

PRÁCTICA 2: LIMPIEZA Y ANÁLISIS DE DATOS

Reinaldo Quintero

Irene Díaz

Descripción del data set

Elección del data set Data set

Para la realización de la práctica, hemos escogido el data set de "[Heart Disease UCI](#)", disponible en [Kaggle](#). La información de este data set se recopiló de distintas fuentes: Cleveland Clinical Foundation, Instituto de Cardiología de Hungría, V.A Centro Médico y la Universidad de Zurich.

Motivo de la elección

Solo en 2012, murieron más de 17 millones de personas en el mundo por enfermedades cardiovasculares, según datos oficiales de la Organización Mundial de la Salud, convirtiéndose en una de las principales causas de den función a nivel mundial. Las enfermedades cardiovasculares son un conjunto del trastorno del corazón y los vasos sanguíneos.

Visto la relevancia con la que cuentas las enfermedades cardiovasculares, disponer de datos que expliquen la naturaleza de estas enfermedades es de máxima relevancia. Pero tenemos que convertir esos datos en información para que se les consiga dar una utilidad real.

Hemos escogido este data set porque se centra en el estudio de las enfermedades cardiovasculares y queremos comprobar cómo la aplicación de las técnicas de análisis de datos puede ser útiles a la hora de identificar y lidiar con enfermedades de este tipo. Buscaremos entonces identificar qué características son delatadoras de las enfermedades cardiovasculares (presión arterial, edad, sexo, niveles de tensión, colesterol...)

Gracias al análisis de datos, información veraz, recopilada de fuentes fiables, con datos una cantidad ingente de datos reales, puede arrojar conclusiones concisas que nos ayuden en el día a día en un campo tan complejo e importante como el de la medicina.

Descripción del data set

Aunque el data set original cuenta con 75 atributos, en Kaggle se encuentra disponible una versión filtrada del mismo con solo 14 variables. Son las siguientes:

- age: edad en años (continua)
- sex: 1 = hombre, 0 = mujer (categórica)

- cp: (chest pain) tipo de dolor de pecho: 1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptotic (categórica)
- trestbps: (Resting Blood Pressure) presión arterial en reposo (in mm Hg a la hora de ser ingresado en el hospital) (continua)
- chol: (Serum Cholestrol) colesterol en mg/dL (continua)
- fbs: (Fasting Blood Sugar) compara el azúcar en sangre estando en ayunas > 120 mg/dl (1 = true, 0 = false) (categórica)
- restecg: resultados del electrocardiograma en reposo: 0 = normal; 1 = havingST-T wave abnormality; 2 = left ventricular hyperthrophy (categórica)
- thalach: (Max heart rate achieved) ritmo cardiaco (continua)
- exang: (Exercise induced angina) angina (1 = yes; 0 = no) (categórica)
- oldpeak: depresión del segmento ST inducida por el ejercicio relativo al reposo (continua)
- slope: (Peak exercise ST segment) pendiente del máximo del ejercicio del segmento ST: 1 = upsloping; 2 = flat; 3 = downsloping (categórica)
- ca: número mayor de vasos (continua)
- thal: 3 = normal; 6 = fixed defect; 7 = reversable defect (categórica)
- goal: si el individuo sufre enfermedad cardiovasclar o no: 0 = no; 1,2,3,4 = sí (categórica)

Contribuciones

Contribuciones	Firma
Investigación previa	RQ; ID
Redacción de las respuestas	RQ; ID
Desarrollo código	RQ; ID