Predicción de complicaciones durante la hospitalización del paciente de infarto de miocardio

Juan Acostupa
Dpto. de Ing. Informática PUCP
Lima, Perú
a20244906@pucp.edu.pe

Breno Muñoz
Dpto. de Ing. Informática PUCP
Lima, Perú
a20162955@pucp.edu.pe

Ronny Huerta
Dpto. de Ing. Informática PUCP
Lima, Perú
a20184255@pucp.edu.pe

I. Introducción

El ataque al miocardio es una de las complicaciones de salud que causan mas muertes a nivel mundial, en el Perú por ejemplo, mas de 4 mil personas mueren al año debido a esta emergencia médica [1]. Las causas que lo originan son diversas como: enfermedad de las arterias coronarias, acumulación de colesterol, artereoesclerosis, coágulos de sangre formados en las arterias, atero-trombosis, la edad, el sedentarismo, y una alimentación desbalanceada. Sin embargo, se ha estudiado que esta enfermedad cardiovascular tiene patrones los cuales puedes ser usados para su análisis mediante el uso de una gran cantidad de datos recogidos de pacientes como en [2], para su aplicación de modelos de aprendizaje automático.

El presente estudio se focaliza en el tratamiento de los datos de dos problemas: el primero, en la predicción de las complicaciones que pueda tener el paciente basado en la información recolectada al momento de su internamiento, y el segundo problema de la predicción de las complicaciones luego de un periodo de tres días de internamiento en el hospital.

El presente informe se organiza de la siguiente manera: en primer lugar, se describe una introducción general a la problemática (sección I) junto con el objetivo principal a abordar en el presente trabajo; en la segunda parte (sección II) se hace una breve síntesis del estado del arte con artículos científicos usando los mismos datos del presente proyecto y con otros diferentes pero con alternativas de solución en relación al problema mencionado; en la tercera parte (sección III) se explica sobre el diseño del experimento, en el cual se ha hace una descripción detallada del conjunto de datos y la metodología a usar para el abordaje del presente proyecto.

II. Estado del arte

Existen múltiples trabajos que tratan sobre diferentes aspectos relacionados a la utilización de diferentes algoritmos de aprendizaje automático y la importancia del rol que juegan estos métodos en la mejora del diagnóstico y prevención de complicaciones. *Quer et. al.* [3] hace un énfasis especial en el rol que juegan el personal médico, en especial los cardiólogos. Asimismo, *Van den Eynde et. al.* [4] hacen mención de un número de ejemplos exitosos de técnicas de inteligencia artificial en casos de cardiología clínica, mientras que *Sevakula et. al.* [5] hacen un estudio sobre nuevos casos que se le pueden dar a estas técnicas, además de las limitaciones que estas tienen. Finalmente, *Benzakour et. al.* [6] hace un estudio extenso de los conjuntos de datos publicados sobre diferentes tipos de enfermedades cardiovasculares, encontrando 39 conjuntos de datos diferentes. Entre ellos, el dataset objeto de estudio del presente trabajo.

El dataset Myocardial Infarction Complications Database se ha utilizado previamente para predecir complicaciones en pacientes que han sufrido infarto del miocardio. A continuación se resumen las metodologías respecto del tratamiento de los datos y los modelos de clasificación sobre estos mismos datos y sobre otras bases de datos [8].

1) Preprocesamiento: Respecto del preprocesamiento de los datos, algunos autores coinciden en tratar los datos faltantes eliminando aquellas variables cuyos faltos exceden un límite porcentual máximo. En cambio, otros utilizan técnicas de reducción de la dimensionalidad [8] como el Análisis de Componentes Principales (PCA) y el catPCA. Teóricamente, este último está pensado para mantener la máxima correlación entre las variables cuando se tienen salidas con datos de tipo categórico [2]. Otros autores han tomado la alternativa de utilizar técnicas innovadoras para intentar imputar valores nulos [9]. Adicionalmente, se observa que se ha intentado abordar el desbalance de clases utilizando métodos conjuntos de sobre o sub muestreo y métricas sensibles a la clase minoritaria [10] [9]

2) Clasificadores: En esta categoría se han implemenado modelos como Regresión Logística [11] [12], Redes Neuronales [11] [12] [9] [8], XGBOOST [12] [8], Máquina de Vectores de Soporte (SVM) [12], Bosques Aleatorios [12] y k-Vecindarios Más Cercanos [13] [10]. Esto puede deberse a que los modelos son adecuados para la clasificación de variables del tipo categórico. En general, se observa un éxito entre moderado y alto el cual puede depender de las variables que se eligen para realizar la predicción, donde individualmente se ha logrado sobrepasar métricas en más de 91%.

III. Diseño del Experimento

A. Descripción del conjunto de datos

El conjunto de datos contiene información sobre pacientes que han sufrido un infarto al miocardio y complicaciones asociadas a esta afección. De acuerdo a [2], este conjunto de datos contiene registros recolectados por el Hospital Clínico

Interdistrital de Krasnoyarsk, en Rusia, durante los años 1992-1995.

1) Número y Tipo de variables: El conjunto de datos consta de dos partes. La primera contiene variables que se pueden considerar "de entrenamiento", conformando un total de 110. Estas variables incluyen:

- 78 variables binarias.
- 16 variables categóricas.
- 9 variables numéricas discretas.
- 7 variables numéricas contínuas.

Algunas de las características más relevantes contienen información demográfica, antecedentes médicos (arritmias, insuficiencias cardiacas, entre otros), factores de riesgo cardiovascular e información de los primeros días de seguimiento al paciente post-infarto.

La segunda parte del conjunto de datos consta de variables que contienen información de diferentes tipos de complicaciones relacionadas a infarto en el miocardio. Estas variables se pueden considerar como "objetivos", y son un total de 12, siendo estas:

- 11 variables binarias.
- 1 variable categórica.

2) Número de muestras: El conjunto de datos contiene un total de 1700 registros. No existe una separación entre conjunto de entrenamiento y conjunto de prueba, pero esta separación es necesaria para cumplir con el objetivo del presente análisis, por lo que se va a realizar siguiendo la estrategia utilizada en la sección de metodología.

3) Número de muestras por clase: Este conjunto de datos contiene un total de 12 variables que se pueden considerar como objetivo, de las cuales 11 son binarias. La distribución de clases de cada una de estas variables binarias es:

TABLA I
NÚMERO DE MUESTRAS POR VARIABLE BINARIA

Variable	Clase = 0	Clase = 1
FIBR_PREDS	1530	170
PREDS_TAH	1680	20
JELUD_TAH	1658	42
FIBR_JELUD	1629	71
A_V_BLOK	1643	57
OTEK_LANC	1541	159
RAZRIV	1646	54
DRESSLER	1625	75
ZSN	1306	394
REC_IM	1541	159
P_IM_STEN	1552	148

Mientras tanto, la variable categórica tiene la siguiente distribución:

Por un lado, las variables de predicción del dataset sufren del problema del desbalance. Generalmente, se observa una proporción de alrededor 1 a 9 entre casos positivos y negativos (Tabla I), respectivamente. En base a esto, será necesario tomar en cuenta el problema de el desbalance de clases durante el entrenamiento para evitar el sesgo hacia la clase negativa.

Por otro lado, las variables de entrada contienen una gran cantidad de valores faltantes. De las 110 variables de entrada,

TABLA II
DISTRIBUCIÓN DE CLASES DE LA VARIABLE CATEGÓRICA LET_IS

Valor de LET_IS	Frecuencia
0	1429
1	110
3	54
7	27
6	27
4	23
2	18
5	12

11 de ellas presentan al menos 20% de valores faltantes (Tabla III).

TABLA III
VARIABLES CON MÁS DE 20% DE VALORES FALTANTES

Variable	Porcentaje de valores faltantes
IBS_NASL	95.7
S_AD_KBRIG	63.2
D_AD_KBRIG	63.2
GIPO_K	21.7
K_BLOOD	21.8
GIPER_NA	22.0
NA_BLOOD	22.0
KFK_BLOOD	99.7
NA_KB	38.6
NOT_NA_KB	40.3
LID_KB	39.8

4) Estadística descriptiva: Este conjunto de datos contiene información de 111 variables descriptivas, que brindan información detallada de varios aspectos relacionados al paciente. Para poder obtener información del conjunto de datos, se ha optado por utilizar un algoritmo de clustering sobre la base de correlaciones, con lo cual, se han agrupado las 111 variables en 5 categorías distintas. La figura 1 muestra una representación visual del método de clustering basado en la correlación entre variables utilizado.

De acuerdo a las variables agrupadas en cada clúster, se puede interpretar a grandes razgos a cada cluster de la siguiente manera:

- Clúster #1: (38 variables) Características generales del paciente, antecedentes médicos, hallazgos electrocardiográficos y manejo inicial en la unidad de cuidados intensivos.
- Clúster #2: (8 variables) Seguimiento de los síntomas y manejo del dolor en los primeros días de hospitalización.
- Cluster #3: (34 variables) Factores de riesgo cardiovascular, hallazgos clínicos y de laboratorio al ingreso, y manejo terapéutico inicial.
- Clúster #4: (27 variables) Antecedentes de arritmias e insuficiencia cardíaca, y hallazgos electrocardiográficos y clínicos relacionados al ingreso.
- Clúster #5: (4 variables) Niveles séricos de electrolitos clave (potasio y sodio).

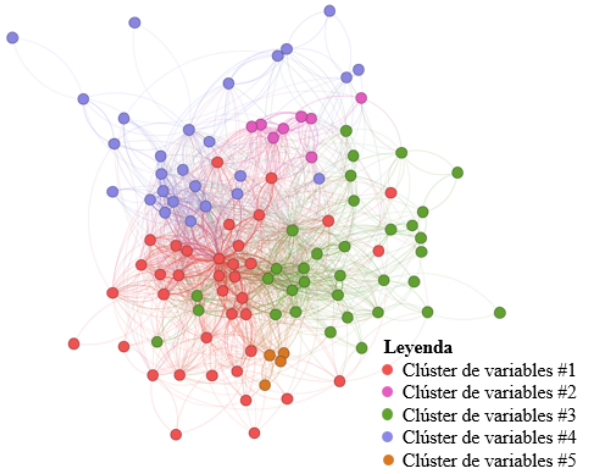

Fig. 1. Clustering de las 111 variables basado en correlaciones.

B. Metodología

El planteamiento del plan de acción está basado en tres aspectos de las características del dataset. A continuación se explica cada uno de ellos.

En primer lugar, respecto de las variables de predicción, el dataset contiene once variables objetivo de naturaleza binaria y una variable categórica que puede tomar 7 categorías diferentes. Por ello, se puede tratar el problema tanto como uno de clasificación con múltiples clases (multiclass) o múltiples etiquetas (multilabel). En tal medida, las clasificaciones de esta naturaleza se pueden abordar con los siguientes modelos

- Regresión logística (Esquema One Vs. The Rest para el caso multiclass).
- Árboles de decisión.
- Redes neuronales.
- Modelos de gradient boosting (XGBoost, LightGBM).

En segundo lugar, las etiquetas están fuertemente sesgadas por la clase negativa. Para abordar este problema se considera probar cada una de las estrategias clásicas: subsampleo, *k-folds* estratificado con validación cruzada, SMOTE y técnicas de *boosting* para árboles de decisión. Adicionalmente, por esta misma razón se entrenará los modelos con las métricas de AUC-PR y F1-beta como objetivo. Esto debido a que se caracterizan por ser sensibles a la clase minoritaria valorando el desempeño de los modelos tanto sobre la clase negativa como positiva [13].

En tercer lugar, es necesario abordar la gran cantidad de valores faltantes en las 11 variables mencionadas previamente. Al respecto, los autores consultados coinciden con eliminar las variables que presentan excesiva cantidad de faltos, cada uno por distintos criterios. A nuestro criterio, se ha decidido probar primero un punto máximo de 30% de faltos de acurdo con Govolenkin et. al. [2] y luego un punto de corte de 50%, planteado por Narayan et. al. [7].

Finalmente, se elegirá al mejor modelo basado en la comparación de las métricas. Luego, este será sometido a una búsqueda en grilla para hallar los mejores parámetros.

Referencias

[1] Ministerio de Salud del Perú-MINSA,"Muertes por infarto al miocardio-Perú," Dia mundial del Corazón, 2012. [En línea]. Disponible en: https://www.gob.pe/institucion/minsa/noticias/34838-al-ano-mas-de-4-mil-personas-mueren-por-infarto-en-el-peru.

[2] S. E. Golovenkin et al.,"Trajectories, bifurcations, and pseudo-time in large clinical datasets: applications to myocardial infarction and diabetes data", GigaScience, vol. 9, n.º 11. Oxford University Press (OUP), nov. 2020. doi: 10.1093/gigascience/giaa128.

[3] G. Quer, R. Arnaout, M. Henne, y R. Arnaout, "Machine Learning and the Future of Cardiovascular Care: JACC State-of-the-Art Review", J Am Coll Cardiol. 2021 Jan, 77 (3) 300–313. doi: 10.1016/j.jacc.2020.11.030

[4] J. Van den Eynde, M. Lachmann, K. Laugwitz, C. Manlhiot y S. Kutty, "Successfully implemented artificial intelligence and machine learning applications in cardiology: State-of-the-art review", Trends in Cardiovascular Medicine, Vol. 33, Issue 5, Pages 265-271, 2023. doi: 10.1016/j.tcm.2022.01.010.

[5] R. Sevakula, W. Au-Yeung, J. Singh, E. Heist, E. Isselbacher y A. Armoundas, "State-of-the-Art Machine Learning Techniques Aiming to Improve Patient Outcomes Pertaining to the Cardiovascular System", Journal of the American Heart Association. Vol. 9, Issue 4, 2020, Pages e013924. doi: 10.1161/JAHA.119.013924.

[6] H. Benzakour, C. Chekira, H. E. Fadili, y K. Zenkouar, "A State of Art of Cardiovascular Diseases Using Machine Learning Algorithms", 2023 7th IEEE Congress on Information Science and Technology (CiSt). IEEE, dic. 16, 2023. doi: 10.1109/cist56084.2023.10409886.

[7] Tiwari, R. (9 de febrero de 2023) "Advanced Evaluation Metrics for Imbalanced Classification Models". [En línea]. Disponible en: https://medium.com/cuenex/advanced-evaluation-metrics-for-imbalanced-classification-models-ee6f248c90ca

[8] A. Moore y M. Bell, "XGBoost, A Novel Explainable AI Technique, in the Prediction of Myocardial Infarction: A UK Biobank Cohort Study", Clinical Medicine Insights: Cardiology, vol. 16. SAGE Publications, p. 117954682211336, ene. 2022. doi: 10.1177/11795468221133611.

[9] M. Mesinovic y K.-W. Yang, "Multi-label Neural Model for Prediction of Myocardial Infarction Complications with Resampling and Explainability", 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE, sep. 27, 2022. doi: 10.1109/bhi56158.2022.9926915.

[10] A. Newaz, M. S. Mohosheu, y Md. A. Al Noman, "Predicting complications of myocardial infarction within several hours of hospitalization using data mining techniques", Informatics in Medicine Unlocked, vol. 42. Elsevier BV, p. 101361, 2023. doi: 10.1016/j.imu.2023.101361.

[11] D. A. Rossiev, S. E. Golovenkin, V. A. Shulman, y G. V. Matjushin, "Neural networks for forecasting of myocardial infarction complications", The Second International Symposium on Neuroinformatics and Neurocomputers. IEEE. doi: 10.1109/isninc.1995.480871.

[12] R. Ghafari et al., "Prediction of the Fatal Acute Complications of Myocardial Infarction via Machine Learning Algorithms", The Journal of Tehran University Heart Center. Knowledge E DMCC, ene. 30, 2024. doi: 10.18502/jthc.v18i4.14827.

[13] Narayan et. al. (2023) "Myocardial Infarction Complications SDS322E - Final Project". [En línea]. Disponible en: https://github.com/sarvagnyanarayan/myocardial-infarctions