Predicción de Mortalidad y Complicaciones Cardiovasculares Intrahospitalarias en Pacientes con Síndrome Coronario Agudo Mediante un Modelo Machine Learning Supervisado

Fabio Andrés Morales Quintero, Investigador principal

Juan Camilo Ortiz Uribe, Cardiólogo Intervencionista, asesor científico

José John Fredy González Veloza, asesor estadístico

Resumen

Introducción: Los síndromes coronarios agudos (en adelante, SCA) representan la principal causa de muerte en el mundo. Los modelos de predicción clínica pueden ser útiles para tomar decisiones principalmente en pacientes de alto riesgo ya que precisan vigilancia temprana y tratamientos más agresivos. Existen numerosos modelos de predicción para los diferentes tipos de SCA, en su mayoría generalizados para predecir el riesgo que cuestiona su utilidad.

Este trabajo propone un modelo de predicción de riesgo simple basado en machine learning (en adelante, ML), que aplica a todos los tipos de SCA y se enfoca en mortalidad y complicaciones relevantes instrahospitalarias.

Objetivo: Evaluar la capacidad de discriminación que tiene un clasificador de machine learning para predecir desenlaces de mortalidad y complicaciones cardiovasculares instrahospitalarias en pacientes con SCA, atendidos en el Hospital San Vicente Fundación de Medellín.

Metodología: Se realizó un estudio de cohorte transversal en pacientes con SCA de enero de 2018 a enero de 2021. Inicialmente, se procedió con un análisis exploratorio, limpieza y transformación del conjunto de datos. Luego se dividió la data en subconjuntos de entrenamiento y testeo, y se implementó la técnica DBMIST-US para sub-muestrear la clase mayoritaria de entrenamiento, tomando una muestra representativa con un 95% de confianza. Por último, se entrenaron los modelos ML usando validación cruzada, y se evaluaron sus rendimientos por medio de la matriz de confusión y las curvas de aprendizaje.

Resultados: Se incluyeron 1783 pacientes, encontrando que para esta población la precisión de los modelos heurísticos es del 7.6%, la sensibilidad es del 96%, y la especificidad del 7.8%. Se encontraron dos problemas en el conjunto de datos que (desbalance y superposición de clases) que afectaron negativamente el rendimiento de los modelos basados en reglas y los de ML. Para enfrentar estas dos complejidades se aplicó la técnica DBMIST-US seleccionando una muestra representativa de la

clase mayoritaria de 92 instancias, con las cuales se entrenaron los modelos teniendo como resultado una precisión del 17%, sensibilidad del 92% y especificidad del 64%.

Conclusión: La técnica DBMIST-US es competitiva frente a otras técnicas de sub-muestreo y sobre-muestreo, para enfrentar los problemas de desbalance y superposición de clases. Se comprobó que no es necesario tener una gran cantidad de datos para predecir mortalidad, sino que basta una muestra representativa para optimizar el rendimiento de los clasificadores. Por otro lado, la práctica clínica está abocada a priorizar la sensibilidad sobre la precisión y la especificidad; sin embargo, con esta técnica se puede desarrollar una calculadora de riesgo que permita salvar una cantidad similar de pacientes disminuyendo los costos debido a que el número de falsos positivos disminuiría significativamente dando lugar a un incremento en la especificidad.

Palabras clave: Síndrome coronario agudo, predicción de mortalidad, DBMIST-US, árbol de expansión mínimo, superposición de clases u overlapping.

Abstract

Introduction: Acute coronary syndromes (hereinafter, SCA) represent the main cause of death in the world. Clinical prognoses can be useful for decision-making, especially in high-risk patients, since they require early surveillance and more aggressive treatment models. There are numerous prediction models for the different types of SCA, mostly generalized to predict risk that calls into question their usefulness.

This work proposes a simple risk prediction model based on machine learning (hereinafter, ML), which applies to all types of ACS and focuses on mortality and relevant in-hospital complications.

Objective: To evaluate the discrimination capacity of a machine learning classifier to predict mortality outcomes and in-hospital cardiovascular complications in patients with SCA, treated at the Hospital San Vicente Fundación de Medellín.

Methodology: A cross-sectional cohort study was conducted in patients with SCA from January 2018 to January 2021. Initially, an exploratory analysis, cleaning, and transformation of the data set was carried out. Then the data was divided into training and testing subsets, and the DBMIST-US technique was implemented to sub-sample the majority class, taking a representative sample with 95% confidence. Finally, the ML models were trained using cross-validation, and their performances were evaluated by means of the confusion matrix and the learning curves.

Results: 1783 patients were included, finding that for this population the precision of the heuristic models is 7.6%, the sensitivity is 96%, and the specificity is 7.8%. Two problems were found in the data set (class imbalance and overlap) that negatively affected the performance of the rule-based and ML models. To face these two complexities, the DBMIST-US technique was applied, selecting a representative sample of the majority class of 92 instances, with which the models were trained, resulting in an accuracy of 17%, sensitivity of 92%, and specificity of 64%.

Conclusion: The DBMIST-US technique is competitive against other sub-sampling and over-sampling techniques, to face the problems of imbalance and class overlap. It was found that it is not necessary to have a large amount of data to predict mortality, but that a representative sample intelligently selected is enough to optimize the performance of the classifiers. On the other hand, clinical practice is bound to prioritize sensitivity over precision and specificity; however, with this technique a risk calculator can be developed that allows a similar number of patients to be saved, reducing costs because the number of false positives would decrease significantly, giving rise to an increase in specificity.

Keywords: Acute coronary syndrome, mortality prediction, DBMIST-US, minimal spanning tree, class overlap.

Predicción de Mortalidad y Complicaciones Cardiovasculares Intrahospitalarias en Pacientes con Síndrome Coronario Agudo Mediante un Modelo de Machine Learning Supervisado

Los SCA representan la principal causa de muerte en el mundo (Martinez Merlo et al., 2019; Gaviria et al., 2020; Infarto agudo de Miocardio. Causas, síntomas y tratamiento. Clínica Universidad de Navarra, s.f.), y existe una variabilidad importante en cuanto a sus complicaciones dependiendo a través del espectro en el SCA (Collet et al., 2021). Los modelos de predicción clínica pueden ser útiles para tomar decisiones médicas principalmente en pacientes identificados como de alto riesgo ya que pueden requerir vigilancia temprana o tratamientos mas agresivos. Usando calculadoras simples validadas, el clínico, de manera mas precisa puede sugerirle al enfermo con SCA cuál sería la probabilidad de tener un evento y cómo esto se traslada a decisiones clínicas.

El SCA abarca un conjunto de condiciones que van desde la angina inestable, SCA sin elevación del ST y SCA con elevación del ST (Cassiani M & Cabrera G, 2009).

Existen numerosos modelos de predicción para los diferentes tipos de SCA (Eagle et al., 2004); la gran mayoría se han desarrollado a partir de estudios clínicos aleatorizados muy generalizados para predecir el riesgo, por lo tanto su utilidad puede ser cuestionable.

Este trabajo propone el desarrollo de un modelo de predicción de riesgo simple basado en ML aplicable a todos los tipos de SCA y que se enfoca en predicción de mortalidad y complicaciones relevantes intrahospitalarias.

Método

Se realizó un estudio retrospectivo de pacientes con SCA de enero de 2018 a enero de 2021, usando las características clínicas de los pacientes (edad, troponina I de alta sensibilidad, FEVI, frecuencia cardiaca, creatinina sérica, presión arterial sistólica, presión diastólica, colesterol HDL, colesterol LDL, Puertabalón, Reinfarto, entre otros) como variables predictoras. La mortalidad es la variable objetivo para este estudio y se caracteriza por tener distribución de Bernoulli, y se obtuvo del historial médico del Hospital.

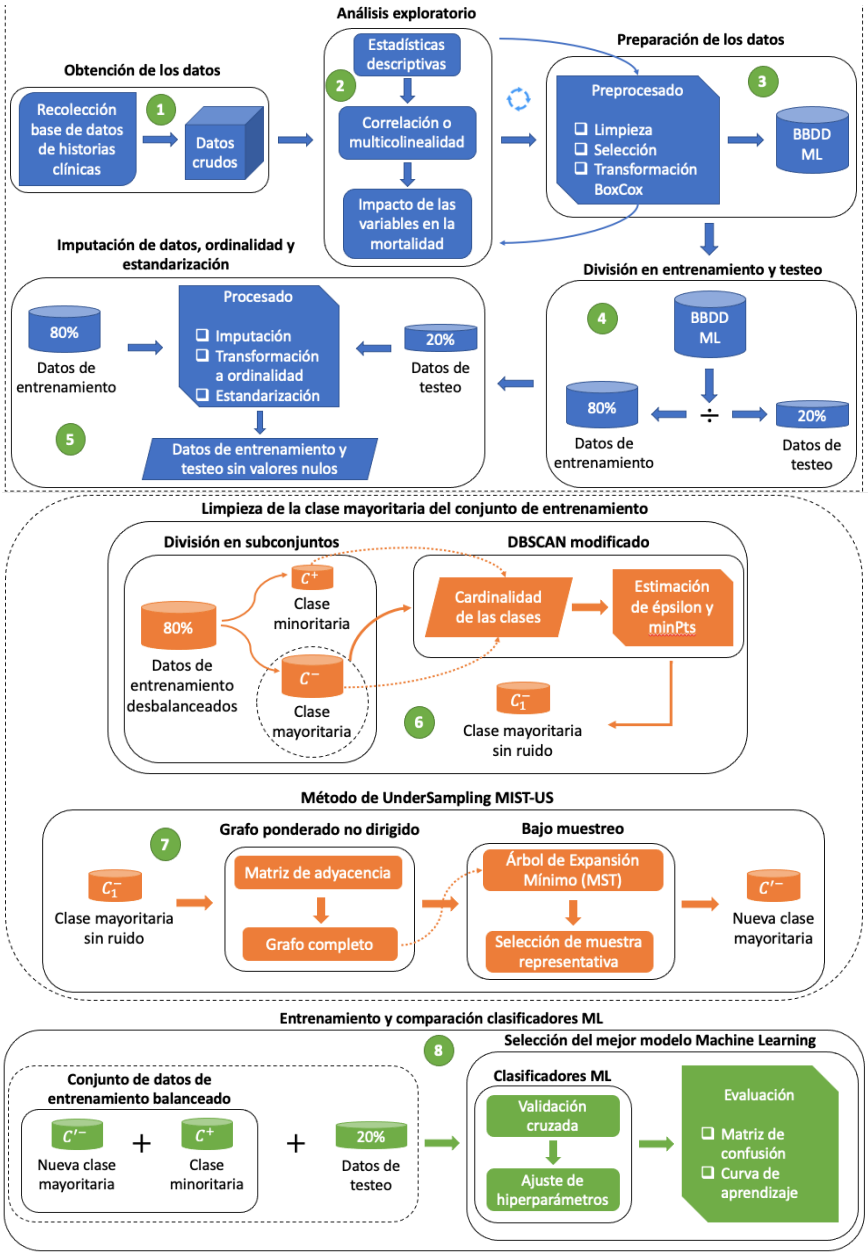
Preprocesamiento de los Datos

La figura 1 muestra las etapas o fases de la metodología implementada en este trabajo: se recolectó una base de datos de historias clínicas de pacientes con SCA (angina inestable, SCA con y sin elevación del ST), del Hospital san Vicente Fundación de Medellín. Después se realizó el tratamiento de las variables para limpiar, imputar datos nulos, analizar correlaciones y hacer transformaciones BoxCox para corrección de asimetrías en las variables numéricas. Paso a seguir se dividió la base en entrenamiento y testeo y se diseñaron modelos heurísticos. Luego se entrenaron los modelos de la librería de PyCaret (Training functions in PyCaret, s.f.).

Con el ánimo de mejorar el desempeño del modelo se aplicó *Synthetic Minority Over-sampling Technique* (en adelante, SMOTE), el cual crea muestras ficticias de la clase minoritaria para igualar la cardinalidad de la clase mayoritaria. No obstante, el rendimiento de los clasificadores no mejoró. Se evidenció mediante visualización, que los datos tenían el problema de superposición o traslape de clases, el cual consiste en que las clases cubren la misma región del espacio por lo que se hace difícil que un clasificador identifique los grupos a los que pertenecen las instancias.

Figura 1

Etapas del Proceso de Entrenamiento de Modelos ML



Nota. La figura muestra las etapas de preprocesamiento, bajo-muestreo y entrenamiento en orden descendente, respectivamente, cada una con una serie de pasos para optimizar el rendimiento de los clasificadores en la predicción de mortalidad en pacientes con SCA. Fuente: Elaboración propia.

Por lo cual, se empleó la estrategia propuesta por Guzmán Ponce (2021) la cual se trata de un método de UnderSampling denominado DBMIST-US que aplica el DBSCAN modificado que se caracteriza por calcular los parámetros épsilon (en adelante, eps) y minPts a partir de la cardinalidad de la clase mayoritaria según las ecuaciones de Smiti & Elouedi (2012), y luego aplicar un árbol de expansión mínimo (en adelante, MST) para seleccionar una muestra representativa de la clase mayoritaria.

Obtención de los Datos

Se realizó la recolección de información de historias clínicas (datos crudos) considerando las características clínicas de los pacientes como variables predictoras. La mortalidad se estableció como variable objetivo, y se caracteriza por ser binaria, es decir, solo tiene dos valores: 1 o 0, el paciente falleció o sobrevivió, respectivamente.

Análisis Exploratorio

Se llevó a cabo la inspección de los datos (verificación del tipo de dato, cantidad de registros duplicados, estadísticas descriptivas, desbalance de clases, porcentaje de datos faltantes), correlación de datos nulos, análisis de multicolinealidad, y una visualización de la distribución de las variables y su impacto en la mortalidad.

Preparación de los Datos

Se hizo limpieza, selección y transformación de variables. Para esto, se realizó la corrección de digitación en variables numéricas y categóricas, eliminación de registros duplicados y de aquellas variables redundantes o con alta correlación que tenían menor impacto sobre la mortalidad. Por último, se aplicó transformación BoxCox para corregir la asimetría de las variables apoyado en el test de kurtosis y skewness con p-valor del 5%.

División en Entrenamiento y Testeo

Se tomó el 80% de los datos como base de entrenamiento y el 20% como testeo.

Imputación de Datos, Ordinalidad y Estandarización

Se usó la moda para imputar datos en variables categóricas, en variables numéricas se imputó la mediana si la cantidad de datos faltantes era menor al 10% y se usó interpolación con datos existentes si superaba el 25%. Las variables categóricas que tienen implícita una jerarquía se convirtieron a ordinales, y las numéricas se estandarizaron.

Tratamiento de las Complejidades de los Datos

En esta etapa se enfrentó el problema del desbalance y la superposición de clases, mediante un algoritmo basado en grafos y clustering conocido como DBMIST-US, el cual resulta de combinar DBSCAN con el MST.

Limpieza de la Clase Mayoritaria del Conjunto de Entrenamiento

Se separaron las clases minoritaria y mayoritaria del conjunto de entrenamiento. Luego se sometió la clase mayoritaria a una limpieza de instancias ruidosas (también llamadas, instancias atípicas) por medio del algoritmo de DBSCAN cuyos parámetros se estimaron usando la cardinalidad de ambas clases. Las instancias ruidosas o *ruido* son aquellas que tienen definida una clase, pero tienen similitudes con instancias de otras clases (Guzmán Ponce, 2021).

Método de UnderSampling MIST-US

Una vez se tuvieron los datos de la clase mayoritaria sin ruido, se construyó un grafo completo ponderado mediante una matriz de adyacencia expresando cada instancia en términos de vértices y aristas. Paso seguido, se construyó un MST generado por el algoritmo de Prim (1957) con el fin de obtener instancias lo más alejadas de la frontera de decisión, donde finalmente se tomaron las S-primeras del MST cuya cantidad se determinó por la ecuación 4 con un 95% de confianza, siendo similar a la cardinalidad de la clase minoritaria.

Entrenamiento y Comparación de Clasificadores Machine Learning

Una vez seleccionada la nueva clase mayoritaria del conjunto de entrenamiento, se unió nuevamente con la clase minoritaria dando origen al conjunto de datos de entrenamiento balanceado, el cual se usó para entrenar los clasificadores ML. Por su parte, los datos de testeo permanecieron intactos, es decir, no se les aplicó limpieza ni submuestreo.

Selección del Mejor Modelo Machine Learning

Para la etapa de entrenamiento de los clasificadores, se implementó validación cruzada de 5-fold y se ajustaron los hiperparámetros de clasificador que se consideró de mejor rendimiento, con el fin de reducir el sobreajuste en el aprendizaje. La evaluación del modelo seleccionado se realizó por medio de la matriz de confusión para calcular métricas tales como precisión, especificidad, sensibilidad y F1_Score. También se analizaron los resultados mediante las curvas de aprendizaje.

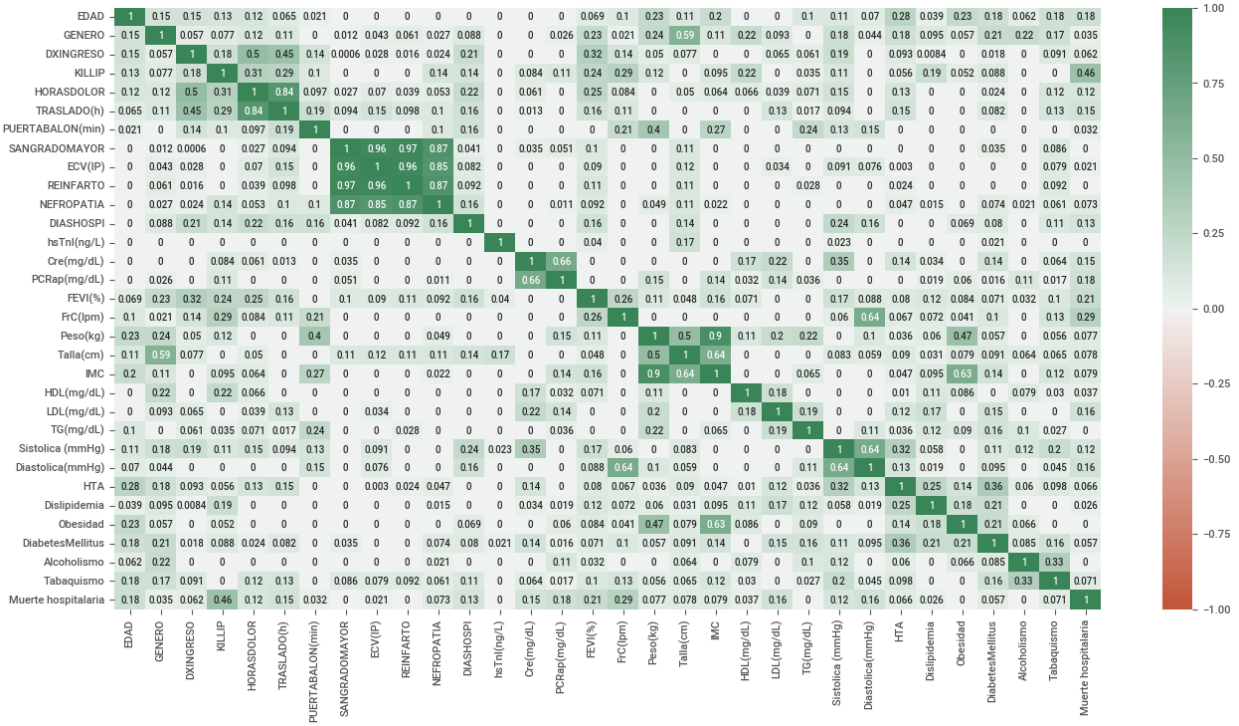
Por último, se realizó un análisis global de los resultados obtenidos al entrenar los clasificadores con la base de entrenamiento en bruto y la base de entrenamiento balanceada con DBMIST-US. De igual forma, se compararon las técnicas de sobremuestreo y submuestreo aplicadas a la base de entrenamiento en bruto con DBMIST-US.

Resultados

Dentro del análisis descriptivo, se encontró que algunas variables presentaban alta correlación (Figura 2). Debido a que la variable IMC está explicada por el peso(kg) y la talla(cm), se decidieron eliminar estas dos últimas. Del mismo modo, se eliminaron las variables SANGRADOMAYOR, REINFARTO, HORASDOLOR y TRASLADO(h).

Figura 2

Correlación de las Variables Categóricas y Numéricas de Pacientes con SCA



Nota. Matriz de correlaciones entre las variables categóricas y numéricas usando el coeficiente de correlación Phi_K, el cual está basado en varios refinamientos del test de hipótesis de Pearson de independencia de dos variables (KPMG Analytics and Visualization Environment, s.f.). Fuente: Elaboración propia.

Se construyeron modelos heurísticos debido a que reflejan lo que se hace en la práctica clínica, por lo que se analizaron algunas variables para evaluar el rendimiento. En la Tabla 1 se observa que las variables KILLIP IV y hsTnI (ng/L) > 5 son buenos predictores de especificidad, mientras que Sistolica (mmHg) < 90 y Edad (años) > 80 son buenos predictores de sensibilidad. La variable KILLIP IV es la única que al parecer guarda un balance entre precisión y recall, debido a que en las otras tres se evidencia que al aumentar una de ellas, la otra resulta en desmejoramiento.

Tabla 1

Rendimiento de Modelos Heurísticos

Variable	Precisión (%)	Sensibilidad (%)	Especificidad (%)	F1-score (%)
KILLIP IV	32.3	38.5	93.65	35.1
hsTnI (ng/L) > 5	28.6	7.7	98.48	12.1
Sistolica (mmHg) < 90	7.4	96.2	5.13	13.7
Edad (años) > 80	7.0	96.2	0.00	13.1

Nota. Fuente: Elaboración propia

Se encontró que los datos presentaban además de desbalance, el problema de superposición de clases u *overlapping,* lo que implica una pérdida de rendimiento en los clasificadores (Guzmán Ponce, 2021).

Debido a lo anterior, se optó por aplicar un método de sub-muestreo conocido como DBMIST-US el cual emplea la técnica de clustering DBSCAN modificado para la limpieza de ruido (Figura 3) y el MST para la selección de una muestra representativa (Figura 4).

La limpieza de ruido se hizo aplicando las ecuaciones propuestas por Smiti & Elouedi (2012), en las cuales se usa la cardinalidad de la clase mayoritaria (ecuación 1) para calcular el parámetro eps, la cardinalidad de la clase minoritaria (ecuación 2) y la ecuación 3 para calcular minPts. La propuesta de

Guzmán Ponce (2021) toma el 20% del valor de eps, sin embargo, en este trabajo se eligió el 70% del valor computado.

$$eps = \sqrt{\frac{\sum_{i=1}^{|C^-|} distancia^2(m, p_i^-)}{|C^-|}} \times 0.7 \quad (1)$$

Donde m es el vector medio de la clase mayoritaria, p_i^- representa una instancia de la clase mayoritaria, y $distancia$ es la distancia euclídea.

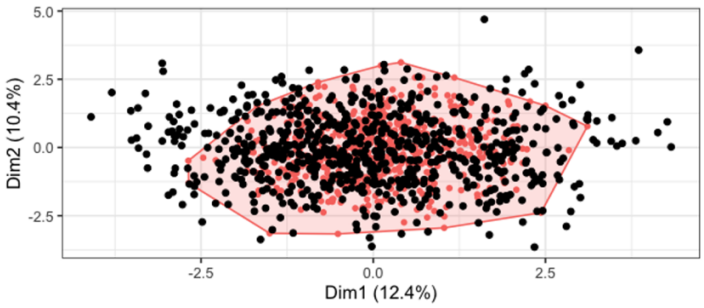
$$minPts = \frac{\pi \times eps^2}{TotalVolume} \times |C^+| \quad (2)$$

Donde,

$$TotalVolume = \frac{4}{3} \times \pi \times eps^3 \quad (3)$$

Figura 3

Diagrama de Dispersión del Ruido de la Clase Mayoritaria Usando DBSCAN Modificado



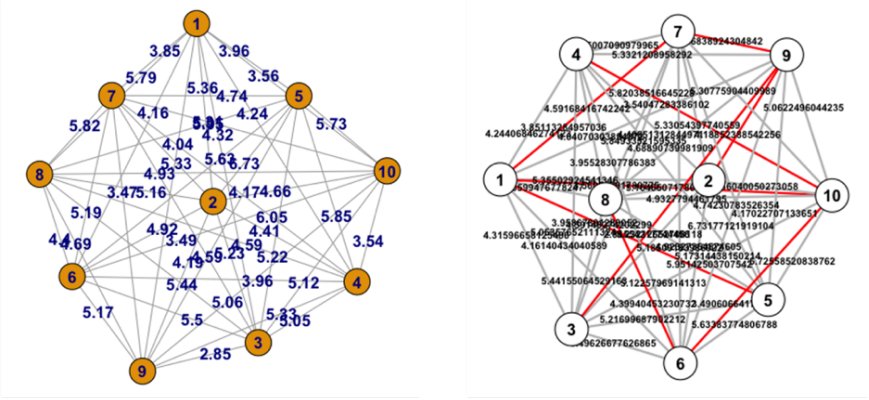
Nota. Representación gráfica de la clase mayoritaria de entrenamiento en las dos primeras componentes principales . Los puntos de color negro son instancias atípicas etiquetadas por el DBSCAN con eps de 2.58 y minPts de 30, que posteriormente fueron eliminadas, quedando solamente las instancias al interior del polígono convexo. Fuente: Elaboración propia.

Luego de aplicar la limpieza de instancia ruidosas, se construyó el grafo completo ponderado no dirigido (Figura 4, inciso a) a partir de la generación de la matriz de adyacencia, expresando las instancias de la clase mayoritaria en términos de vértices y aristas. La matriz de adyacencia es una matriz cuadrada que contiene los pesos de las aristas, es decir, las distancias euclídeas entre cada par de vértices.

El inciso b de la Figura 4 representa el MST generado por el algoritmo de Prim (1957), el cual toma los *S*-primeros vértices determinados por la ecuación 4 de la matriz de adyacencia ordenada de manera ascendente. Para construir el MST se usó la librería de R llamada *optrees,* diseñada por Fontenla Cadavid (2014).

Figura 4

Grafo Completo y Sub-grafo MST de la Clase Mayoritaria sin Instancias Ruidosas



Nota. Se construyó el (a) grafo completo de la clase mayoritaria de entrenamiento limpia de ruido por DBSCAN modificado, y a partir de este se obtuvo el (b) sub-grafo del MST (aristas de color rojo). Se seleccionaron solo diez instancias para poder visualizar con claridad los vértices y aristas con sus respectivos pesos. Fuente: Elaboración propia.

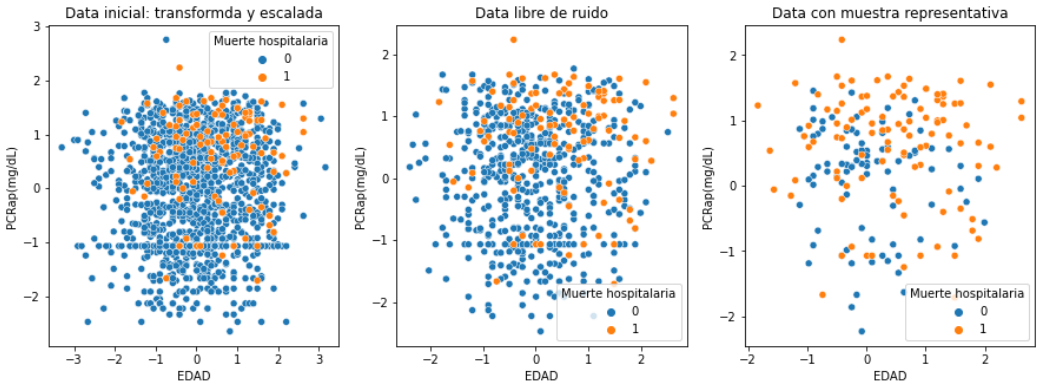
Por último, se tomó una muestra mediante la ecuación 4 propuesta por Guzmán Ponce (2021), donde $Z = 1.96$, $e = 0.05$ y $\sigma = 0.5$ para el 95% de confianza de acuerdo al procedimiento de Torres et al. (2006). Como resultado, la técnica DBMIST-US redujo la clase mayoritaria inicial de 1400 datos aproximadamente a 92 instancias (Figura 5), la cual es una cantidad similar a la clase minoritaria con 104 instancias.

$$S = \frac{S \quad \sigma^2 Z^2}{e^2(|C^-| - 1) + \sigma^2 Z^2} \tag{4}$$

Donde *S* es el número de instancias de la muestra representativa.

Figura 5

Proceso de Sub-muestreo de la Clase Mayoritaria de Entrenamiento



Nota. Después de hacer el preprocesamiento de datos de entrenamiento de la clase mayoritaria (izquierda) se usó el DBSCAN modificado para limpiar instancias ruidosas (centro), y por último, se seleccionó una muestra representativa (derecha, color azul) mediante el MST. Fuente: Elaboración propia.

Para la fase de entrenamiento, se escogió el clasificador Random Forest (en adelante, RF) debido a su alta precisión y rendimiento para conjuntos de datos grandes y complejos (InteractiveChaos, s.f.); sin embargo, se ajustaron los parámetros de max_depth a 4, n_estimators a 80, max_leaf_nodes 9, min_samples_split a 4 y min_samples_leaf a 2 para ayudar a reducir el sobreajuste en el modelo; el ajuste se realizó mediante pruebas de ensayo y error.

Se seleccionaron umbrales de decisión con el objetivo de equilibrar los niveles de precisión y sensibilidad, y se encontró que, para las técnicas como SMOTE y Convolutional Neural Network (en adelante, CNN) aplicadas a la data en bruto, los umbrales con mejor precisión y recall eran inferiores a 0.5, mientras que para la base transformada con DBMIST-US el mejor umbral fue de 0.53 (Tabla 2).

Tabla 2

Rendimiento del RF con Diferentes Técnicas Aplicadas a los Datos en Bruto

Técnica	Umbral	Precisión (%)	Sensibilidad (%)	Especificidad (%)	F1-score (%)	AUC (%)
SMOTE	0,30	15,3	92,3	59,8	26,2	85
CNN	0,20	17,2	100,0	62,2	29,3	88
NearMiss	0,50	17,6	92,8	66,2	29,7	84
DBMIST-US	0,53	17,0	92,3	64,6	28,7	84

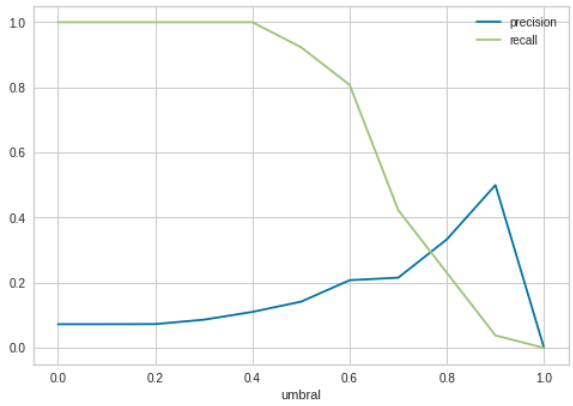
Nota. Estas técnicas se aplicaron sobre la clase mayoritaria de entrenamiento en su estado de impureza.

Fuente: Elaboración propia.

Los umbrales inferiores a 0.5 indican que los modelos favorecen la sensibilidad, lo que significa que se disminuyen los falsos negativos y se aumentan los falsos positivos, mientras que umbrales superiores a 0.5 mejoran la precisión al hacer que disminuyan los falsos positivos con la consecuencia de aumentar los falsos negativos (Figura 6).

Figura 6

Balance entre Precisión y Sensibilidad



Nota. En el eje X están representados los umbrales de decisión, y en el eje Y los valores de precisión y sensibilidad. Fuente: Elaboración propia.

De acuerdo con la Tabla 2, las técnicas NearMiss y DBMIST-US producen un rendimiento similar en el clasificador. Sin embargo, en la fase de entrenamiento con validación cruzada de 5 fold, las métricas de DBMIST-US son mejores que las de NearMiss (Tabla 3).

Tabla 3

Rendimiento del RF en la Fase de Entrenamiento

Técnica	AUC (%)	Precisión (%)	Sensibilidad (%)	F1-score (%)
NearMiss	78.57	13.14	78.81	22.46
DBMIST-US	90.8	82.62	85.52	83.47

Nota. Fuente: Elaboración propia.

Sumado a esto, cabe resaltar que NearMiss elimina aleatoriamente muestras de la clase más grande cuando éstas se encuentran muy cerca a muestras de la otra clase (Madhukar, 2020), mientras

que DBMIST-US selecciona de manera inteligente instancias por medio del MST (Guzmán Ponce, 2021). Ambas técnicas se realizaron con hiperparámetros modificados del RF.

Tabla 4

Rendimiento del RF con Diferentes Técnicas Aplicadas a los Datos sin Ruido

Técnica	Umbral	Precisión (%)	Sensibilidad (%)	Especificidad (%)	F1-score (%)	AUC (%)
SMOTE	0,40	17,7	84,6	69,2	29,3	86
CNN	0,40	17,6	96,2	64,6	29,8	86
TomekLinks	0,18	16,7	84,6	66,8	27,8	83
NearMiss	0,50	18,3	88,5	68,9	30,3	85

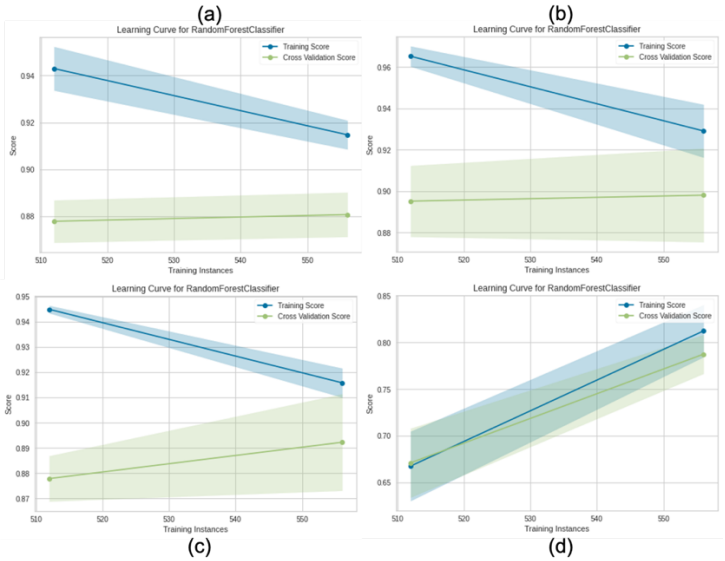
Nota. Estas técnicas se aplicaron sobre las clase mayoritaria de entrenamiento sin instancias atípicas.

Fuente: Elaboración propia.

Por otra parte, al aplicarlas a la base limpia por medio de DBSCAN modificado, se observó que las curvas de aprendizaje seguían una pendiente positiva o negativa sin llegar a una constante, indicando que los clasificadores necesitaban más datos para seguir aprendiendo (Figura 7), por lo que el rendimiento del RF que se observa en la Tabla 4 no es determinante para asegurar que el clasificador está generalizando.

Figura 7

Aprendizaje del Modelo RF con Técnicas Aplicadas a los Datos sin Ruido



Nota. Curvas de aprendizaje del modelo RF aplicando las técnica (a) SMOTE, (b) CNN, (c) TomekLinks y (d) NearMiss a la clase mayoritaria de entrenamiento sin instancias ruidosas. Fuente: Elaboración propia.

Teniendo en cuenta que la clase minoritaria cuenta con 104 instancias y que la clase mayoritaria resultante de aplicar DBMIST-US consta de 92, se aplicaron las técnicas de sobremuestreo Robot Operating System (en adelante, ROS) y SMOTE para conseguir un grado de desbalance igual a 1 (Tabla 5).

Con un umbral de 0.55 en ROS y SMOTE, se encontró que el modelo tiene un rendimiento similar a cuando se aplica DBMIST-US, favoreciendo la precisión pero disminuyendo el nivel de sensibilidad aproximadamente en un 4%.

Tabla 5

Rendimiento del RF con DBMIST-US, ROS y SMOTE

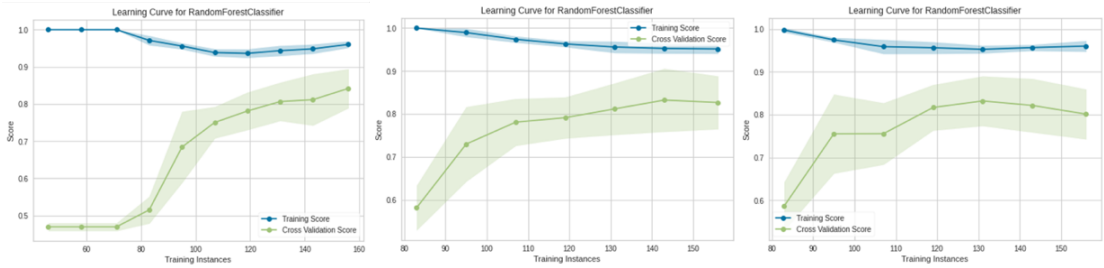
Técnica	Umbral	Precisión (%)	Sensibilidad (%)	Especificidad (%)	F1-score (%)	AUC (%)
DBMIST-US	0,53	17,0	92,3	64,6	28,7	84
ROS	0,55	19,5	88,5	71,3	31,9	85
SMOTE	0,55	17,8	88,5	67,9	29,7	84

Nota. Fuente: Elaboración propia

En la Figura 8 se puede observar que las curvas de aprendizaje conservan un sobreajuste reducido. Sin embargo, la curva de ROS (inciso b) parece ser más suave y estable, la curva de SMOTE (inciso c) parece desmejorar el aprendizaje cuando el modelo es entrenado con instancias numerosas, mientras que la curva de DBMIST-US tiende a necesitar más instancias para seguir aprendiendo. Esto puede significar que el desbalance igual a 1 es determinante para mejorar el rendimiento de un clasificador.

Figura 8

Aprendizaje del Modelo RF con Grado de Desbalance de Clase Igual a la Unidad



Nota. Estas tres curvas muestran el aprendizaje del modelo RF cuando se aplica (a) DBMIST-US a la conjunto de datos en bruto, y luego se usa (b) ROS y (c) SMOTE a la nueva clase mayoritaria para conseguir un grado de desbalance de 1. Fuente: Elaboración propia.

Discusión

La superposición de clases afecta significativamente el rendimiento de los clasificadores debido a que estos son incapaces de etiquetar correctamente las instancias. Las técnicas de bajomuestreo y sobremuestreo conocidas como CNN, TomekLinks, NearMiss, SMOTE, ROS, no fueron eficaces para enfrentar el problema de *overlapping*; no obstante, se observó que estas técnicas mejoraban la precisión con la implicación de que las curvas de aprendizaje presentaban sobreajuste, es decir, el modelo de RF estaba memorizando más no generalizando.

Por otro lado, DBMIST-US resultó ser eficaz para enfrentar el desbalance y la superposición de clases al mismo tiempo, debido a que mejoró la precisión del modelo haciendo que éste tuviera la capacidad de generalizar. El problema de superposición de clases es adecuadamente tratado con DBSCAN modificado, mientras que el desbalance es abordado de forma favorable con los grafos usados. Esta técnica ayudaría a disminuir los costos del hospital salvando la misma cantidad de pacientes.

Se evidenció que DBSCAN modificado ayuda a mejorar la precisión debido a que disminuye instancias ruidosas de la clase mayoritaria, que pueden ser consideradas como falsos positivos por los clasificadores debido a que se encuentran muy cercanas a muestras de la clase minoritaria; en otras palabras, se produce un mal etiquetado que conduce a una clasificación incorrecta (Guzmán Ponce, 2021). Cabe señalar, que los modelos ML tienen un mejor rendimiento cuando el conjunto de datos se encuentra libre de ruido.

La ecuación 1 indicó con un 95% de confianza la cantidad de 92 instancias de la clase mayoritaria como muestra representativa, reduciendo el volumen de los datos de la clase negativa a un 6.57% aproximadamente, logrando incrementar el rendimiento del clasificador. Se concluye que, de acuerdo con Maillo et al. (2020), la reducción del conjunto de datos mediante la selección inteligente de instancias ayuda a los clasificadores a tener un rendimiento similar o mejor.

También se concluye que los modelos entrenados y optimizados con DBMIST-US, son mejores en términos de métricas que los modelos basados en reglas los cuales representan la forma empírica de clasificar a los pacientes en el Hospital.

Los modelos heurísticos demostraron que en la práctica la evaluación clínica está supeditada a escoger entre la precisión y la sensibilidad ya que se observó que al mejorar el rendimiento de una de ellas automáticamente la otra se veía afectada; sin embargo, el RF optimizado con DBMIST-US, tiene mejor rendimiento dado que incrementa la precisión de un 7.6% a un 17% con una leve disminución en la sensibilidad del 96% al 92%, por lo que la toma de decisiones clínicas puede hacerse con más eficiencia y mayor discriminación; en otras palabras, se clasifican menos personas vivas como fallecidas.

También se evidenció que las variables de manera aislada no son buenos predictores de mortalidad, que la Troponina y la frecuencia cardíaca son buenos predictores de especificidad y la presión sistólica es buen predictor de sensibilidad.

Sumado a esto, se logró crear una herramienta de clasificación de pacientes con SCA validada para la población interna.

Referencias

Martinez Merlo, J. A., Lastre Amell, G. E., & Cassiani, C. (2019). Cuidados de enfermería en pacientes con Síndrome Coronario Agudo (SCA). Ene, 13(2), 1-13.

Gaviria, S., Ramírez, A., Alzate, M., Contreras, H., Jaramillo, N., & Muñoz, M. C. (2020). Epidemiología del síndrome coronario agudo. Medicina UPB, 39(1), 49–56. https://doi.org/10.18566/medupb.v39n1.a08

Infarto agudo de Miocardio. Causas, síntomas y tratamiento. *Clínica Universidad de Navarra.* (s.f.). Cun.es. Recuperado el 27 de marzo de 2023, de https://www.cun.es/enfermedades-tratamientos/enfermedades/infarto-miocardio

Collet, J.-P., Thiele, H., Barbato, E., Barthélémy, O., Bauersachs, J., Bhatt, D. L., Dendale, P., Dorobantu, M., Edvardsen, T., Folliguet, T., Gale, C. P., Gilard, M., Jobs, A., Jüni, P., Lambrinou, E., Lewis, B. S., Mehilli, J., Meliga, E., Merkely, B., … Siontis, G. C. M. (2021). Guía ESC 2020 sobre el diagnóstico y tratamiento del síndrome coronario agudo sin elevación del segmento ST. *Revista española de cardiología*, 74(6), 544.e1-544.e73. https://doi.org/10.1016/j.recesp.2020.12.024

Cassiani M, C. A., & Cabrera G, A. (2009). Síndromes coronarios agudos: epidemiología y diagnóstico. *Salud Uninorte, 25*(1), 118-134.

Eagle, K. A., Lim, M. J., Dabbous, O. H., Pieper, K. S., Goldberg, R. J., Van de Werf, F., Goodman, S. G., Granger, C. B., Steg, P. G., Gore, J. M., Budaj, A., Avezum, A., Flather, M. D., Fox, K. A. A., & GRACE Investigators. (2004). A validated prediction model for all forms of acute coronary syndrome: estimating the risk of 6-month postdischarge death in an international registry: Estimating the risk of 6-month postdischarge death in an international registry. *JAMA: The Journal of the American Medical Association*, 291(22), 2727–2733. https://doi.org/10.1001/jama.291.22.2727

Training functions in PyCaret. (s.f.). *PyCaret*: https://pycaret.gitbook.io/docs/get-started/functions/train

Guzmán Ponce, A. (2021). *Nuevos algoritmos basados en grafos y clustering para el tratamiento de complejidades de los datos.* México.

Smiti, A., & Elouedi, Z. (June de 2012). DBSCAN-GM: An improved clustering method based on Gaussian Means and DBSCAN techniques. *IEEE 16th International Conference on Intelligent Engineering Systems*, 573-578.

Prim, R. C. (1957). Shortest Connection Networks And Some Generalizations. *The Bell System Technical Journal, 36*(6), 1389-1401.

KPMG Analytics and Visualization Environment. (s.f.). Phi_K Correlation Analyzer Library. https://phik.readthedocs.io/en/latest/

Fontenla Cadavid, M. (2014). *Optimal Trees: Programación en R de problemas de búsqueda de árboles óptimos.* Coruña, España.

Torres, M., Paz, K., & Salazar, F. (2006). Tamaño de una muestra para una investigación de mercado. *Boletín Electrónico, 2*, 1-13.

InteractiveChaos. (s.f.). *Random Forest*. https://interactivechaos.com/es/wiki/random-forest

Madhukar, B, (October 29, 2020) Using Near-Miss Algorithm For Imbalanced Datasets https://analyticsindiamag.com/using-near-miss-algorithm-for-imbalanced-datasets/

Maillo, J., Triguero, I., & Herrera, F. (2020). Redundancy and Complexity Metrics for Big Data Classification: Towards Smart Data. *IEEE Access, 8*, 87918-87928.