ENTREGA 2: PROYECTO DE ANALÍTICA DE DATOS

PREDICT CO2 EMISSIONS IN RWANDA.

INTEGRANTES:

LAURA CRISTINA DIAZ OSORIO.

C.C: 1018351214

JUAN FELIPE ESCOBAR RENDÓN.

C.C:1001416321

CURSO:

INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

TUTOR:

RAÚL RAMOS POLLÁN

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

AMALFI-ANTIOQUIA

2023

PREDICT CO2 EMISSIONS IN RWANDA.

1. Exploración de datos.

Iniciamos la visualización y exploración de los datos obtenidos, realizando, histogramas, gráficos de correlación e interacciones entre las variables para tratar de identificar patrones y distribuciones que estos tengan.

1.1. Variable predictora.

Procedemos con la exploración de la variable a predecir, con esto queremos observar, la distribución que sigue la variable 'emission', la presencia de valores atípicos los cuales podrían afectar la precisión en las predicciones futuras.

1.1.1 Observación de datos atípicos en la variable respuesta.

En el histograma se puede observar que la variable a predecir no sigue una distribución normal y si bien la mayoría de los datos se encuentran a la izquierda del histograma, es claro que estos tienen un sesgo bastante amplio a la derecha, indicando que se presentan valores muy grandes lo que puede generar un problema a la hora de observar la precisión en las predicciones; sin embargo, para esta primera versión de limpieza del dataset, se optó por no realizar ninguna transformación y conservar todos los datos atípicos de la variable respuesta.

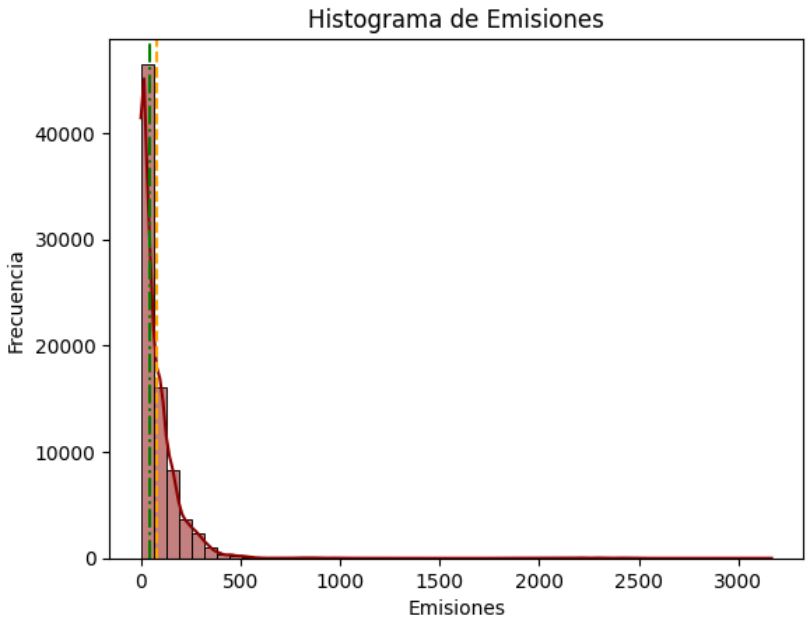

Fig. 1

1.1.2 Gráfica de serie de tiempo.

Se grafica la variable emisión como una serie de tiempo para observar patrones temporales que esta posee en los datos, lo que puede ser beneficioso al momento de escoger y descartar posibles modelos predictivos, para este caso se identifican algunos datos atípicos en la varianza de los datos, lo que nos indicaría la posibilidad de usar algún tipo de transformación o diferenciación para los datos en caso de realizar un modelo de serie de tiempo.

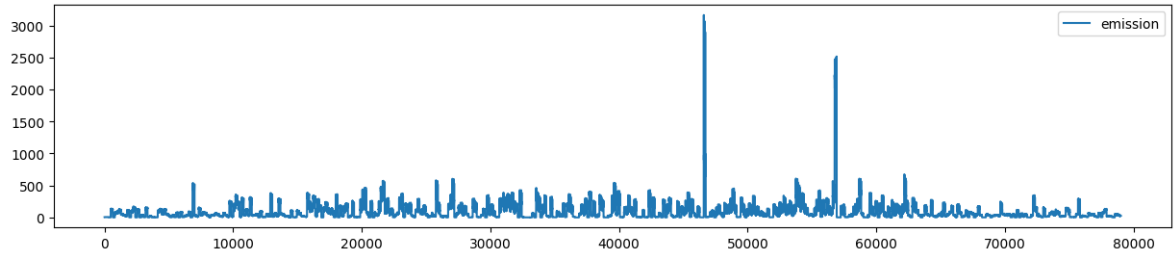

Fig. 2

1.2 Observación general.

Ahora procedemos a visualizar el comportamiento de las otras variables que conforman la base de datos en los que podremos observar correlaciones que las variables tienen entre ellas y con la variable predictora.

	Variable	Correlación
0	emission	1.000000
1	longitude	0.102746
2	UvAerosolLayerHeight_aerosol_height	0.069008
3	Cloud_surface_albedo	0.046587
4	Formaldehyde_tropospheric_HCHO_column_number_d...	0.040263
...	...	...
70	Formaldehyde_tropospheric_HCHO_column_number_d...	-0.033333
71	NitrogenDioxide_solar_azimuth_angle	-0.033417
72	CarbonMonoxide_CO_column_number_density	-0.041328
73	CarbonMonoxide_H2O_column_number_density	-0.043217
74	UvAerosolLayerHeight_aerosol_pressure	-0.068138

75 rows × 2 columns

Fig. 3

Se puede visualizar que la correlación lineal general que tiene la variable a predecir con las otras es bastante baja por lo que se podría decir que un modelo de regresión lineal no sería óptimo ya que las variables se pueden estar teniendo interacciones más complejas entre ellas.

2. Preprocesado y limpieza de datos

2.1 Manejo de datos faltantes.

Para este punto encontramos que el dataset cumple con el requisito de al menos 5% de datos faltantes en al menos tres columnas, igualmente se observó que en algunas de ellas el porcentaje de datos faltantes llegaba hasta más del 50% por lo que optamos por eliminar las columnas que superan este valor y se procedió con la imputación de datos. El método usado para la imputación de datos fue el de crear una muestra con distribución normal a partir la media y desviación de cada columna de para así evitar sesgos en la estimación de la media y varianza de estos.

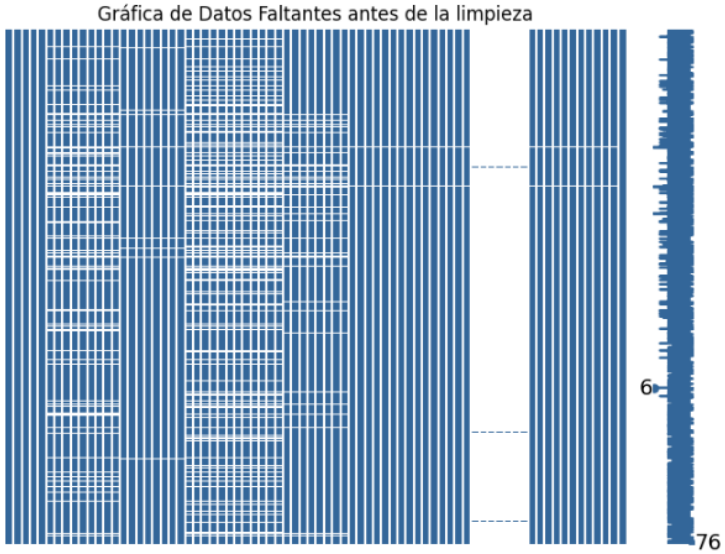

Fig. 4

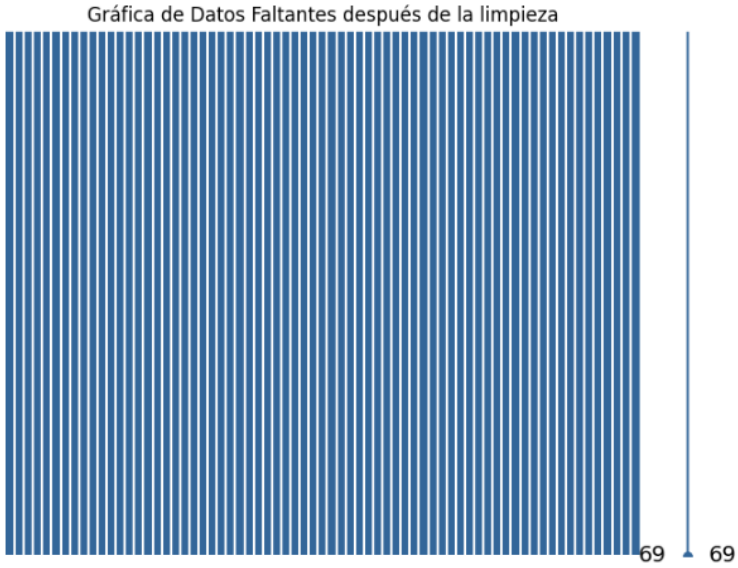

Fig. 5

2.2 Columnas categóricas.

Como el dataset no cumple con el requisito del 10% de columnas categóricas mínimas, por ende para cumplir con este requisito escogimos las columnas con menor correlación lineal, para proceder a discretizar estas, en intervalos y en cada intervalo categorizar con un número entero.

Comparación gráfica de columnas originales vs categorizadas

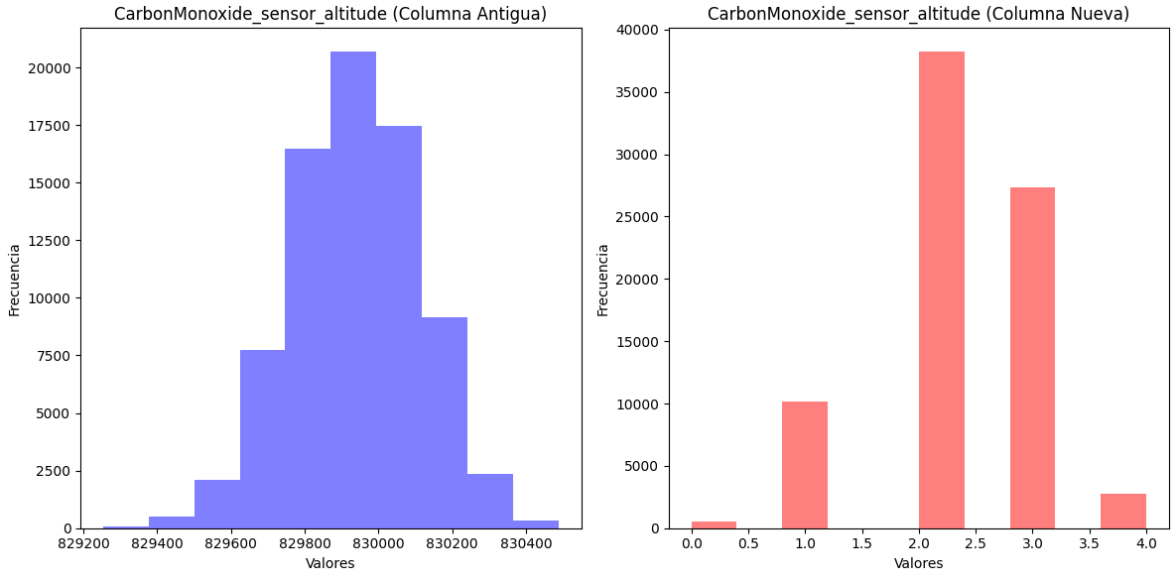

Fig. 6