Universidad Catolica "Nuestra señora de la Asuncion"
Facultad de ciencias y tecnologia
Complementos de Informatica

Preparación y limpieza de los datos extraídos.

Entrega 3 - Data mining

Alumno: Alain Vega
Profesor: Wilfrido Felix Inchaustti Martínez

20 de Noviembre del 2023

Índice

1. Preparacion de los datos

El dataset, asi como se expuso en la fase 2, no presenta muchas dificultades en cuanto a la calidad de los datos, esto resalta su confiabilidad para la construccion posterior de un modelo que ayude a la deteccion de posibles insuficiencias cardiacas.

1.1. Tratamiento de 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]

Como se expuso en el documento de la entrega 2, en el *dataset heart.csv* no se encuentran *missing values* por lo cual el tratamiento es no hacer nada.

1.2. Tratamiento de ruidos

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

En las columnas *RestingBP* y *Cholesterol* se detectaron ruidos. La estrategia empleada de abordaje consiste en reemplazar los registros ruidosos con el promedio de valores libre de ruido.

1.3. Tratamiento de anomalias

Un valor atípico o dato anomali (*outlier*, en inglés) es una observación que numéricamente es muy distinta al resto de elementos de una muestra. [10]

En la columna *Cholesterol* se detectaron anomalias, como se expuso en la entrega 2. La estrategia empleada de abordaje consiste en reemplazar los registros anomalos con el promedio de valores libre de anomalias.

1.4. Manejo de variables categoricas

Las columnas *Sex*, *ChestPain*, *RestingECG*, *ExerciseAngina* y *ST_Slope* eran de tipo string que correspondian a diferentes categorias de cada correspondiente concepto.

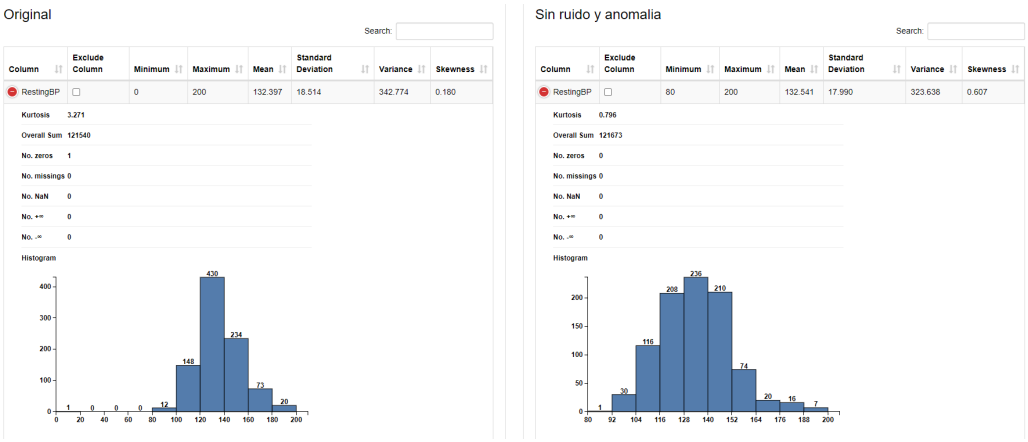

Figura 1: Comparativa de la distribucion del RestingBP (presion arterial), antes despues del tratamiento de rudios y anomalias

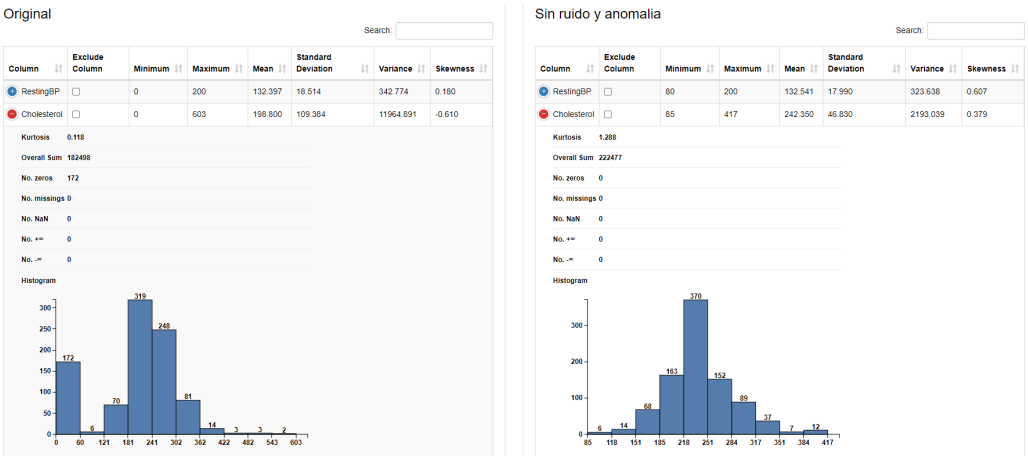

Figura 2: Comparativa de la distribucion del Cholesterol, antes despues del tratamiento de rudios y anomalias

1.5. Normalizacion

No se aplico ninguna normalizacion al *dataset* debido a la buena calidad del conjunto de datos.

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 luego de la preparacion y limpieza cuenta con: 12 columnas, las cuales:

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

   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:

   - **0**: normal

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

   - **1**: 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 isoeléctrica 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]

- **2**: 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:

   - **1**: Si
   - **0**: 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:

    - **1**: ascendente
    - **0**: plano
    - **-1**: descendente

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

    - **si**: insuficiencia cardiaca
    - **no**: normal

El *dataset* dispone de 918 muestras en total para ser analizadas. La figura 3 muestra el workflow correspondiente.

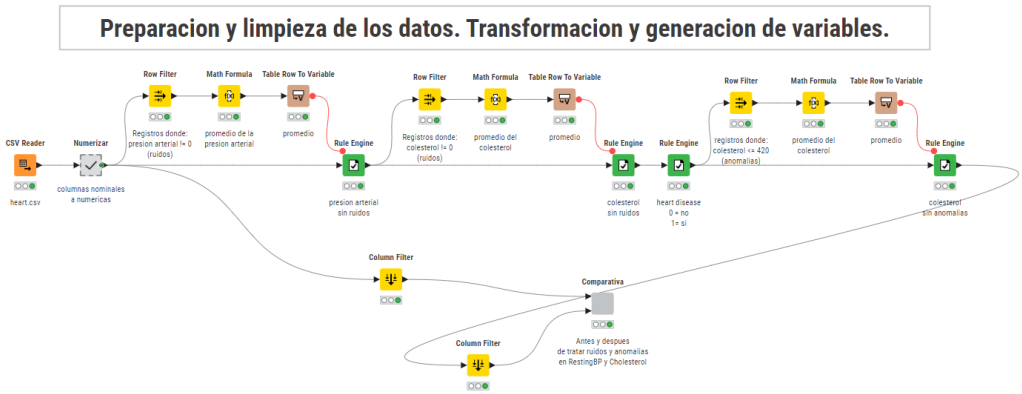

Figura 3: Worflow en knime

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/colesterol/.

[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] 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.