Practica 2: Tipologia y ciclo de vida de los datos
UOC

Álvaro Monforte Marín

January 13, 2023

Contents

1 Pregunta 1

Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El dataset usado en esta practica se encuentra alojado en esta pagina web proporcionado en el documento de la practica. A su vez su origen y pertenencia se debe a la Universidad de California Irvine es a un repositorio que se puede encontrar aqui pagina web del repositorio original.

Este dataset nos presenta 14 atributos, entre los cuales tenemos el catorceavo (output) como label objetivo para estudiar como los otros 13 a priori pueden relacionarse o no entre si con tal de obtener una estimacion acerca de si la persona tiene mas riesgo o no de padecer un ataque cardiaco. Ademas posee 303 pacientes implicados (las instancias).

Atributo	Tipo de dato	Informacion ofrecida
Age	int64	Edad del paciente
Sex	int64	Sexo del paciente
cp	int64	Dolor de pecho $\in [0, 4]$
trestbps	int64	Presion de sangre en reposo
chol	int64	Serum de colesterol
fbs	int64	Azucar en sangre (ayunas)
restecg	int64	Electrocardiograma en reposo
thalach	int64	Maximo ratio de pulsaciones
exang	int64	Angina inducida por ejercicio
oldpeak	float64	Depresión del ST ind. ej. rel.
slp	int64	Pendiente del pico de ejercicio de ST
caa	int64	Numero de tejidos coloreados
thall*	int64	”Thallium Stress Test”
output**	int64	Estrechamiento del diámetro

- '*' : Nos informa sobre la condicion del flujo de la sangre en ese momento. Es un parametro medico bastante complejo.
- '**': Nos habla de la probabilidad de padecer un ataque cardiaco. Es el label.

Las muertes por causas isquemicas en el corazon superan los 8 millones de personas anuales. Luego es de alta importancia estudiar cualquier dataset que permita un entendimiento temprano de este tipo de padecimientos. Fuente

;

---

2 Pregunta 2

Integración y selección de los datos de interés a analizar. Puede ser el resultado de adicionar diferentes datasets o una subselección útil de los datos originales, en base al objetivo que se quiera conseguir.

Respeto a la integracion y seleccion, creo que todos los datos son sumamente importantes. Quizas ya que tengo en cuenta la presion en sangre, ademas del dano en los tejidos, voy a obviar el azucar en sangre al menos en este estudio. Vease cierta informacion que ayuda a entender cuan correlados puenden estar estos eventos.

Luego prosigo de inicio con las 303 instancias y ahora 13 atributos a investigar.

3 Pregunta 3

Limpieza de los datos.

- 3.1. ¿Los datos contienen ceros o elementos vacíos?
  Gestiona cada uno de estos casos.

  El codigo se ha realizado en lenguaje Python, ayudandome de las libreria Pandas en general aunque se ha hecho tambien uso de la libreria Scipy para algunos cometidos de forma principal, ademas de sklearn y pingouin para algunos calculos estadisticos.

  Los datos actualmente no contienen elementos vacios pero si ceros. Esto se ha logrado mediante la ejecucion de la suma de todos los posibles Nan dentro de un dataframe en el que se encuentra la seleccion de todos los datos. Respecto a los ceros, hay que tener en cuenta que algunos datos en su forma original eran categoricos sin orden (taggeados) finitos 0 o 1; Sin embargo, otros eran datos de tipo likert. Hay alguna excepcion de categoricos sin orden de tipo finito por clase, que ayudaran a establecer diferencias respecto a un grupo control con valor 0. Luego los elementos nulos o cero no han sido tratados con dinamicas especiales. Se podria haber tambien estipulado si existia ciertas correlaciones a la hora de eliminar NaN pero se ha procedido de tal manera que se ha eliminado la instancia completa perdiendo ciertos datos que podrian ser de interes.

- 3.2. Identifica y gestiona los valores extremos

  Para la identificacion de valores extremos se podria haber procedido por un metodo grafico Q-Q que ademas de ayudar a la eliminacion de ciertos outliers tambien nos podria haber ayudado a entender la normalidad de los datos. Pero se ha hecho uso de un analisis numerico eliminando aquellos en cuantiles superiores al 75% o al 25%, que tambien seria visible en un diagrama de Box-plot. Con ello, aunque perdiendo informacion (no siempre los outliers son inutiles), pero como estudio temprano esta bien ya que podemos centralizar los datos.

Tras la realizacion de la limpieza se ha obtenido un total de 229 instancias.

4 Pregunta 4

Análisis de los datos.

- 4.1. Selección de los grupos de datos que se quieren analizar/comparar (p. ej., si se van a comparar grupos de datos, ¿cuáles son estos grupos y qué tipo de análisis se van a aplicar?)
- 4.2. Comprobación de la normalidad y homogeneidad de la varianza.
- 4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

La seleccion de los grupos de interes se puede realizar entorno a las variables categoricas finitas sin orden, para asi formalizar subgrupos. En este caso le doy cierto peso a la variable exang que determina si hay angina inducida por ejercicio, el por que de su importancia, tenemos varios papers (estudios) que presentan a esta sintomatologia o padecimiento como algo relevante. Lease el articulo [1]. Por ello puedo comparar exang contra trtbs, chol y thalachh. Donde tanto como trtbs como chol son normales mientras que thalachh no se distribuye de forma normal. Luego esto seria un indicio para no estudiar de forma parametrica a exang contra thalachh si hago una comparacion de medias (donde ahora comentare la homocedasticidad) o al menos usar un test que sea robusto como el de Levene.

Una vez se vea si se prestan a una correlacion entre ellas, estaria bien contrastar como se comportan respecto al target de la variable output, ya que nos queremos adentrar en el estudio de la varianza explicada por estos de la variable a predecir output.

Para la normalidad se hizo uso del test de D'Agostino, este tiene como parametros cruciales la curtosis y la asimetria, hay que tener en cuenta que al descartar outliers centralizando los datos respecto a la mediana podemos provocar una ayuda en la mejor de la curtosis ya que este mide la 'densidad' de puntos respecto al primer momento de inercia, luego al centralizarse a priori deberia haber una mayor densidad respecto al momento de giro.

Para la homocedasticidad se hace uso de Bartlett cuando haya normalidad y del test de Levene cuando no la haya, el test de Levene ademas tiene la propiedad de gran robustez en el momento de que no se presente normalidad.

Para la correlacion de las variables, tenemos que tener en cuenta el grupo de supuestos que se cumple, en este caso cuando sean normales sea hace uso de la correlacion de pearson, cuando no sea asi se hara uso de la de spearman que es no parametrica.

Para el test de diferencia de medias t-test, debemos tener en cuenta la homocedasticidad de los datos, cuando esta no se cumple se debe anadir una correcion de levene

a los grados de libertad de las variables para generar la nueva hipotesis de que sean homocedasticas y asi ejecutarlo de tal manera que lo sean, en el algoritmo de python, gracias al modulo pingouin ya esta implementada esta correccion pero se debe senalar en mi codigo con un flag.

Para el estudio de la regresion lineal, las variables deben ser independientes cosa que se puede contrastrar con la correlacion que existan entre ellas en un mapa de calor. Tambien se debe cumplir que haya una distribucion normal de los errores (cosa que normalmente en el campo medico se puede asumir). Para ayudarme a realizar la regresion lineal con el modulo de sklearn me he ayudado de la pagina web cienciadedatos que lo explica detalladamente con todo lujo de detalles.

5 Pregunta 5

Representación de los resultados a partir de tablas y gráficas. Este apartado se puede responder a lo largo de la práctica, sin necesidad de concentrar todas las representaciones en este punto de la práctica.

6 Pregunta 6

Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

7 Pregunta 7

Código. Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

8 Pregunta 8

Vídeo. Realizar un breve vídeo explicativo de la práctica (máximo 10 minutos), donde ambos integrantes del equipo expliquen con sus propias palabras el desarrollo de la práctica, basándose en las preguntas del enunciado para justificar y explicar el código desarrollado. Este vídeo se deberá entregar a través de un enlace al Google Drive de la UOC (https://drive.google.com/. . .), junto con enlace al repositorio Git entregado.

References

[1] Alon Eisen, Deepak L. Bhatt, P. Gabriel Steg, Kim A. Eagle, Shinya Goto, Jianping Guo, Sidney C. Smith, E. Magnus Ohman, Benjamin M. Scirica, and null null. Angina and future cardiovascular events in stable patients with coronary artery disease: Insights from the reduction of atherothrombosis for continued health (reach) registry. *Journal of the American Heart Association*, 5(10):e004080, 2016. doi: 10.1161/JAHA.116.004080. URL https://www.ahajournals.org/doi/abs/10.1161/JAHA.116.004080.