Práctica 18:

Elaboró: Carlos Alejandro Jarero Gonzalez al255813@alumnos.uacj.mx

Matrícula: 255813

Marzo 19, 2025

Reporte

(A) Análisis Exploratorio de Datos (EDA):

- ¿Qué patrones o tendencias observaste en los histogramas y gráficas de densidad (PDF)?

Antes del manejo de los valores perdidos.

En los histogramas y en las gráficas de Probability Density Function (PDF) podemos visualizar las distribuciones de nuestras variables. También podemos evaluar visualmente el sesgo y curtosis. Observamos dos cosas importantes:

- Existen datos extremos que llaman la atención, más precisamente el valor de: `-200`. Este valor parece a simple vista ruido y pudieran ser outliers. Esto llamó la atención lo que me llevó a buscar la fuente del Dataset donde se describe este valor como la etiqueta un valor perdido.
- Otra cosas que podemos observar direcamente, es que muchas de las variables tienen sesgo a la derecha e izquierda, aunque esto esta afectado en este momento por los valores perdidos etiquetdos con el `-200`. En cuanto a la curtosis se observa un predomino de la leptocurtosis y otras con tendencia a la mesoscurtosis. En el notebook se pueden verfificar los valores de curtosis y sesgo.

Después del manejo de los valores perdidos.

Posterior a retirar los valores perdidos e imputarlos, encontramos que las distribuciones cambiaron. Con la ausencia de los `-200`, se observan menos sesgos, en otras palabras distribuciones más centradas, sin embargo la mayoria siguen teniendo sesgo a la derecha.

- ¿Alguna variable parece seguir una distribución normal?

Antes del manejo de los valores perdidos.

Visualmente ningúna parece ses una distribución normal posiblemente secundairo a los valores perdidos etiquetados con `-200`. Esto lo podremos evlauar una ves tratados e imputados los datos.

Después del manejo de los valores perdidos.

Posterior a retiro imputación de los **NaN** y los **-200** encontramos diferencias marcadas contrastando con la primer gráfica. Pero, aún así ninguna pareciera

tener distribución normal en cuanto a sesgo la mas central es rh, pero tiene tendencia hacia lo platicúrtica.

- ¿Qué información útil obtuviste de las gráficas de barras para las variables categóricas (por ejemplo, DayOfWeek)?

Muestra gráficamente el número de clases en la variable categrorica y la frecuencia de estas clases. En este caso vemos que los siete días de la semana tienen una distribución uniforme casi perfecta, que sería lo esperado de la medición continua de sensores. Los días que tienen más mediciones son los viernes, sábados, domingos y jueves.

- ¿Identificaste outliers en los boxplots?

Antes del manejo de los valores perdidos.

Se observan una cantidad importante de outliers principalmente en **nox_gt** y **nmhc_gt**, pero recordemos que en este punto aún existen los valores perdididos marcados como **-200** y con este ruido podemos mal interpretar estos gráficos.

Después del manejo de los valores perdidos.

Posterior al tratamiento datos e imputación de los datos perdidos obtenemos ahora si una visión más real de los posibles outliers de nuestro DataSet. Y seguimos observando outliers en todos menos en **rh**, es importante tomar en cuenta que las imputaciones que se hicieron fueron simples y desconocemos con certeza la causa de la falta de algunos datos y por lo tanto la imputación pudiera no ser la adecuada.

(B) Pruebas de Normalidad:

- ¿Qué conclusiones obtuviste de las pruebas de normalidad (Shapiro-Wilk, Anderson-Darling, Kolmogorov-Smirnov)?

Antes del manejo de los valores perdidos.

Nota: como en el primer paso no se habían tratado los valores perdidos incluidos los NaN se tuvieron que eliminar para realizar cada uno de los cálculos de lo contrario no era posible obtener los valores del estádistico ni de p.

Como era de esperarse y acorde a lo observado en los histogramas y las gráfica de PDF, vemos que en todos los casos las distribuciones parecen no ser normales. En todos los casos de variables númericas se rechaza la hipotesis nula, por lo tanto tenemos evidencia suficiente que apoya que los datos no tienen una distribución normal. Los puntos de corte evaludos (alphas) en Shapiro-Wilk y Kolmogorov-Smirnov fueron $p<0.5$ y para Anderson-Darling los valores críticos evaluados fueron para 0.01, 0.025, 0.05, 0.1 y 0.15, donde todos fueron significativos y por lo tanto ninguna variable cumplió la normalidad.

Después del manejo de los valores perdidos.

Posterior al tratamiento de los datos y a pesar de la mejora en curtosis y sesgos en ningún caso fue posible no rechazar la hipótesis nula. Concluimos lo mimso que antes del tratamiento, ninguna variable parece tener una distribución normal. Los puntos de corte evaludos (alphas) en Shapiro-Wilk y Kolmogorov-Smirnov fueron $p<0.05$ y para Anderson-Darling los valores críticos evaluados fueron para 0.01, 0.025, 0.05, 0.1 y 0.15, donde todos fueron significativos y por lo tanto ninguna variable cumplió la normalidad.

- ¿Qué variables no siguen una distribución normal?

Antes del manejo de los valores perdidos.

Todas

- co_gt
- pt08_s1_co
- nmhc_gt
- c6h6_gt
- pt08_s2_nmhc
- nox_gt
- pt08_s3_nox
- no2_gt
- pt08_s4_no2
- pt08_s5_o3
- t
- rh
- ah

Después del manejo de los valores perdidos.

Todas

- co_gt
- pt08_s1_co
- c6h6_gt
- pt08_s2_nmhc
- nox_gt
- pt08_s3_nox
- no2_gt
- pt08_s4_no2
- pt08_s5_o3
- t
- rh
- ah

Nota: recordemos que nmhc_gt fue eliminada por tener más del 20% de valores perdidos antes del tratamiento esta tampoco tenia una distribución normal.

- ¿Cómo interpretas los QQplots?

Antes del manejo de los valores perdidos.

En ningúno de los gráficos QQ se observa una distribución normal. Lo esperado para decir que se trata de una distribución normal es que los quantiles teoricos y los observados sean iguales o muy similares, gráficamente que siguan la linea roja. En todos los casos vemos que ninguna sigue este patrón y además se observa el ruido que meten los valores perdidos etiquetados en **-200**.

Se concluye que gráficamente ninguno cumple con la distribución normal.

Después del manejo de los valores perdidos.

Posterior al tratamiento hubo mejoría, sin embargo ninguna parece cumplir con la distribución normal. Obvervamos varios cambios, el primero es que efectivamente los patrones exóticos provocados por los valores -200 han desaparecido y que ahora se muestran tanto patrones de sesgo los puntos forman una curva y de colas pesadas que los puntos forman una S. Ningúno cumple con los criterios gráficos de normalidad.

- ¿Qué variables se desvían significativamente de la normalidad?

Antes del manejo de los valores perdidos.

Todas

- co_gt
- pt08_s1_co
- nmhc_gt
- c6h6_gt
- pt08_s2_nmhc
- nox_gt
- pt08_s3_nox
- no2_gt
- pt08_s4_no2
- pt08_s5_o3
- t
- rh
- ah

Después del manejo de los valores perdidos.

Todas

- co_gt
- pt08_s1_co
- c6h6_gt
- pt08_s2_nmhc
- nox_gt
- pt08_s3_nox
- no2_gt
- pt08_s4_no2
- pt08_s5_o3
- t

- rh
- ah

Nota: recordemos que nmhc_gt fue eliminada por tener más del 20% de valores perdidos antes del tratamiento esta tampoco tenia una distribución normal.

(C) Tratamiento de Datos Faltantes:

- ¿Qué estrategia utilizaste para manejar los datos faltantes? ¿Por qué elegiste esa estrategia?

Para columnas y filas totalmente vacías se decidió eliminarlas, fueron 2 columnas y 114 filas que se encontraban totalmente vacías. Por ser registros totalmente vacíos no existe razón de por que usar alguna imputación.

Las columnas que tenían más de 20% de datos perdidos se decidió eliminarlas ya que no se cuenta con datos suficientes para su imputación. Fue el caso de mmhc_gt.

Para las columnas numéricas se decidió por ForwardFill ya que eran datos de sensores a través del tiempo relacionados con contaminantes y este toma el último valor válido, los datos entaban en orden ascendente por fecha y hora. De usar una medida de tendencia central en este caso podíamos contaminar aún más los datos con una medición que no reflejaría la medición mas cercana. Es más probable que la medición previa esté mas cerca de la realidad temporal que la media de todo el dataset.

- ¿Cómo cambió el EDA después de la imputación de datos?

Hubo cambios notables en los estadísticos descriptivos por ejemplo los valores mínimos de pasar de -200 en todas las variables ya vemos los reales. La media se vio altamente impactada en algnos casos paso de 10.19 a -159.09, y posterior al tatamiento e imputación de los datos podemos decir que haora su valor es más cercano a la realidad observada.

- ¿Observaste diferencias significativas en las distribuciones de las variables?

Las categóricas siguieron un patrón uniforme antes y despúes del tratamiento de los datos, ya que no había datos perdidos.

En el caso de las numéricas si hubo un cambio importante, primero las distribuciones que tenian dendencia o parecian bimodales desaparecieron por la falta de los datos etiquetados con -200. Segúndo las distribuciónes mejorar a lepto y mesocurticas, pero en la mayoría de los casos continuaron siendo muy sesgadas. A peaar del tratmaiento, ninguna demostró seguir la distribución normal.

(D) Matriz de Correlación y Pairplot:

- ¿Qué relaciones lineales identificaste en la matriz de correlación y el pairplot?

Antes del manejo de los valores perdidos.

Matríz de correlación. Se seleccionó el metodo de spearman por que parece que nuestros datos no siguen una distribución normal. En la matríz d correlacion vemos varias variables con buenas correlaciones tomando en cuenta un umbral de $\geq |0.7|$ y que esto fue antes del tratamiento de datos. Las que tuvieron buenas correlaciones fueron: co_gt con nox_gt y no02_gt; pt08_s1_co con pt08_s5_03, pt08_s2_nmhc y c6h6_gt;c6h6_gt con pt08_s2_nmhc (perfecta), pt08_s4_no2 y pt08_s5_o3; pt08_s2_nmhc con pt08_s4_no2 y pt08_s5_o3; nox_gt con no2_gt; y, t con ah; todas fueron correlaciones positivas.

Nota: Recuerde antes del tratamiento de datos.

Pairplot Antes del tratamiento de los datos antes que notar relacionels lineales se observa como los valores -200 están metiendo ruido en nuestos datos y en gran parte de los casos generan dos grupos de mediciones, los correrctos reales y los datos perdidos etiquetados como -200.

Nota: Recuerde antes del tratamiento de datos.

Después del manejo de los valores perdidos.

Matríz de correlación. Se seleccionó el metodo de spearman por que parece que nuestros datos no siguen una distribución normal. Posterior al tratamiento de los datos las correlaciones cambiaron bastante, lo que cofnirma que los datos nulos marcados con -200 modificaban nuestros datos. Encontramos lo siguiente:

Buenas correalaciones positivas (tomando en cuenta un umbral de $\geq |0.7|$):

- co_gt con pt08_s1_co, c6h6_gt, pt08_s2_nmhc, nox_gt, no2_gt, pt08_s5_o3
- pt08_s1_co con c6h6_gt, pt08_s2_nmhc, pt08_s5_o3
- c6h6_gt con pt08_s2_nmhc, no2_gt, pt08_s4_no2, pt08_s5_o3 (perfecta)
- pt08_s2_nmhc con pt08_s4_no2, pt08_s5_o3
- nox_gt con no2_gt, pt08_s5_o3
- t con ah

Buaenass correlaciones negativas (tomando en cuenta un umbral de $\geq |0.7|$): - co_gt con pt08_s3_nox - pt08_s1_co con pt08_s3_nox - c6h6_gt con pt08_s3_nox - pt08_s2_nmhc con pt08_s3_nox - nox_gt con pt08_s3_nox - pt08_s3_nox con pt08_s5_o3

Pairplot Ahora logramos ver mejor la relación entre las variables ya que despues de retrirar los -200 que metían ruido a nuestro análisis las relaciones entre variables de expresan.

En el Pairplot, que es básicamente una representación gráfica de la matríz de correlación encontramos datos bastante similares que en la explicación pasada

de la matríz de correlación. Este tipo de graficas aportan un poco más de información ya que podemos ver no solo la correalación si no la fomra puede que la correlacion sea buena en el centro pero no en los extremos. Tambien podemos ver en estaas gráficas valores atípicos que modifican nuestras correlaciones.