Proyecto:

SIN BAJAR LA GUARD.IA ~ CONSTRUYENDO CONOCIMIENTO AL SERVICIO DE LA SALUD

Una vez que se ha realizado la limpieza y preprocesamiento de los datos, es importante realizar una exploración detallada de los mismos, para comprender mejor su distribución y características. Esto puede incluir gráficos, tablas, estadística descriptiva y otros métodos de visualización y exploración. Todo esto pensando en un aprendizaje más magro (más puro) de los datos bajo estudio.

Los aspectos más relevantes sobre los que se busca investigar son:

Zonas de “acceso desigual”

Identificar y, eventualmente, priorizar las "zonas calientes" que presentan bajo acceso a servicios sanitarios. En relación a los accesos a los servicios de salud de alta complejidad en los lugares menos poblados, analizar si se puede calcular la cercanía a los establecimientos que brindan prestaciones de mayor complejidad. Este indicador busca ser una medida de accesibilidad física a dichos centros de salud y reflejar de algún modo que la cercanía a un hospital de alta complejidad implica un acceso a prestaciones de salud potencialmente mayor que la cercanía a una posta sanitaria o una “salita”, ya que, como podemos asumir, el tipo de emergencia atendida y la atención que pueden brindar estos establecimientos difieren notablemente.

Perfil sanitario para cada provincia (o en su defecto para cada región del país)

A través de los perfiles sanitarios, se pretenderá brindar un recurso para que los tomadores de decisiones y especialistas en salud, cuenten con información clave, concreta y resumida sobre la realidad sanitaria de cada distrito. Algunos aspectos que sería relevante contrastar con los datos del Censo Nacional son: tasa de natalidad, tasa de mortalidad (por causas), prevalencia de obesidad, hipertensión arterial, consumos de sustancias, entre otras…

1. Mapping de activos para la salud:

a. ¿Existe alguna relación entre la cantidad de establecimientos de salud y la densidad de médicos por especialidad en cada provincia?
b. ¿Hay diferencias significativas en la distribución de establecimientos de salud financiados públicamente y privadamente en todo el país?
c. ¿Se pueden identificar áreas geográficas con una escasez de establecimientos de salud en relación con la población atendida?
d. ¿Cuál es la relación entre la complejidad de los establecimientos de salud (con o sin internación) y la presencia de médicos especialistas en esas áreas?
e. ¿Cuáles son las especialidades médicas a nivel provincial?
f. ¿Existen especialidades médicas con una distribución desigual en comparación con otras en términos de ubicación geográfica de los médicos? (cantidad de profesionales por “N” habitantes)

2. Visualizaciones:

a. Usar gráficos para entender la distribución de la población, la cantidad de viviendas, y los recursos de salud en diferentes provincias y localidades.
b. Mapas Geoespaciales: representar los datos en mapas para visualizar la proximidad a los centros de atención y áreas con alta vulnerabilidad sanitaria

3. Feature Engineering:

Crear nuevas características que podrían ser útiles, como la relación nacimientos/defunciones, densidad de población, y distancia a centros de atención sanitaria.

4. Correlaciones:

Verificar mediante una matriz de correlación, la correlación entre las variables bajo estudio.

a. ¿Hay datos fuertemente correlacionados? ¿y si apuntamos a una variable específica, hay alguna correlación más marcada?
b. Analizar si se debe eliminar las features fuertemente correlacionadas (una de cada par), ya que mantener columnas altamente correlacionadas, puede ocasionar un comportamiento no deseado en los modelos de clasificación.

5. Encoding:

Pasar las variables categóricas (strings) a numéricas. Analizar diferentes métodos para elegir el más adecuado (One hot encoding, Label encoding, Getdummies).

6. Escalamiento de los datos:

Transformar las features para que tengan distribuciones más cercanas a la normal (elegir qué método es más conveniente: logaritmica, normalizar, estandarizar). En el caso de usar PCA, este paso debe realizarse posteriormente de aplicarlo, para que en las componentes PCA las variables sean “pesadas” de manera similar, y no tenga alto impacto la varianza de las columnas originales

Características que debe cumplir el entregable:

- Generar un dataset “limpio”, con todos los pasos aplicados, ya que será el que utilizaremos en el siguiente TP para los modelos de clasificación/clusterizacion. Concluir luego de ésta “limpieza” cuántos registros hemos mantenido/eliminado, con el fin de no quedarnos con muy pocos registros para avanzar más adelante con algún modelo de clasificación.
- Se debe ir desarrollando cada punto en la misma notebook donde se escriba el código. Dicho notebook debe contar con un índice, con sus diferentes apartados y el código debe ser fácil de leer, estar probado y comentado (esto último, en función de la necesidad).
- Se debe enviar el link directo del archivo .ipynb ó alternativamente subir el entregable a un repositorio GitHub mediante la integración con Google Colab. Recordar que al compartir el notebook, queden habilitados los permisos de edición, para poder dejar comentarios/correcciones.
- Tener en cuenta que si bien, pueden realizar diversos análisis y visualizaciones, se debe dejar en el entregable sólo aquello que sea relevante.
- Luego de cada análisis es importante poder obtener una conclusión de lo observado y/o breve interpretación de los resultados.

Repositorio:

https://github.com/NoeliaFerrero/Proyecto_MentoriaFAMAF_2024

Deadline: 29/07