

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



Análisis exploratorio y selección de los datos a utilizar para los modelos.

Entrega 2 - Data mining

Alumno: Alain Vega

Profesor: Wilfrido Felix Inchausti Martínez

16 de Octubre del 2023

Índice

1. Datos a utilizar	2
2. Analisis exploratorio	4
2.1. Missing values	5
2.2. Ruidos y anomalias	5
2.2.1. Ruidos	6
2.2.2. Anomalias	6
2.3. Redundancia	6
2.4. Correlacion	6
3. Conclusiones	6

1. Datos 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 salguineo 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. [6]

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. [7]

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. [8]
- **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.

2. Analisis exploratorio

En esta seccion se realizara un analisis exploratorio al *dataset* con la plataforma Knime, con el fin de cubrir:

2.1. Missing values

Un valor faltante puede significar varias cosas diferentes. Quizás el campo no era aplicable, el evento no ocurrió o los datos no estaban disponibles. Podría ser que la persona que ingresó los datos no conocía el valor correcto o no le importaba si un campo no estaba completado. [9]

El *dataset* no posee valores faltantes en ningún registro, como lo muestra la figura 1

Name	Type	# Missing values
Age	Number (integer)	0
Sex	String	0
ChestPainType	String	0
RestingBP	Number (integer)	0
Cholesterol	Number (integer)	0
FastingBS	Number (integer)	0
RestingECG	String	0
MaxHR	Number (integer)	0
ExerciseAngina	String	0
Oldpeak	Number (double)	0
ST_Slope	String	0
HeartDisease	Number (integer)	0

Figura 1: Missing values en el dataset heart.csv

2.2. Ruidos y anomalías

Un ruido es un dato o un conjunto de datos que agregan información sin sentido a la muestra. [5]

Un valor atípico o dato anómalo (*outlier*, en inglés) es una observación que numéricamente es muy distinta al resto de elementos de una muestra. [11]

2.2.1. Ruidos

En la columna *RestingBP* se encontro un registro cuyo valor es 0, cosa que es imposible.

En la columna *Cholesterol* se encontro un registro cuyo valor es 0, cosa que es imposible.

2.2.2. Anomalias

En la columna *Cholesterol* se encontraron varios registros cuyos valores son mayores a 400.

2.3. Redundancia

La redundancia a nivel de datos en un conjunto de datos hace referencia a que 2 o mas datos presentan la misma informacion.

Durante la exploracion no se encontraron valores o campos redundantes misma informacion

2.4. Correlacion

La correlación mide la fuerza de la relación entre dos variables. Proporciona una idea de cómo se relacionan las variables y cómo se afectan entre sí. [10]

3. Conclusiones

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. *Cholesterol*. <https://www.texasheart.org/heart-health/heart-information-center/topics/cholesterol/>.
- [5] javatpoint. *What is Noise in Data Mining?* <https://www.javatpoint.com/what-is-noise-in-data-mining>.
- [6] MayoClinic. *Angina de pecho*. <https://www.mayoclinic.org/es/diseases-conditions/angina/symptoms-causes/syc-20369373>. 2022.
- [7] NHS. *What is blood pressure?* <https://www.nhs.uk/common-health-questions/lifestyle/what-is-blood-pressure/>. 2022.
- [8] Ed Burns y Robert Buttner. *The ST Segment*. <https://litfl.com/st-segment-ecg-library/>. 2022.
- [9] Minewiskan y TimShererWithAquent. *Missing Values (Analysis Services - Data Mining)*. <https://learn.microsoft.com/en-us/analysis-services/data-mining/missing-values-analysis-services-data-mining?view=asallproducts-allversions>. 2022.
- [10] Utkarsh. *Association and Correlation in Data Mining*. <https://www.scaler.com/topics/association-and-correlation-in-data-mining/>. 2023.
- [11] Victor Yepes. *¿Qué hacemos con los valores atípicos (outliers)?* <https://victoryepes.blogs.upv.es/2022/02/21/que-hacemos-con-los-valores-atipicos-outliers/>. 2022.