Ciencia de datos aplicada al estudio de la Obesidad y

otras enfermedades crónicas en Córdoba

# Práctico N°2

# Análisis y curación

Grupo 2

➢ Fernando Apaza

➢ Edgardo Garrigo

➢ Nicolas E Ponce

**Objetivo y alcance**

En esta segunda aproximación (que suele ir de la mano con el análisis exploratorio y visualización) se pretende empezar a transformar el dataset para los primeros experimentos de aprendedores en las siguientes materias.

En este laboratorio el primer gran objetivo va a ser lograr un único dataset, que condense toda la información importante para el posterior entrenamiento y aprendizaje que permitirá conocer si algunas de las variables incluidas en el mismo tienen influencia sobre el desarrollo de obesidad.

Se deberá explorar la base de datos y detectar variables erróneas, o información repetida, para que, luego de haber obtenido el dataset deseado, guardarlo en un archivo nuevo.

**Método**

Para este laboratorio pueden usar tanto Python como R, y además de generar las variables nuevas que puedan ser necesarias más adelante, hay que limpiar los datos realizando un tratamiento similar de los datos al [laboratorio 1 dictada en las clases de A y C] (https://github.com/DiploDatos/AnalisisYCuracion/blob/master/notebooks/Limpieza.ipynb)

Recuerden que la base para este práctico es diferente! descargarla en este [link](https://github.com/Mentoria-obesidad/base-obesidad-segundo-practico).

**Lista de tareas**

1) Verificar los tipos de datos detectados por pandas. Asegurar tipo de datos correctos para las variables que son continuas y aquellas que son categóricas. En el caso de no coincidir se deberán modificar la naturaleza de esas variables.

2) Verificar y asegurar que no existan datos duplicados.

3) Verificar que no existan caracteres "raros" para los datos de los features "ubic" y "cod\_encuestador". Tip: Pueden hacerlo importando la librería ftfy.badness

4) Codificar las categorías de la variable "sexo", "ubic" y "nes" para su posterior uso en los próximos prácticos.

5) Verificar espacios y normalizar los nombres de al menos 10 features que crean convenientes para un mejor entendimiento de su significado. Por ej: "riesgo" cambiarla por "cat\_cintura", entre otros que crean convenientes y vayan a ser usados por ustedes.

6) Supongamos que necesitamos compartir públicamente todo el dataset y que los features "cod\_encuestador" y "dni" son sensibles y confidenciales. Elijan y apliquen dos técnicas de despersonalización para solucionar este aspecto.

7) Analizar y normalizar los datos faltantes (nulos).

En este punto encontrarán que hay muchos features con datos faltantes, les pedimos que elijan los 3 features con mayor cantidad de registros nulos, para los cuales se espera un análisis sobre las posibles causas de cada uno (calculando la proporción que significan sobre todo el dataset y buscando posibles patrones respecto a otros features relacionados). Si tienen dudas antes de aplicarles un tratamiento a cada uno nos pueden consultar, tenemos a una experta en imputación en el grupo! :)

8) Aplicar reglas de integridad sobre los datos, validando y normalizando al menos las siguientes:

a) Existen sujetos que en la variable "cancer" pusieron que no, pero en "ubic" hay un número mayor al 1? o que haya un dato faltante en "cancer" pero exista un número mayor al 1 en "ubic"?

b) Con respecto a los meses y estaciones del año, todos coinciden? hay algún mes o estación en la que no se hayan realizado encuestas?

9) Outliers: Durante el Análisis que realizaron en nuestro Práctico anterior, descubrieron que existen valores atípicos aplicar una solución para normalizar solo esos Outliers. Recuerden que pueden consultarnos lo que haga falta!

10) Guardar el dataset resultante con un nuevo nombre, para trabajarlo más adelante.

**Se evaluarán los siguientes aspectos:**

1- Que se apliquen los conceptos vistos con los profes en el teórico y en el práctico.

2- Que el entregable no sea solo la notebook. El informe debe tener un mensaje claro y debe presentarse en un formato legible para cualquier tipo de stakeholder. Además, se deberá entregar el dataset trabajado.

3- Capacidad de Análisis.

4- Criterio para elegir qué solución aplicar en cada caso y con qué método implementarla.

Deadline tentativo para la entrega 19/07/2020

**Respuestas**

* Tarea 1

Se convierten varias variables a distintos formatos: categoricas, numericas y string, según corresponde.

Las dos estrategias a utilizar:

-ds.stress.astype(int)

-ds.stress=ds.stress.apply(pd.to\_numeric)

Nota: Las variables actih, actch, y senth están en formato datetime64 y deberían ser numéricas float. Tratamos de aplicar diferentes estrategias para pasarlo a numérico, pero ninguna resultó.

* Tarea 2

Se detectan 6 pares de datos duplicados, evidenciado por el dni de los entrevistados. Se los elimina del dataset.

Existe una persona que fue entrevistada en dos ocasiones diferentes, julio y noviembre. No se eliminan los datos.

Se detectan dos features identicos, “dni” y “dni\_”, se procede a eliminar la columna correspondiente a “dni\_”

* Tarea 3

Se verifica que no existen caracteres raros en las dos variables evaluadas.

* Tarea 4

Se codifican las categorías de las variables "sexo", "ubic" y "nes"

* Tarea 5

Se verifica que no existen espacios en los nombres de los variables y se renombran los siguientes

'ec':'ecar',

'ic':'icar',

'cant' : 'cant\_cig',

'antepa' : 'ante\_ecar\_pad',

'antema' : 'ante\_ecar\_mad',

'anteher' : 'ante\_ecar\_her',

'Tiempo' : 'tiempo\_dbt',

'antepad' : 'ante\_dbt\_pad',

'antemad' : 'ante\_dbt\_mad',

'anteher' : 'ante\_dbt\_her',

'antehij' : 'ante\_dbt\_hij',

'charla' : 'charla\_dbt',

'prog' : 'prog\_dbt',

'ubic' : 'tum\_ubic',

'nombre' : 'actfis\_nombre',

'act': 'actfis\_nivel',

'tipo' : 'actfis\_tipo',

'durac' : 'actfis\_durac',

'veces' : 'actfis\_veces',

'carga' : 'actfis\_carga',

'riesgo' : 'riesgo\_ecar',

"tecar":"ecar\_trata",

"tipo1":"dbt\_tipo1",

"tipo2":"dbt\_tipo2",

"tratadbt": "dbt\_trata"

* Tarea 6

Se realizó la despersonalización de dos features con información sensible a identificación mediantes dos algoritmos hash: md5 y el blake2b. El último se escogió para guardar los datos en base de datos, en las columnas dni y cod\_encuestador.

* Tarea 7

A continuación se muestran los 3 features con mayor cantidad de datos faltantes.

grupoedad 4298 datos faltantes

fgins1 1614 datos faltantes

actfis\_carga 195 datos faltantes

En el feature grupoedad todos los campos están vacíos. Es una variable que no aparece en el codificador. Es posible que no haya estado en la planilla de encuesta y pero por alguna razón apareció en el ds. por lo que se elimina de la base de datos

El 62.5% de los valores de fgins1 están completos. fgins1 hace referencia a "ingesta de grasas insaturadas según frecuencia (bajo,adecuado, alto)". Es posible que los encuestados no tengan conocimiento sobre qué es una grasa insaturada por lo que les resulta difícil definir su frecuencia de ingesta.

Solo 42 encuestados no contestaron las variables fgins1 & fgins2. Es curioso y difícil de comprender cómo el entrevistado puede no haber contestado fgins1 y luego si contestar fgins2. ¿ error en la carga de datos?

Realizando una inspección de la tabla en donde actfis\_carga está vacío, se observa que la variables fgins1 también está vacía en las primeras 5 filas. ¿coincidencia? Lo evaluamos:

empleando el dataset 'dsz' todas las encuestas donde actfis\_carga está vacío, también está vacía la variable fgins1. No parece haber una razón lógica, pero definitivamente es un patrón.

Además, las otras variables (sentmin, actmh, actimin, actih, actch, actid, actmmin y actmd), que figuran como las más incompletas, también están relacionadas con actifis. Esto es otro patrón que gira entorno a actfis.

En principio no imputaremos ningún dato. Tendremos presente que existen y cuales son los valores faltantes, y evaluaremos su tratamiento de acuerdo al análisis que deseemos hacer en el futuro.

El valor al lado de cada uno representa el porcentaje de datos que tienen cada uno en su columna

fgins1: 62.5 %

actfis\_carga: 95.5 %

sentmin: 96.2 %

* Tarea 8

**A) Conclusión**: de los 9 casos, 7 se explican fácilmente: Las personas declararon la ubicación de sus tumores benignos. La variable tum\_ubic no especifica si es tumor maligno o benigno. Decidimos no descartar la encuesta, si tendremos en cuenta esto para más adelante.

Sin embargo, hay dos registros anómalos:

1. La n\_encuesta 2887, el cual declara no tener cáncer, pero Sí un tumor maligno en mamas y recibir tratamiento en instituto privado (tratinst = 2.00). Es muy probable que el entrevistado se equivocó en el casillero de cáncer, porque los otros datos parecen indicar que realmente tiene cáncer maligno. No se descarta la encuesta, si tendremos en cuenta esto para más adelante.
2. La n\_encuesta 555, el cual declara no tener cáncer, ni tumor maligno, benigno o recibir tratamiento alguno, pero Sí un tumor en cuello de útero. Es probable que hubiera una confusión en el entrevistado y este declaró cuello uterino erróneamente. También es posible que alguna vez tuviera un tumor en el cuello de útero y que fue extirpado en el pasado. La variable tum\_ubica no aclara si se debe contestar sobre tumores del presente o del pasado. No se descarta la encuesta, si tendremos en cuenta esto para más adelante.

Con respecto a los datos faltantes, 2 personas no completaron el casillero cáncer, pero declararon tener un tumor benigno y su ubicación. No se descarta la encuesta, si tendremos en cuenta esto para más adelante.

Por otro lado, la encuesta 606 no declaró nada relacionado a cáncer, como tampoco nada sobre diabetes. El resto de la encuesta está completa. No se descarta la encuesta.

**B) Conclusión:** Antes de hacer el análisis, dos puntos a considerar:

1. Todo el análisis será bajo la premisa que la encuesta fue realizada en Argentina, de manera de relacionar los meses con las estaciones locales.
2. Los meses pueden, aparentemente, no coincidir con la estación ya que cambian el dia 21 de un mes. Por lo que existe la posibilidad de declarar 2 estaciones en un mismo mes dependiendo si es a principio o fin de mes.

Podemos observar en los resultados anteriores que se declararon 9 meses y solo 3 estaciones. Los datos reflejan que no hubo encuestas en otoño, y concuerda con el hecho que no aparecen los meses de abril mayo o junio. Tener en cuenta que aparece marzo, pero los primeros 21 días son verano.

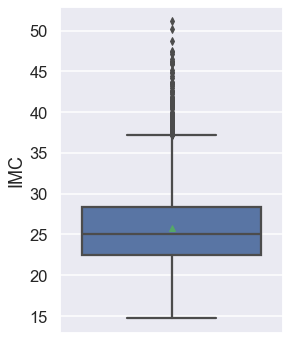
Por otro lado, existen 83 meses que no coinciden con su estación del año:

* hay 62 entradas que no coincide la estación Verano.
* hay 1 entrada que no coincide estación Invierno.
* hay 20 entradas que no coincide estación Primavera.

En base a los resultados creemos que la variable "est" debe ser eliminada del dataset, ya que la variable mes es más específica sobre la realización de la encuesta. También creemos que habría sido buena variable incluir fecha de la encuesta.

* Tarea 9

Las variables seleccionadas en el práctico pasado fueron: sexo, IMC, eent, actfis y edad. Sólo IMC y edad son variables continuas.



Para la variable IMC, obtenemos los estadisticos:

count 4292.00

mean 25.82

std 4.78

min 14.76

25% 22.49

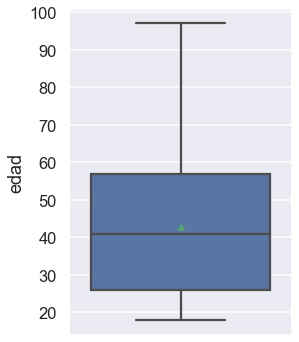
50% 25.10

75% 28.38

max 51.11

Figura 1: Boxplot de IMC donde se pueden observar los outliers por arriba del bigote superior (IMC mayor a 37.86)

En la figura 1 se puede observar el boxplot para la variable IMC. Los outliers detectados fueron 96 y la función reconoce a los índice mayores a 37.86 como outliers, siendo que la media está en 25.82. Se decide no eliminar los outliers de IMC ya que representa a la población con índice de obesidad grado II y III. Serán importantes para nuestros estudios futuros.



Para la variable edad, los estadísticos calculados son los siguientes

count 4292.00

mean 42.66

std 17.95

min 18.00

25% 26.00

50% 41.00

75% 57.00

max 97.00

Figura 2: Boxplot de edad donde se pueden observar que no existen outliers

En la figura 2 se puede observar el boxplot para la variable edad. No se detectan outliers y, siendo la media de 42.66 años.

* Tarea 10

Se guarda la nueva base de datos se llama ‘Nueva\_Base\_Obesidad\_CURADA.csv’.