

Universidad Católica "Nuestra señora de la Asunción"
Facultad de ciencias y tecnología
Complementos de Informática



Preparación y limpieza de los datos extraídos.

Entrega 3 - Data mining

Alumno: Alain Vega

Profesor: Wilfrido Felix Inchausti Martínez

19 de Noviembre del 2023

Índice

1. Preparacion de los datos	2
1.1. Tratamiento de missing values	2
1.2. Tratamiento de ruidos	2
1.3. Tratamiento de anomalias	2
1.4. Manejo de variables categoricas	2
1.5. Normalizacion	2
1.6. Nuevas variables	2
2. Diccionario de datos finales a utilizar	2

1. Preparacion de los datos

- 1.1. Tratamiento de missing values
- 1.2. Tratamiento de ruidos
- 1.3. Tratamiento de anomalias
- 1.4. Manejo de variables categoricas
- 1.5. Normalizacion
- 1.6. Nuevas variables

2. Diccionario de datos finales a utilizar

Los datos a utilizar en el proyecto sera un archivo .csv (*Comma Separated Values*) llamado *heart.csv* que se obtuvo de la plataforma web *Kaggle* [2] el cual de inicio contiene 12 columnas, las cuales:

- 1. **Age**. edad del paciente
- 2. **Sex**. sexo del paciente, donde:
 - M: Masculino
 - F: Femenino
- 3. **ChestPain**. tipo de dolor en el pecho, el cual puede ser:
 - TA: *Typical Angina*, es decir angina tipica
 - ATA: *Atypical Angina*
 - NAP: *Non-Anginal Pain*
 - ASY: *Asymptomatic*

La angina de pecho es un tipo de dolor de pecho causado por la reduccion del flujo sanguineo al corazon. El dolor a menudo se describe como un dolor constrictivo, presión, pesadez, opresión o dolor en el pecho. El paciente siente como si tuviera un gran peso apoyado en el pecho. [5]

4. ***RestingBP***. presión arterial en reposo (sistólica) medido en mililitros de mercurio [mmHg]

La presión arterial es una medida de la fuerza que utiliza el corazón para bombear sangre por el cuerpo. Se mide en milímetros de mercurio [mmHg] y se expresa en 2 cifras:

- presión sistólica: la presión cuando el corazón expulsa la sangre
- presión diastólica: la presión cuando el corazón descansa entre latidos

Por ejemplo, una presión arterial de “140 sobre 90” o 140/90 mmHg, significa una presión sistólica de 140 mmHg y una presión diastólica de 90 mmHg. [6]

5. ***Cholesterol***. Colesterol serico medido en miligramos por decilitro de sangre [mm/dl]

El colesterol es una sustancia grasa (un lípido) presente en todas las células del organismo. Los niveles de colesterol en sangre, que indican la cantidad de lípidos o grasas presentes en la sangre, se expresan en miligramos por decilitro [mg/dl] La sangre lleva el colesterol a las células en partículas transportadoras especiales denominadas «lipoproteínas». Dos de las lipoproteínas más importantes son:

- lipoproteína de baja densidad (LDL) - tambien conocida como colesterol malo
- lipoproteína de alta densidad (HDL) - tambien conocida como colesterol bueno

El colesterol total (serico) en sangre es la suma del colesterol transportado en las partículas de LDL, HDL y otras lipoproteínas. [4]

6. ***FastingBS***. azúcar en sangre (glucosa) en ayunas medido en miligramos por decilitro [mg/dl], puede ser:

- 1: Si $FastingBS > 120$ [mg/dl]
- 0: Caso contrario

7. ***RestingECG***. resultados del electrocardiograma en reposo, puede ser:

- **Normal:** normal

Indica que no se observaron anomalías significativas en el electrocardiograma. Las ondas y complejos están dentro de los rangos normales.

- **ST:** tener anomalía de la onda ST-T (inversiones de la onda T y/o elevación o depresión del ST $> 0,05$ mV)

El segmento ST es la sección plana e isoeletrica del ECG entre el final de la onda S (el punto J) y el comienzo de la onda T. El segmento ST representa el intervalo entre la despolarización y la repolarización ventricular. [7]

- **LVH:** muestra probable o definitiva hipertrofia ventricular izquierda según los criterios de Estes

Se refiere a un aumento en el tamaño de las fibras miocárdicas en la cámara de bombeo cardíaca principal. Esto sugiere un agrandamiento anormal del músculo del ventrículo izquierdo del corazón. [1]

Un ECG de diagnóstico en reposo (electrocardiograma) registra la actividad eléctrica del corazón mientras está en reposo. Proporciona información sobre su frecuencia y ritmo cardíaco y también puede mostrar si hay agrandamiento del corazón o evidencia de un ataque cardíaco previo. [3]

8. **MaxHR.** frecuencia cardíaca máxima alcanzada medido en pulsaciones por minuto [ppm]

9. **ExerciseAngina.** angina inducida por el ejercicio, puede ser:

- **Y:** Si
- **N:** No

10. **OldPeak.** pico antiguo = ST [Valor numérico medido en depresión]

11. **ST_Slope.** pendiente del segmento ST durante un ejercicio físico máximo en una prueba de esfuerzo cardíaco. Puede ser:

- **Up:** ascendente
- **Flat:** plano

- **Down:** descendente

12. ***HeartDisease***. sufrio de insuficiencia cardiaca

- **1:** insuficiencia cardiaca
- **0:** normal

El *dataset* dispone de 918 muestras en total para ser analizadas.

Referencias

- [1] MD Ary L Goldbeguer. *Left ventricular hypertrophy: Clinical findings and ECG diagnosis*. <https://www.uptodate.com/contents/left-ventricular-hypertrophy-clinical-findings-and-ecg-diagnosis>. 2022.
- [2] fedesoriano. *Heart Failure Prediction Dataset*. <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction>. 2021.
- [3] ascot cardiology group. *Diagnostic Resting ECG*. <https://ascotcardiologygroup.co.nz/services/diagnostic-resting-ecg/>.
- [4] The Texas Heart Institute. *Colesterol*. <https://www.texasheart.org/heart-health/heart-information-center/topics/cholesterol/>.
- [5] MayoClinic. *Angina de pecho*. <https://www.mayoclinic.org/es/diseases-conditions/angina/symptoms-causes/syc-20369373>. 2022.
- [6] NHS. *What is blood pressure?* <https://www.nhs.uk/common-health-questions/lifestyle/what-is-blood-pressure/>. 2022.
- [7] Ed Burns y Robert Buttner. *The ST Segment*. <https://litfl.com/st-segment-ecg-library/>. 2022.