Predicción de insuficiencia cardiaca como herramienta de tamizaje

El objetivo principal de este estudio es la predicción de eventos de insuficiencia cardíaca, con énfasis en la identificación temprana del riesgo de mortalidad en pacientes con afecciones cardiovasculares. La insuficiencia cardíaca constituye una de las complicaciones más frecuentes y severas de las enfermedades cardiovasculares, las cuales representan la principal causa de muerte a nivel mundial. Frente a esta realidad clínica, la utilización de datos estructurados y herramientas de análisis avanzado se vuelve crucial para detectar patrones relevantes, estratificar el riesgo clínico y apoyar la toma de decisiones médicas mediante modelos predictivos.

El conjunto de datos utilizado en este taller proviene de un estudio observacional y contiene información clínica de pacientes con diagnóstico o sospecha de insuficiencia cardíaca. Cada registro corresponde a un paciente único y contempla un total de 13 variables clínicas que describen aspectos fisiológicos, antecedentes médicos y parámetros de seguimiento. Estas variables son clave para evaluar el estado general del paciente, monitorear su evolución y predecir la probabilidad de desenlaces adversos, como la muerte durante el período de observación.

Variables del estudio

Cada variable registrada en el estudio aporta información clínica clave sobre el estado de salud del paciente y su posible evolución. Comprender el significado y la función de estas variables es fundamental para interpretar adecuadamente los datos, identificar factores de riesgo y construir modelos de predicción robustos en el contexto de la insuficiencia cardíaca. A continuación, se describen las variables implicadas:

Variable	Descripción	Importancia para el estudio
edad	Edad del paciente en años.	La edad avanzada es un factor de riesgo clave en la mortalidad por insuficiencia cardíaca.
anemia	Presencia de anemia (Sí/No).	La anemia puede agravar la insuficiencia cardíaca al reducir la capacidad de oxigenación.
fosfocinasa_creatinina	Nivel de la enzima CPK (mcg/L).	Elevaciones pueden indicar daño muscular o infarto, relevantes para diagnóstico y pronóstico.
diabetes	Diagnóstico de diabetes (Sí/No).	Factor de riesgo importante que acelera el deterioro cardiovascular.

fracción_eyección	Porcentaje de sangre que expulsa el corazón en cada latido.	Indicador directo de función cardíaca; valores bajos indican insuficiencia cardíaca.
presión_arterial_alta	Hipertensión arterial (Sí/No).	Condición frecuente asociada a insuficiencia cardíaca y eventos cardiovasculares.
plaquetas	Conteo de plaquetas (mil/µL).	Puede reflejar alteraciones hematológicas o inflamatorias que afectan el pronóstico.
creatinina_suero	Nivel de creatinina en sangre (mg/dL).	Indicador de función renal; el deterioro renal es frecuente en pacientes con insuficiencia cardíaca.
sodio_suero	Nivel de sodio en sangre (mEq/L).	Niveles bajos (hiponatremia) están asociados a mayor mortalidad en insuficiencia cardíaca.
sexo	Sexo biológico del paciente (Masculino/Femenino).	Existen diferencias fisiopatológicas y de respuesta al tratamiento entre hombres y mujeres.
fumador	Tabaquismo (Sí/No).	Factor de riesgo modificable que empeora la evolución cardiovascular.
tiempo_seguimiento	Días de seguimiento hasta la muerte o final del estudio.	Permite medir la duración de la observación y calcular tasas de eventos.
evento_muerte	Indicador de mortalidad durante el seguimiento (Sí/No).	Variable objetivo del estudio; permite construir modelos de predicción de riesgo.

Análisis de datos e identificación de tendencias

En esta actividad, los estudiantes deberán explorar la base de datos con el objetivo de identificar tendencias, patrones y asociaciones relevantes entre las variables clínicas y el desenlace de mortalidad. Este análisis descriptivo permitirá comprender mejor el perfil de los pacientes con insuficiencia cardíaca y detectar factores de riesgo predominantes. Identificar estas tendencias no solo es clave para el desarrollo de modelos predictivos, sino que tiene un alto valor aplicado en el contexto de la atención primaria en salud, donde los equipos clínicos enfrentan el desafío de tomar decisiones

oportunas con información limitada. Reconocer tempranamente a los pacientes en riesgo permite priorizar intervenciones, optimizar recursos y prevenir complicaciones mayores, mejorando así los resultados en salud poblacional.

1. Preprocesamiento de la base de datos

a. Manejo de valores nulos

Identificar los registros que contienen datos faltantes es un paso esencial en el preprocesamiento. Sin embargo, en el contexto clínico, surge un dilema: la imputación de valores ausentes no siempre es recomendable, ya que la alta variabilidad fisiológica y la complejidad de las variables pueden hacer que métodos como la media o la mediana distorsionen la interpretación clínica y comprometan la validez del análisis. Por tanto, se deja a criterio de los estudiantes decidir entre imputar o eliminar registros incompletos, debiendo justificar su elección en función de la proporción de datos faltantes y la relevancia clínica de cada variable.

b. Detección y tratamiento de valores atípicos

La detección y tratamiento de valores atípicos debe abordarse desde dos enfoques complementarios: por un lado, el análisis estadístico mediante herramientas como boxplots, percentiles e IQR; y por otro, el juicio clínico, dado que en medicina existen valores extremos que, aunque inusuales, son fisiológicamente plausibles según el contexto individual del paciente. En este sentido, no todos los outliers deben eliminarse, ya que algunos pueden representar estados críticos reales. Por ello, se proporciona una tabla con rangos clínicamente válidos que servirá como guía para ayudar a los estudiantes a tomar decisiones informadas sobre qué valores conservar, revisar o corregir, según su naturaleza y significado clínico.

Variable	Rango válido	Justificación fisiológica o clínica
edad	40 – 95 años	El estudio se enfoca en adultos con insuficiencia cardíaca. No hay pacientes pediátricos ni centenarios extremos (poco comunes y mayor riesgo de error de ingreso).
fosfocinasa_creatinina	23 – 7861 mcg/L	Valores muy bajos son normales en reposo. Valores altos indican daño muscular severo (infarto, trauma, miopatías). ≈ 7880 es elevado pero posible

		clínicamente, y se observa en rabdomiólisis, infarto agudo de miocardio o miopatías inflamatorias.
fracción de eyección	14% – 80%	<40% indica insuficiencia cardíaca. <20% es grave. >70–80% se puede observar en cardiomiopatías hipertróficas o atletas. Valores fuera de este rango son muy sospechosos.
plaquetas	25,100 – 850,000 /µL	Valores <150,000 se consideran trombocitopenia. <30,000 puede ser crítico. >≈ 850,000 indica trombocitosis o enfermedades mieloproliferativas leucemia mieloide crónica.
creatinina en suero	0.5 – 9.4 mg/dL	Normal: 0.6–1.3. >1.5 indica disfunción renal. >≈ 9.5 puede verse en insuficiencia renal crónica avanzada.
sodio en suero	113 – 148 mEq/L	135–145 es normal. <125 es hiponatremia severa (puede causar convulsiones). 113 es crítico pero se puede observar. >145 es hipernatremia.
tiempo de seguimiento	4 – 285 días	Tiempo registrado de seguimiento del paciente. Mínimo: 4 días (alta temprana o muerte rápida). Máximo: 285 días (casi 10 meses), es razonable para un estudio longitudinal.

c. Categorización y transformación de variables

En esta etapa, los estudiantes deben interpretar el funcionamiento de técnicas de codificación como Label Encoding y One-Hot Encoding, y considerar cuál de ellas resulta más pertinente según la naturaleza de las variables, los requerimientos del análisis exploratorio y las necesidades del posible modelo computacional a implementar.

2. Análisis grafico

El análisis visual de los datos permite detectar patrones, asociaciones y comportamientos que pueden no ser evidentes en los resúmenes numéricos. En esta etapa, se propone a los estudiantes abordar los siguientes retos gráficos:

1. **Distribución de la edad y su relación con la mortalidad:** Analizar cómo se comporta la edad en la población estudiada y si existe una tendencia clara en los desenlaces según grupos etarios.
2. **Distribución de variables clínicas frente al indicador de mortalidad:** Comparar gráficamente variables continuas entre pacientes fallecidos y no fallecidos.
3. **Evaluar asociaciones entre condiciones clínicas y mortalidad:** Examinar si variables categóricas como anemia, diabetes, hipertensión, tabaquismo y sexo presentan una distribución diferente según el desenlace, lo cual puede sugerir riesgo asociado.
4. **Relación combinada entre variables fisiológicas y mortalidad:** Graficar combinaciones clave como edad vs creatinina en suero y edad vs fracción-eyección, diferenciadas por moralidad, para identificar posibles interacciones que indiquen mayor riesgo.
5. **Mapa de calor de correlaciones:** Construir un heatmap para visualizar qué variables están correlacionadas entre sí, lo cual puede ser útil tanto para la interpretación como para la preparación de datos para modelos computacionales.

Desarrollo y evaluación de los modelos computacionales

Una vez completado el preprocesamiento y el análisis exploratorio de los datos, se procederá al desarrollo de modelos computacionales con el fin de predecir la mortalidad por insuficiencia cardíaca. Para este proceso, los estudiantes deberán considerar los siguientes aspectos clave:

1. **Balanceo de clases:** Dada la posible desproporción entre los pacientes fallecidos y no fallecidos, es necesario aplicar técnicas de balanceo que eviten sesgos en el entrenamiento del modelo. Los estudiantes deberán interpretar el funcionamiento de métodos como SMOTE, SMOTE-Tomek y ADASYN, y seleccionar la técnica que consideren más adecuada según la naturaleza de los datos y la distribución de la variable objetivo.

2. **Escalamiento de variables:** Muchas técnicas de clasificación se ven afectadas por diferencias en la escala de las variables numéricas. Se deberá aplicar escalamiento utilizando métodos como StandardScaler o RobustScaler, analizando su funcionamiento y eligiendo el más conveniente en función de los datos.
3. **Entrenamiento de modelos supervisados:** Se entrenarán tres algoritmos de clasificación: Decision Tree, Random Forest y Support Vector Machine (SVM).
4. **Evaluación de desempeño:** Se deberán calcular las principales métricas de evaluación para cada modelo: accuracy, f1-score, recall, y AUC (Área bajo la curva ROC). Además, se debe graficar la matriz de confusión para visualizar la distribución de verdaderos positivos, falsos negativos y otros errores de clasificación.
5. **Selección del modelo final:** Con base en la interpretación comparativa de las métricas de rendimiento, los estudiantes deberán seleccionar el modelo que ofrezca el mejor equilibrio en el escenario de atención.
6. **Trabajo futuro:** Explorar y mencionar como se podrían mejorar los resultados de los modelos computacionales.

Lecturas de apoyo

1. The anesthesiologist's guide to critically assessing machine learning research: a narrative review (https://doi.org/10.1186/s12871-024-02840-y).
2. Predicting no-shows at outpatient appointments in internal medicine using machine learning models (10.7717/peerj-cs.2762).
3. A Comparative Study of SMOTE, Borderline-SMOTE, and ADASYN Oversampling Techniques using Different Classifiers (10.1109/ICSMDI57622.2023.00060).