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

Construccion del modelo para obtencion de los resultados

Entrega 4 - Data mining

Alumno: Alain Vega
Profesor: Wilfrido Felix Inchaustti Martínez

20 de Noviembre del 2023

Índice

1. Caso de estudio: Insuficiencia cardiaca

1.1. Problema

Las enfermedades cardiovasculares (ECV) son la principal causa de muerte a nivel mundial, cobrando la vida de aproximadamente 17.9 millones de personas cada año, lo que representa el 31 % de todas las muertes en todo el mundo. Cuatro de cada 5 muertes por ECV son causadas por ataques cardíacos y accidentes cerebrovasculares, y un tercio de estas muertes ocurren prematuramente en personas menores de 70 años.

La insuficiencia cardíaca es un evento común causado por las ECV. Las personas con enfermedades cardiovasculares o que tienen un alto riesgo cardiovascular (debido a la presencia de uno o más factores de riesgo como hipertensión, diabetes, hiperlipidemia o enfermedades ya establecidas) necesitan una detección temprana y gestión, en la cual un modelo de aprendizaje automático puede ser de gran ayuda. [2]

1.2. Conjunto de datos

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 son:

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 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

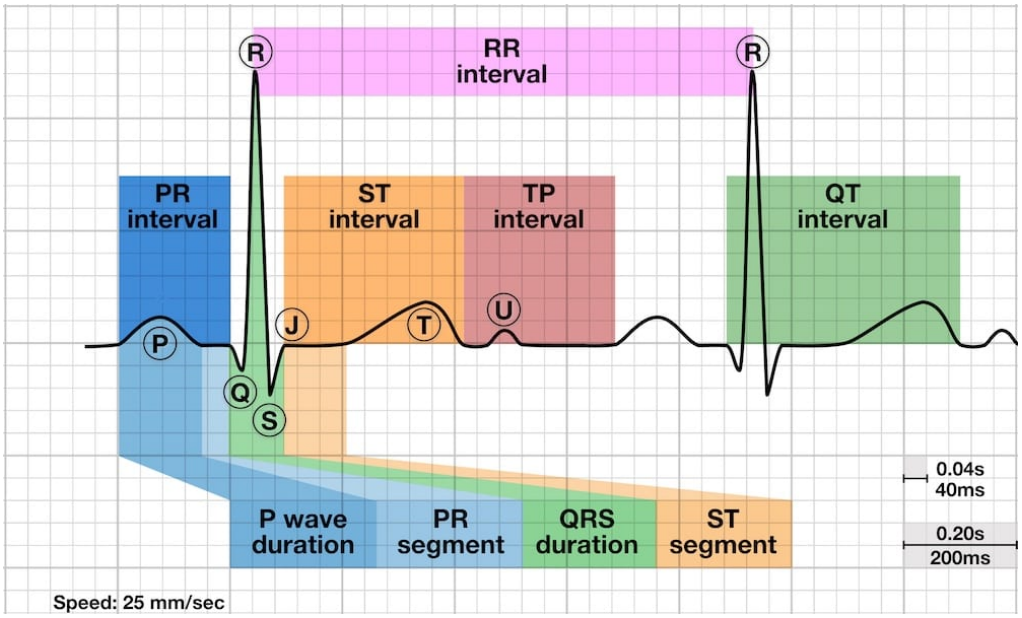

Figura 1: ECG de un corazon [8]

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

   - **Normal**: normal

     Indica que no se observaron anormalidades 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 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]

   - **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

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 ningun registro, como lo muestra la figura 2

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 2: Missing values en el dataset heart.csv

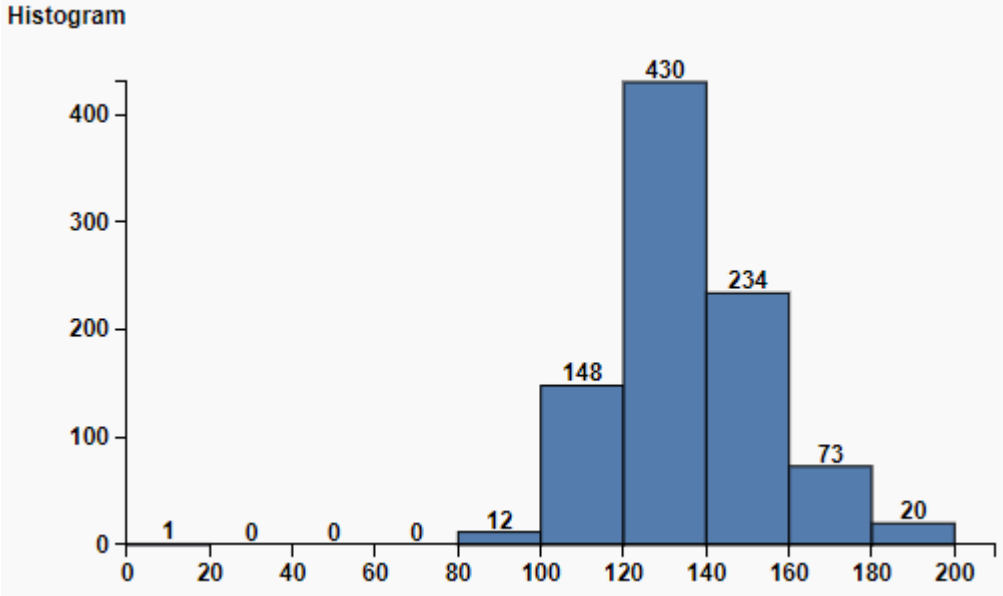

Figura 3: Histograma de la presion arterial

2.2. Ruidos y anomalias

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

Un valor atípico o dato anomali (*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. Se puede apreciar en la figura 3

En la columna *Cholesterol* se encontraron una cantidad de 172 registros cuyos valores son 0, cosa que es imposible. Se puede apreciar en la figura 4

2.2.2. Anomalias

En la columna *Cholesterol* se encontraron 8 registros cuyos valores son mayores a 420, donde se los considera anomalias. Se puede apreciar en la figura 4

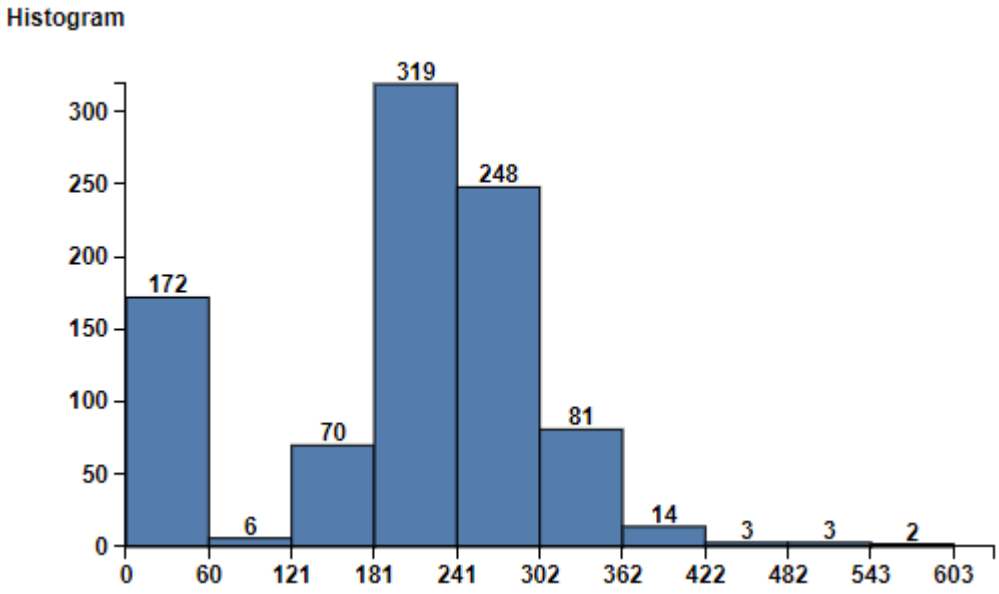

Figura 4: Histograma del colesterol

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, esto se visualizara mejor a continuacion.

2.3.1. 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]

Las correlaciones de datos se pueden visualizar en la figura 5.

La correlacion mas fuerte en el conjunto de datos se da entre las columnas: *heartDisease, ST_Slope* con valor de -59.19 %. Como el maximo valor no llega siquiera al 60 % se deduce que no se disponen de datos redundantes.

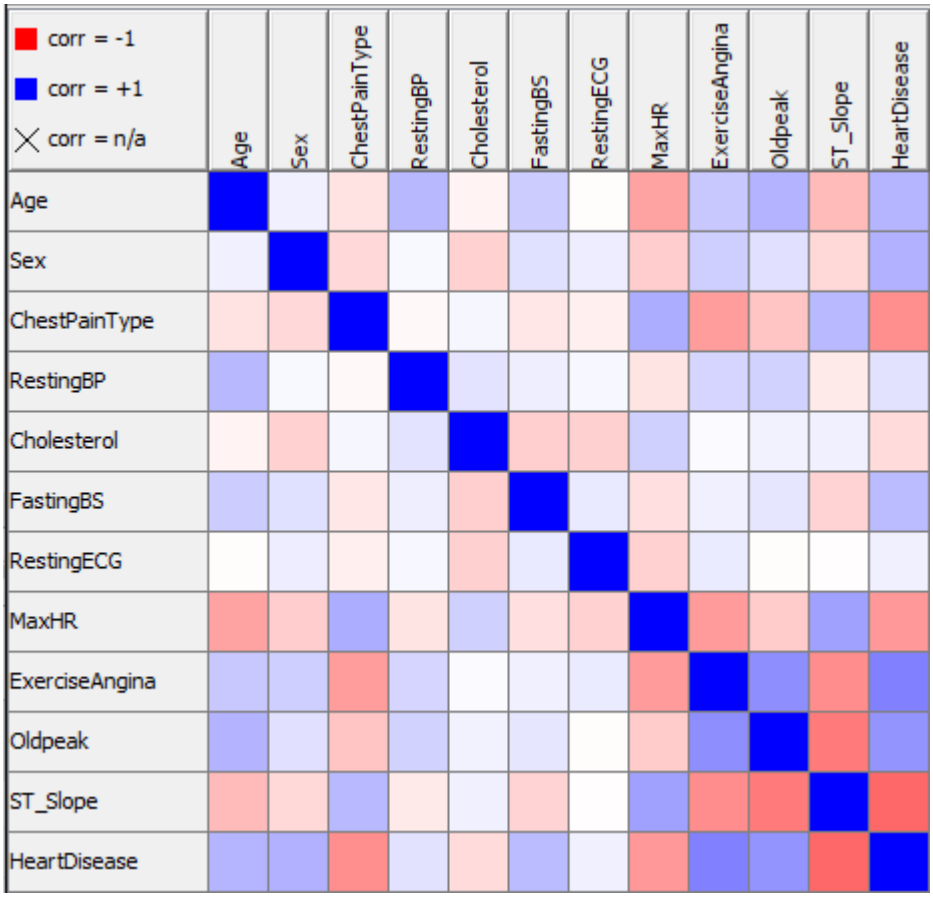

Figura 5: Matriz de correlacion en el dataset heart.csv

3. 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.

3.1. 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.

3.2. 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.

3.3. 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. [11]

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.

3.4. 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.

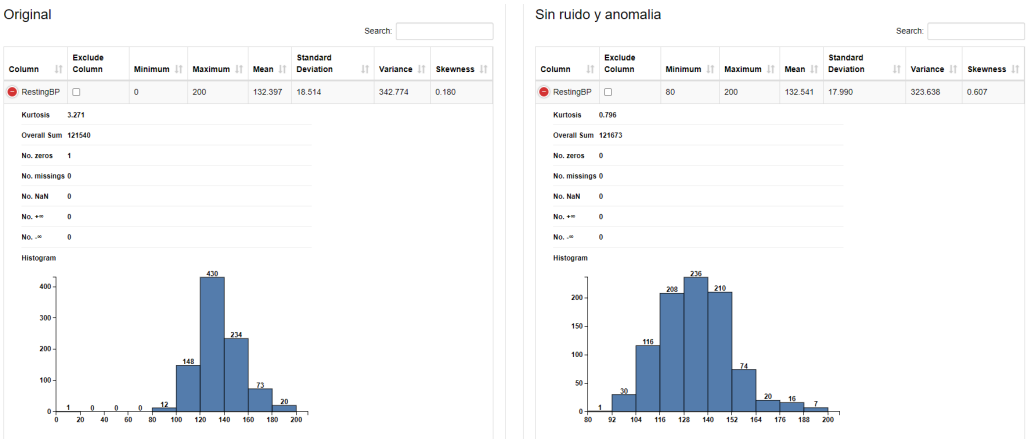

Figura 6: Comparativa de la distribucion del RestingBP (presion arterial), antes despues del tratamiento de rudios y anomalias

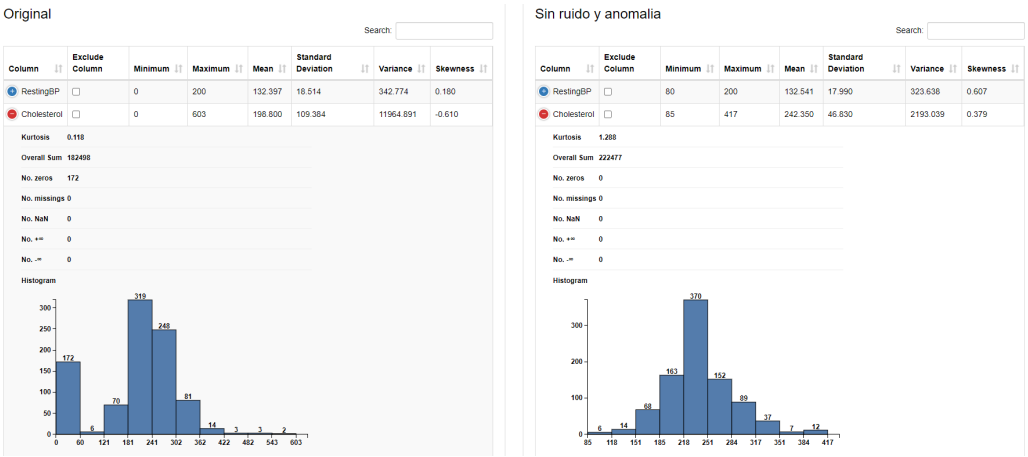

Figura 7: Comparativa de la distribucion del Cholesterol, antes despues del tratamiento de rudios y anomalias

3.5. Normalizacion

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

4. 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 8 muestra el workflow correspondiente.

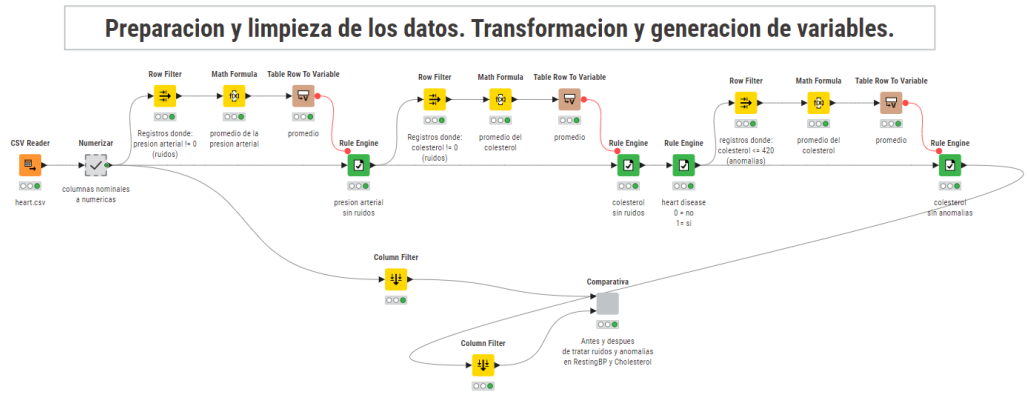

Figura 8: Workflow de la prepacacion y limpieza de datos, en knime.

5. Construccion del modelo

Para la construccion del modelo vamos a utilizar el conjunto de datos ya procesados en la etapa previa. Con motivo de mejorar la confiabilidad del modelo, vamos a implementar varias tecnicas para cumplir con nuestra tarea: predecir cuando una persona puede sufrir o no de insuficiencia cardiaca.

Para lo cual, la siguiente tecnica de validacion fue elegida: **Cross Validation** con 5 particiones del conjunto de datos. Ya que el conjunto de pruebas de la validacion cruzada es significativamente mayor a que simplemente realizar un tipico *train/test split* con 70 - 30. Se compararon las siguientes tecnicas:

- **Decision tree**
- **Logistic regression**
- **Random forest**
- **Gradient boosting (trees)**
- **Naive bayes**
- **Fuzzy logic**
- **Neural network**
- **Support vector machine (SVM)**
- **K nearest neighbor**

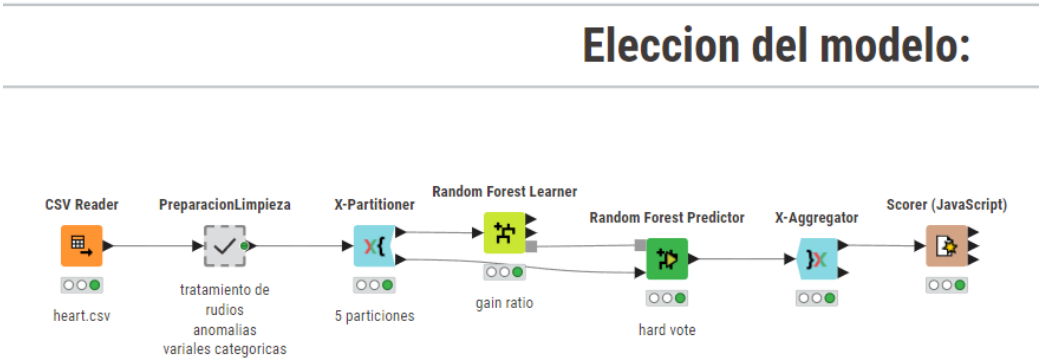

Figura 9: Workflow de la prepacacion y limpieza de datos, en knime.

las figuras 10, 11 y 12 muestran la comparativa entre estas tecnicas utilizando la validacion cruzada.

La tecnica elegia fue el random forest ya que obtiene mejores estadisticas en general en la serie de pruebas que se ejecutaron. Tambien y mas imporante, es la tenica que minimiza los falsos negativos, para este particular problema un falso negativo puede costar vidas.

Por lo tanto una vista simplificada del workflow lo muestra la figura 9

6. Conclusiones

En resumen, en este proyecto de data mining desarrollamos un modelo de predicción para la insuficiencia cardiaca, haciendo uso de técnicas convencionales de análisis de datos. La exploración detallada de conjuntos de datos ha permitido identificar patrones relevantes y relaciones entre variables, respaldando la construcción de modelos de machine learning con resultados aceptables.

Aunque el modelo ha demostrado ser prometedor en la predicción de la insuficiencia cardiaca, se reconoce la necesidad de una validación continua.

Arbol de decision

Confusion Matrix

	no (Predicted)	si (Predicted)	
no (Actual)	318	92	77.56%
si (Actual)	90	418	82.28%
	77.94%	81.96%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (ð□□¹)	Correctly Classified	Incorrectly Classified
80.17%	19.83%	0.599	736	182

Regresion logistica

Confusion Matrix

	no (Predicted)	si (Predicted)	
no (Actual)	334	76	81.46%
si (Actual)	62	446	87.80%
	84.34%	85.44%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (ð□□¹)	Correctly Classified	Incorrectly Classified
84.97%	15.03%	0.695	780	138

Random forest

Confusion Matrix

	no (Predicted)	si (Predicted)	
no (Actual)	335	75	81.71%
si (Actual)	44	464	91.34%
	88.39%	86.09%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (ð□□¹)	Correctly Classified	Incorrectly Classified
87.04%	12.96%	0.736	799	119

Figura 10:

Gradient boosting (tree)

Confusion Matrix

	no (Predicted)	si (Predicted)	
no (Actual)	340	70	82.93%
si (Actual)	69	439	86.42%
	83.13%	86.25%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (ð□□¹)	Correctly Classified	Incorrectly Classified
84.86%	15.14%	0.694	779	139

Naive bayes

Confusion Matrix

	no (Predicted)	si (Predicted)	
no (Actual)	344	66	83.90%
si (Actual)	79	429	84.45%
	81.32%	86.67%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (ð□□¹)	Correctly Classified	Incorrectly Classified
84.20%	15.80%	0.681	773	145

Logica difusa

Confusion Matrix

	no (Predicted)	si (Predicted)	
no (Actual)	327	71	82.16%
si (Actual)	90	399	81.60%
	78.42%	84.89%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (ð□□¹)	Correctly Classified	Incorrectly Classified
81.85%	18.15%	0.635	726	161

Figura 11:

Red neuronal

Confusion Matrix

	no (Predicted)	si (Predicted)	
no (Actual)	330	80	80.49%
si (Actual)	68	440	86.61%
	82.91%	84.62%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (ð□□¹)	Correctly Classified	Incorrectly Classified
83.88%	16.12%	0.673	770	148

Maquina de vector de soporte

Confusion Matrix

	no (Predicted)	si (Predicted)	
no (Actual)	333	77	81.22%
si (Actual)	57	451	88.78%
	85.38%	85.42%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (ð□□¹)	Correctly Classified	Incorrectly Classified
85.40%	14.60%	0.703	784	134

K vecino mas cercano

Confusion Matrix

	no (Predicted)	si (Predicted)	
no (Actual)	241	169	58.78%
si (Actual)	146	362	71.26%
	62.27%	68.17%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (ð□□¹)	Correctly Classified	Incorrectly Classified
65.69%	34.31%	0.302	603	315

Figura 12:

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] 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.