Introducción

El periódico 20 minutos el 12 de diciembre 2017 abría su sección de salud con una notica muy impactante:

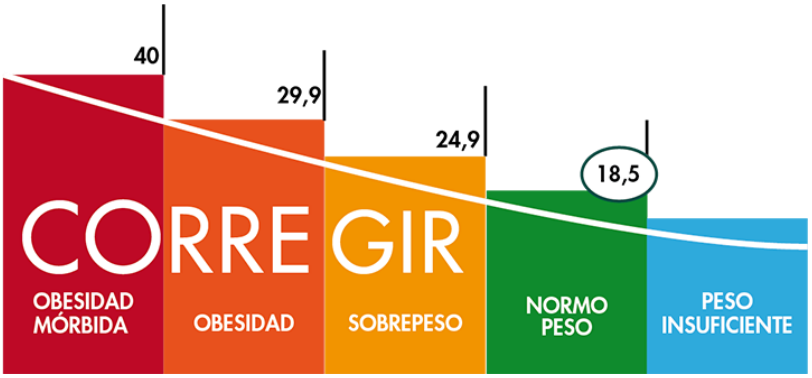
- "La tasa de obesidad se ha duplicado en España en las últimas dos décadas."
- "Un 53% de la población está por encima de su peso."
- "En el caso de los niños, el 12% padece sobrepeso y el 14% obesidad."

https://www.20minutos.es/noticia/3210348/0/tasa-obesidad-duplica-espana-ultimas-dos-decadas/#xtor=AD-15&xts=467263

En el trabajo se analizará el peso de la población mundial, el crecimiento de casos que ha habido en Reino Unido, si tiene alguna relación el peso con la situación económica del país, predecir el sobrepeso con diferentes variables y también la diabetes que va totalmente ligada al sobrepeso.

El trabajo girara sobre IMC o BMI, que es el índice de masa corporal adoptado por la OMS utilizado para el diagnóstico del sobrepeso y de la obesidad. Se puede calcular fácilmente el IMC a partir de dos datos simples: altura y peso. Formula y tabla:

IMC = peso (kg) ÷ altura² (metros)



1 Imc_World

En el primer notebook python se unen dos ficheros, uno de ellos tenemos todos los IMCs mundiales y en el segundo la longitud y latitud de los países, formaremos un tercer datasaet "imc_world_map.xlsx" que usaremos en Tableau.

Dataset usados:

- World_Map.xls
- imc_mundiales.xlsx

Datos obtenidos en:

- https://www.wikipedia.org

Sergio Muñoz Elvira

1.2IMC_mundiales

Con la herramienta Tableau se analiza el dataset "imc_world_map.xslx". En el se puede ver IMC medio en cada país tanto el del hombre como de la mujer.

Dataset usados:

- imc_world_map.xlsx

Datos obtenidos en:

- 1_Imc_world.ipynb

2 Estudio_UK

El fichero recopila los ingresos en hospitales de Reino unido con problemas de peso entre 2006 - 2007. Hay 3 tipos de tablas, ingresos directamente por peso, ingresos y uno de los problemas es el peso y el último los que acaban con cirugía bariátrica.

Dataset usados:

- data_uk.xls

Datos obtenidos en:

- http://data.gov.uk/dataset/statistics_on_obesity_physical_activity_and_diet_england/

Conclusiones:

Aunque la recopilación de los datos ha mejorado en los últimos años se podrá apreciar que cada vez hay mas casos en hombres y mujeres independientemente de la edad que tengan.

De diagnóstico primario:

- El número de casos masculino es mucho más alto que el de mujeres independientemente del año.
- La franja edad más propensa a ingresar hospitalizado por diagnóstico primario es entre 45 y 54 años.

De diagnóstico primario y secundario:

- El número de casos masculino es mucho más alto que el de mujeres independientemente del año.
- La franja edad más propensa a ingresar hospitalizado por problemas de peso es entre 25 y 24 años y de 55 a 74 años, siendo la última franja la que más casos tiene.

Ingresos hospitalarios con cirugía bariátrica:

- El número de casos masculino es mucho más alto que el de mujeres independientemente del año algo que se ha repetido en los casos anteriores.
- La franja edad más propensa a ingresar hospitalizado que acaban en cirugía bariátrica es entre 45 y 54.

3 IMC_PIB

El objetivo es unir los 3 ficheros que contienen información diferente, el primero imc mundial, el segundo el PIB de las poblaciones y el tercero el numero de habitantes para obtener un nuevo dataset que estudiaremos en Tableau. También haremos un modelo KNeighborsClassifier.

Dataset usados:

- bmi_mundial_wikipedia.xlsx
- pib_mundial.xls
- poblacion_mundial.xls

Datos obtenidos en:

- https://www.wikipedia.org/
- https://datacatalog.worldbank.org/dataset/population-ranking
- https://datacatalog.worldbank.org/dataset/gdp-ranking

Conclusiones:

Tras realizar un contar de IMC de todos los países los resultados son:

Insuficiencia_Ponderal →0
Normal → 55
Sobrepeso → 112
Obesidad_I → 4

Hay un alto número de países donde la media de su población sufre sobrepeso.

También hay un contar del PIB por población donde los resultados fueron:

Nivel_1 → 121
Nivel_2 2 →3
Nivel_3 → 4
Nivel_4 → 20
Nivel_5 →6

Donde podemos apreciar que son muy pocos países los que tienen la riqueza mundial y muchos los carecen de recursos siendo la población mundial muy desigual.

Para comprobar la relación entre PIB/población contra BMI medio use el método KNeighborsClassifier, pero para obtener un scoring alto necesitamos crear muchos grupos diferentes sin obtener gran precisión con menos vecinos.

3.2 IMC_vs_pib

Con la herramienta Tableau se analiza el dataset "imc_vs_pib.xlsx". En el que comparará y visualizará IMC medio y el PIB/población mundial.

Dataset usados:

- imc_vs_pib.xlsx

Datos obtenidos en:

- 3_imc_pib.ipynb

Conclusiones:

Se representa gráfica y visualmente el IMC medio y PIB/población mundial y aunque no es lineal en un alto número países con problemas económicos como son los países de Latinoamérica, África y países asiáticos tienen el IMC más bajo que viene ligado a los problemas de falta de alimento de dichos países.

4 Limpieza_Dataset1

Notebook de Python de limpieza del documento principal y se creará un nuevo fichero Excel "Dataset_limpio1" sin NaN y con las columnas que se necesitan.

Dataset usados:

- Dataset1.xls

Datos obtenidos en:

- https://figshare.com/articles/Body_Mass_Index_Is_Better_than_Other_Anthropometric_Indices_for_Identifying_Dyslipidemia_in_Chinese_Children_with_Obesity/3109123

5 Modelización

En este caso se usa RStudio, el objetivo es encontrar las variables mas fuertes que luego se implementaran en modelos de machine learning .

Dataset usados:

- Dataset_limpio1.xlsx

Datos obtenidos en:

- 4_Limpieza_Dataset1.ipynb

Conclusiones:

Las variables para IMC más fuertes son peso y altura, sobre todo esta última, ya que son las variables con las que se calcula .

HIP y WC también son bastantes significativas ya que son medidas del cuerpo, cuanto mas grandes el IMC crece.

En cambio, si lo que buscamos son variables fuertes para un alimento como por ejemplo “Meat” las más significativas son de la misma rama.

Pasa lo mismo si cogemos como referencia el colesterol, el resto de las variables del mismo grupo son las más significativas.

Todas las variables se podrían juntar en 3 grupos, medidas del cuerpo, variables de sangre y alimentación.

6 Machine_learning_WC_HIP

Con el notebook de python implementando herramientas de machine learning predeciremos si una persona sobrepeso o no con las variables “WC” y “HIP.

Métodos:

- Decision Tree
- Random Forest
- Logistic Regression
- K nearest neighbors
- Bagging

Dataset usados:

- Dataset_limpio1.xlsx

Datos obtenidos en:

- 4_Limpieza_Dataset1.ipynb

Conclusiones:

El “accuracy” obtenido en todos los modelos están por encima 90% es debido a “WC” y “HIP” son dos variables que cuanto mas altas sean el IMC va también va a serlo, con ellas es muy fácil obtener con precisión si una persona sufre sobrepeso o no.

7 Machine_learning_alimentación

Continuando con el notebook de python implementando herramientas de machine learning predeciremos si una persona sufre insuficiencia ponderal, peso normal, sobrepeso, obesidad I u obesidad II con las variables “Fruit”,”Vegetable” y “Meat”

Métodos:

- Random Forest
- Logistic Regression
- K nearest neighbors
- Bagging

Dataset usados:

- Dataset_limpio1.xlsx

Datos obtenidos en:

- 4_Limpieza_Dataset1.ipynb

Conclusiones:

El “accuracy” obtenido en todos los modelos están por encima 60% que podría ser una cifra significativa para predecir el estado de peso de una persona, pero al hacer pruebas y cambiando extremos la cantidad de alimentos el modelo solo te muestra dos resultados , insuficiencia ponderal y peso normal, esto es debido a que el dataset el 90% de las personas se encuentran en este estado ya que el estudio es realizado en china donde la población sufre una gran desnutrición. Para que el modelo fuese mucho mas fuerte y los resultados mas cercanos a la realidad necesitaría más muestras de personas que sufran sobrepeso, obesidad I y obesidad II.

8 Machine_learning_sangre

Continuando con el notebook de python implementando herramientas de machine learning predeciremos si una persona sufre insuficiencia ponderal, peso normal, sobrepeso, obesidad I u obesidad II con las variables “LDL”,”TG” y “HDL” además también se intenta predecir el IMC exacto.

Métodos:

- Decision Tree
- Random Forest
- Logistic Regression
- K nearest neighbors
- Bagging

Dataset usados:

- Dataset_limpio1.xlsx

Datos obtenidos en:

- 4_Limpieza_Dataset1.ipynb

Conclusiones:

Predecir estado:

El “accuracy” obtenido en todos los modelos están por encima 60% pero como pasa en el caso anterior solo te muestra dos resultados , insuficiencia ponderal y peso normal, es debido a que el fichero el 90% de los datos se encuentran en este estado.

Predecir BMI:

El “accuracy” obtenido en los modelos ronda el 10%, podríamos descartarlo, ya que obtener el IMC correcto es casi imposible.

9 Machine_learning_diabetes

Uno de los mayores problemas del sobrepeso viene ligado con la diabetes, con un notebook de Python y modelos de machine learning se predice la posibilidad de ser diabático con diferentes parámetros de salud.

Métodos:

- Decision Tree
- Random Forest
- Logistic Regression
- Bagging

Dataset usados:

- diabetes.csv

Datos obtenidos en:

- https://www.kaggle.com/johndasilva/diabetes

Conclusiones:

Todos los modelos tienen un “accuracy” por encima del 78%, el porcentaje de precisión es bastante alto con todas las variables.