

Vamos a empezar a pensar como un científico de datos.

Para ello, en esta primera hoja de trabajo vamos a trabajar los tres niveles de extracción de información de los datos.

Para ello te voy a proponer que uses algún de estos datasets:

Te propongo los siguientes datasets:

- <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>
- <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction>
- <https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists>
- <https://www.kaggle.com/datasets/benroshan/factors-affecting-campus-placement>
- <https://www.kaggle.com/datasets/ayessa/salary-prediction-classification>

Elige alguno de ellos o usa tus propios datos.

De uno de los datasets vamos a plantear lo siguiente:

- El problema y objetivo que se pretende con el dataset
- Con los conocimientos que tienes ahora trata de definir qué vas a hacer en cada etapa de la pirámide:
 - descriptiva
 - inferencial
 - reconocimiento de patrones

Vamos a ello:

1. Define el problema y el objetivo que pretendes con el dataset

Describe con palabras sencillas el problema y el objetivo que pretendes con tu dataset.

Se quiere conocer si una serie de hábitos de vida y de condicionantes personales y de salud están relacionadas con el hecho de sufrir o no un ictus.

El objetivo principal es reconocer patrones y predecir el ictus a partir de una serie de hábitos de vida y condicionantes personales y de salud. También se va a comparar los hábitos de vida y condicionantes personales y de salud en función del género, la edad y de si han sufrido o no un ictus.

2. Suponiendo que el dataset está limpio explica los pasos que vas a seguir en las tres etapas:

ETAPA DE ANÁLISIS DE DATOS	¿QUÉ VAS A HACER EN CADA ETAPA? Escribe lo que se te ocurra sin demasiados filtros.
Descriptivo.	Describir las variables en función del género, en función de la edad (categorizada en grupos) y

El objetivo es lograr explicar los datos con gráficos y estadísticos descriptivos que nos permitan encontrar tendencias y empezar a sacar valor de los datos.	de si han sufrido o no un ictus mediante estadísticos descriptivos y gráficos en función de la naturaleza de las variables.
Inferencial. Recuerda que aquí puedes comparar, puedes relacionar variables o incluso crear modelos entradas - salida que nos permiten obtener patrones clave de los datos.	Calcular los diferentes pruebas estadísticas para valorar si las relaciones tienen o no una significación estadística. Y, analizar mediante un modelo de regresión logística qué variables pueden predecir el ictus.
Reconocimiento de patrones. Puedes solucionar problemas de: - Clasificación - Regresión - Clustering - agrupado de datos - Reglas de asociación - Reducir dimensiones - Predicción de series temporales	Mediante Machine Learning, entrenar un modelo con algoritmos supervisados de clasificación para predecir el ictus.

Este pequeño ejercicio es el que hago mentalmente cuando empiezo un nuevo proyecto.

No en todos los proyectos es necesario aplicar las tres etapas pero sí que es importante pensar hasta dónde puedes llegar de los datos.

Es un excelente ejercicio para empezar.

¡Vas por buen camino!

Seguimos avanzando.

1 abrazo y hasta pronto.
Brindo por tus éxitos :)

Jordi Ollé

Informació d'atributs

- 1) id: identificador únic
- 2) gènere: "Home", "Dona" o "Altres"
- 3) edat: edat del pacient
- 4) hipertensió: 0 si el pacient no té hipertensió, 1 si el pacient té hipertensió
- 5) heart_disease: 0 si el pacient no té cap malaltia cardíaca, 1 si el pacient té una malaltia cardíaca
- 6) ever_married: "No" o "Yes"

7) work_type: "children", "Govt_jov", "Never_worked", "Privat" o "Autònom"

8) Tipus de residència: "Rural" o "Urbà"

9) Nivell_de_glucosa mitjà: nivell mitjà de glucosa a la sang

10) IMC: índex de massa corporal

11) smoking_status: "abans fumat", "mai fumat", "fuma" o "Desconegut"*

12) ictus: 1 si el pacient va tenir un ictus o 0 si no

*Nota: "Desconegut" a smoking_status significa que la informació no està disponible per a aquest pacient