de 2021, evidenciando que el proceso de vacunación inició y se vacunan más personas, presentando que la cantidad de casos nuevos y muertes han bajado, especialmente en abril, estos números no tienen precedentes y el número de personas que reciben la primera dosis de la vacuna es ligeramente inferior al 50%. Como resultado, el número de muertos aún no se ha disparado a pesar de que la omicronia se ha convertido en la enfermedad más infecciosa de todos los tiempos.

Algo destacable en los patrones de Argentina es que indica que hay 14 olas durante 3 meses, esto casi coincide con el número de semanas, presentando como curva más preocupante la del número de nuevas muertes, porque es extraño que ese número suba y baje tan violentamente cada semana.Resaltando que se debe realizar un estudio más exhaustivo para comprender mejor los datos presentados.