Práctica 19:

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

Matrícula: 255813

Marzo 19, 2025

Reporte

(A) Análisis Exploratorio de Datos (EDA):

- ¿Qué información relevante obtuviste de los histogramas y gráficas de densidad (PDF)?

Antes del manejo de los valores perdidos.

De las tres variables numéricas encontramos que los years y los nodes no presentan una distribución normal, years es platicurtica y nodes leptocurtica con un sesgo muy alto a la izquierda.

Vemos que la variable age parece tiene cierta tendencia a la normalidad, pero por su curtosis posiblemente no, tiene un valor muy bajo de 0.5894.

Después del manejo de los valores perdidos.

Posterior al tratamiento de datos, eliminación de duplicados, no hubo cambio sifnificativo en los histogramas ni en las graficas PDF.

- ¿Alguna variable parece seguir una distribución normal?

Antes del manejo de los valores perdidos.

La que parece que pudiera tal vez ser normal es la variable de age, sin embargo, por curtosis es poco probable.

Después del manejo de los valores perdidos.

Posterior al tratamiento de datos, eliminación de duplicados, se obaserva lo mismo, la única que al menos gráficamente parece normal es age.

- ¿Qué insights obtuviste de las gráficas de barras para la variable categórica status?

Antes del manejo de los valores perdidos.

Que la variable categórica tiene dos clases, y que perdomina los que sobrevivieron 5 años o más. También vemos que existe en desvalanceo importante entre las clases. Además esta sería una buena variable para ser resultado o para ser predecida. Que es una clase binaria.

Después del manejo de los valores perdidos.

No hubo cambios significativos posterior al tratamiento de los datos.

- ¿Cómo se distribuyen los pacientes según su estado de supervivencia?

Antes del manejo de los valores perdidos.

Tiene una distribución desvanceada hacia la supervivencia.

Después del manejo de los valores perdidos.

Igual que antes del tratamiento de los datos: Tiene una distribución desvanceada hacia la supervivencia.

(B) Pruebas de Normalidad:

- ¿Qué variables no siguen una distribución normal según las pruebas de Shapiro-Wilk, Anderson-Darling y Kolmogorov-Smirnov?

Antes del manejo de los valores perdidos.

Las tres variables antes del tratamiento de los datos mostraron que se rechazaba la H0 y por lo tanto no siguen una distribución normal según las pruebas Shapiro-Wilk y Kolmogorov-Smirnov con un alpha de menor a 0.05. Sin embargo, para la prueba Anderson-Darling no fue posible rechazar la H0 a un nivel de 0.05. Técnicamente tenemos cierta avidencia de que age sigue una distribución normal, pro también tenemos dos pruebas que rechazan esto.

Después del manejo de los valores perdidos.

Posterior al tratamiento de datos se encontraron los mismos hallazgos, solo con tendencia a la normalidad en al variable age, con dos pruebas que rechazan su normalidad y Anderson-Darling que en puntos críticos de 0.05, 0.025 y 0.01 no pieden rechazar su normalidad.

- ¿Qué conclusiones puedes extraer de los QQplots?

Antes del manejo de los valores perdidos.

En la variable year podemos ver un patron S caracteristico de colas cargadas que podemos ver en la platicurtosis y en la variable nodes encontramos una parabola caracteristico de las distribuciones sesgadas como la que vemos en su histograma. Ambos patrones se observan en distribuciones no normales.

Sin embargo, en la variable age vemos que los puntos casi siguen perfectamente los valores teóricos por los osbservados solo en los extremos existe cierta desviación. En esta variable parece que hay tendencia a la normalidad, al igual que lo obtenido por las pruebas estadísticas.

Después del manejo de los valores perdidos.

No hubo cambios significativos posterior al tratamiento de los datos.

- ¿Qué variables tienen una distribución cercana a la normal?

Antes del manejo de los valores perdidos.

La variable age tiene evidencia de normalidad por QQ plot y Anderson Darling ($p<0.05$). Pero existe evidencia de que no sigue la distribución normal por Shapiro-Wilk y Kolmogorov-Smirnov. Así mismo por curtosis y sesgo queda indeterminado. Concluyo que podemos considerar que esta varibale tiende a la normalidad y si la medolodogía de recolección fue buena por la cantidad de datos podríamos tratarla como normal en algunas situaciones particulares, pero no podemos decir que tenemos evidencia contundente de su normalidad.

Después del manejo de los valores perdidos.

No hubo cambios significativos posterior al tratamiento de los datos.

(C) Tratamiento de Datos Faltantes:

- ¿Qué columnas tenían todos los valores faltantes?

No hubo columnas con datos faltantes.

- ¿Cómo manejaste estas columnas?

No hubo necesidad de manejarlas.

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

No hubo necesidad de hacer imputaciones.

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

No, al eliminar los datos duplicados no se observaron cambios significativos en las distribuciones de las variables.

(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.

Se usó spearman ya que las variables no tienen una distribución normal. En no se encuentra ningúna buena correlación en la matriz de correlación el mejor valor fue -0.1 lo cual es muy bajo o tiene una correlación negativa muy débil entre edad y nodulos.

En el gráfico pairplot vemos exactamente lo mismo un gráfico de puntos muy disperso que no muestra ninún patrón importante.

Después del manejo de los valores perdidos.

Posterior al tratamiento de los datos. Se usó spearman ya que las variables no tienen una distribución normal. En no se encuentra ningúna buena correlación en la matriz de correlación el mejor valor fue -0.1 lo cual es muy bajo o tiene una correlación negativa muy débil entre edad y nodulos.

En el gráfico pairplot vemos exactamente lo mismo un gráfico de puntos muy disperso que no muestra ninún patrón importante.