Predicción de muertes por enfermedades cardiovasculares

Javier Alberto Ibáñez Bolaños – ibanez.ja@javeriana.edu.co
Pontificia Universidad Javeriana - Bogotá D.C.

Resumen – El presente documento tiene como objetivo presentar los algoritmos de aprendizaje de regresión logística, Knn, Naive Bayes y árboles de decisión con los cuales se pretende estimar los eventos de muerte generados por enfermedades cardiovasculares (ECV) que son la principal causa de fallecimiento a nivel mundial según la OMS, para esto se ha utilizado un dataset que contiene 12 características medicas de 299 pacientes, registrando cuantos de ellos han muerto en periodo de estudio. Con lo cual se desea que el algoritmo pueda predecir bajo que condiciones una persona puede estar en riesgo de muerte.
Keywords*:* Regresión logística, Knn, Naive Bayes, árboles de decisión, ECV, dataset, predecir, riesgo.

I. BUSINESS UNDERSTANDING

Las enfermedades cardiovasculares (ECV) son un grupo de trastornos del corazón y de los vasos sanguíneos, se generan por un decrecimiento u obstrucción del flujo de sangre en las venas y arterias del cuerpo. Estas enfermedades son la principal causa de muerte a nivel mundial, cobrando un estimado de 17.9 millones de vidas cada año según la OMS, duplicando por ejemplo la mortalidad generada por todos los tipos de cáncer unidos, lo que representa el 31% de todas los fallecimientos en el mundo.

Por lo anterior, se requiere un sistema de predicción el cual permita estimar el riesgo de muerte de las personas a partir de un conjunto de características medicas de entrada, las cuales deben ser seleccionadas con precisión y por personal experto, quienes conocen los principales factores que producen las enfermedades y de esta manera obtener una base de datos lo suficientemente robusta, con la cual es posible llevar a cabo un modelo de aprendizaje de clasificación supervisado que permite contribuir a la prevención de eventos de muerte producidos por las *ecv*. Para lo cual se realizarán diversos métodos de clasificación como *Logistic regression*, *Knn*, *Naive bayes* y *decisión tree* que serán evaluados por medio de métricas como *Mcc*, *F1 score* y *Auc-Roc* con el objetivo de determinar el mejor modelo para la prevención de muertes producidas enfermedades cardiovasculares.

II. DATA UNDERSTANDING

La base de datos usada para este proyecto "*Heart Failure Prediction*" fue extraída de la plataforma de *Kaggle* [1] la cual cuenta con 12 características de entrada y una etiqueta de salida tal y como se puede observar a continuación en la tabla 01.

Características			**Etiqueta**
Edad	Anemia	Creatinina_f	Evento de muerte
Diabetes	Eyección	Presión arterial	
Plaquetas	Creatinina_s	Sodio	
Sexo	Tabaquismo	Tiempo	

Tabla 01. Estructura de la base de datos.

Los valores dados en el *dataset* para los parámetros mostrados en la tabla 1 son en total 299, datos médicos que al ser de pacientes son diferentes, ya que para el caso de características como anemia, diabetes, presión arterial alta y tabaquismo tienen valores binarios donde un "1" representa que hay presencia de este factor en el paciente mientras que un "0" es lo opuesto, así mismo para el sexo un "1" representa que la persona es un hombre mientras que un "0" se refiere a una mujer, además para el caso de la etiqueta de evento de muerte un "1" representa fallecimiento y un "0" que no lo hay. Mientras que, características como la creatinina de ambos tipos, fracción de eyección, plaquetas, sodio y tiempo en días están dadas en su convención natural por lo que la normalización de los datos es un factor muy importante antes de realizar los modelos de aprendizaje.

Por otra parte, del conjunto de datos se pudo extraer información importante que se puede contrastar con el resultado de los modelos posteriormente. Por ejemplo, el estudio cuenta con 194 hombres y 105

mujeres los cuales tienen una edad promedio de 60 años con un tiempo de seguimiento promedio de 130 días y un gran porcentaje de ellos cuentan con condiciones fáciles de medir como las que se muestran en la tabla 02.

Condición	Anemia	Diabetes	Presión arterial	Tabaquismo
Pacientes	129	125	105	96

Tabla 02. Pacientes con condiciones especiales.

Finalmente, con el objetivo de analizar todo el conjunto de datos y ver la relación entre ellos se realizó el cálculo del coeficiente de correlación, el cual permite estimar la relación entre dos variables donde un valor de 1 representa una asociación positiva muy fuerte, -1 representa una asociación negativa muy fuerte y valores cercanos a 0 representan que la asociación es muy débil tendiendo a ser nula entre más cercano este a cero el coeficiente.

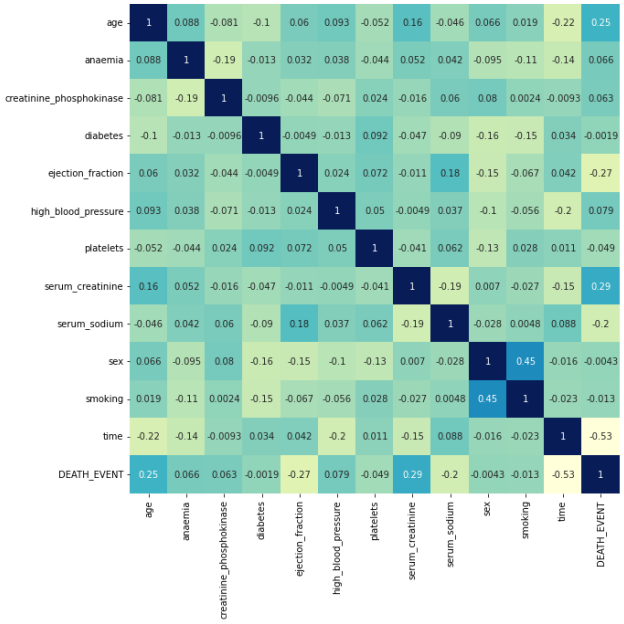

Figura 01. Correlación del dataset.

La figura 01 permite observar la correlación entre las variables del conjunto de datos, donde la tendencia es que no hay una correlación fuerte entre la mayoría de los pares. Sin embargo, en algunos casos como el de la edad y la muerte, la fracción de eyección y el sodio, la creatinina y la muerte se observa una correlación fuerte en relación con las demás.

III. DATA PREPARATION

Para el caso específico de este proyecto no fue necesario un proceso extenso de preparación de los datos del dataset como en otros casos, ya que la base de datos con la cual se trabajo estaba ajustada de manera correcta. La única adecuación que se llevó a cabo al conjunto de datos para empezar a desarrollar los diversos modelos de aprendizaje fue la conversión de tipos de datos, ya que para el caso de la característica de años se tenían algunos pocos valores de tipo flotante, cuando la mayor parte estaba en entero por lo que se realizó el ajuste de todos los datos a tipo entero. Posteriormente, se creó un dataset nuevo ya con el ajuste realizado debido a que el conjunto no presento ningún otro tipo de anormalidad como valores tipo NaN o nulos. Además de que debido a la complejidad del problema en cuestión no es sencillo extraer nuevas características a partir de otras, esto se debe a que por ejemplo la correlación entre las variables no permite determinar un nuevo parámetro de utilidad y que otros parámetros que pueden ayudar a predecir la muerte debido a ataques cardiovasculares deben ser producto de mediciones clínicas a los pacientes.

IV. MODELLING

Los modelos de aprendizaje implementados para este proyecto son Regresión logística, KNN, Naive bayes y Árboles de decisión, para estos cuatro modelos se realizó la división del conjunto de datos de entrenamiento y prueba en 75% y 25% respectivamente, además de realizar la normalización de todos los datos con la librería de *sklearn.preprocessing*. Por otra parte, para el caso del modelo de regresión logística se llevó a cabo el análisis por componentes principales (PCA), con el objetivo de observar el comportamiento del modelo ante la reducción de características manteniendo una varianza mayor a 90% debido a que la recomendación general de ser mayor al 97% no se cumple para el data set aun disminuyendo solo una característica. Así mismo, es importante resaltar que las métricas de evaluación para todos los modelos fueron el MCC, F1 score, Accuracy y Auc-Roc con los cuales se pretende establecer el modelo que mejor se ajusta al problema en cuestión.

- **Regresión logística**

La regresión logística aplicada al conjunto de datos se realizó haciendo uso se la librería *sklearn.linear_model* e importando la función de *LogisticRegression*, con lo cual se tomó un *random_state* de 0 para que todas las compilaciones llevadas a cabo siempre den igual, así mismo se tomó una regularización con penalización L2 con el parámetro (*penalty='l2'*) ya que esto permite evitar el sobre ajuste del modelo, para lo cual se utiliza un solucionador que admita esta penalización siendo para este caso *solver='lbfgs'* y con un máximo de iteraciones de 100k en las cuales se espera que converja el modelo. Por otra parte, como había sido mencionado en apartados anteriores para el caso de la regresión logística se aplicó PCA con el objetivo de comparar la respuesta del modelo en sus métricas de evaluación, con lo cual al reducir la características a de 12 a 11 se tiene una varianza de aproximadamente 96% y con 10 características se tiene un varianza cercana a 90%. Teniendo en cuenta lo anterior, la tabla 03 presenta las métricas de evaluación obtenidas con 12, 11 y 10 características del conjunto de datos.

Características	MCC	Accuracy	F1 score	Auc-Roc
12	0.66	0.86	0.75	0.90
11	0.66	0.86	0.75	0.89
10	0.59	0.84	0.70	0.87

Tabla 03. Métricas de evaluación para regresión logística.

- **KNN**

El modelo de aprendizaje KNN fue llevado a cabo mediante la función *KNeighborsClassifier* de la librería *sklearn.neighbors*, donde en los parámetros se realizó la configuración de pesos uniforme con *weights='uniform'* así como también, la distancia escogida mediante *metric='minkowski'* y un algoritmo para calcular los vecinos más cercanos de fuerza bruta con *algorithm='brute'*. De igual manera, el parámetro más importante del algoritmo que es el K se itero en un rango de 1 a 50, con el objetivo de escoger el valor optimo en el cual se obtiene la mayor precisión con el menor K posible, con lo cual se utiliza el coeficiente de silueta, obteniendo así el valor de *K=22* tal y como se puede observar en la figura 02 a continuación.

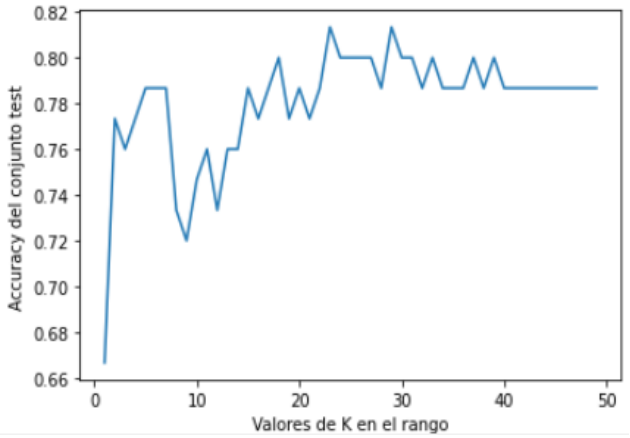

Figura 02. Valor de K por coeficiente de silueta.

Teniendo en cuenta lo anterior, se realizó la evaluación del modelo de aprendizaje con las mismas métricas que para los otros modelos, obteniendo los resultados de la tabla 04.

MCC	Accuracy	F1 score	Auc-Roc
0.38	0.79	0.38	0.86

Tabla 04. Métricas de evaluación para KNN.

- **Naive bayes**

El algoritmo de aprendizaje Naive bayes fue llevado a cabo mediante la variante de Gauss, usando la función *GaussianNB* de la librería *sklearn.naive_bayes* y aunque es un método paramétrico no tiene hiperparámetros que iterar, con lo cual solo se realiza el ajuste de los conjuntos de entrenamiento y prueba al modelo que se dividen en 75% y 25% respectivamente, para posteriormente hacer uso de la función que es común a todos los modelos de *predict* con la que se obtiene el resultado del modelo y a partir de este resultado obtenido de los dos conjuntos de datos se evalúa el modelo, teniendo los resultados mostrados en la tabla 05.

MCC	Accuracy	F1 score	Auc-Roc
0.53	0.83	0.75	0.74

Tabla 05. Métricas de evaluación para Naive Bayes G.

- **Árboles de decisión**

Finalmente se realizó el modelo de aprendizaje de árboles de decisión haciendo uso de la función *DecisionTreeClassifier* de la librería *sklearn.tree*, con un *random_state*=0 para fijar la semilla y obtener los mismos valores siempre en el modelo. De igual manera se estimó la profundidad máxima del árbol en un rango de 1 a 50, buscando el valor para el cual se obtuviese el máximo MCC del conjunto de prueba con lo cual se obtuvo un valor de 2 obteniendo el árbol mostrado a continuación en la figura 03.

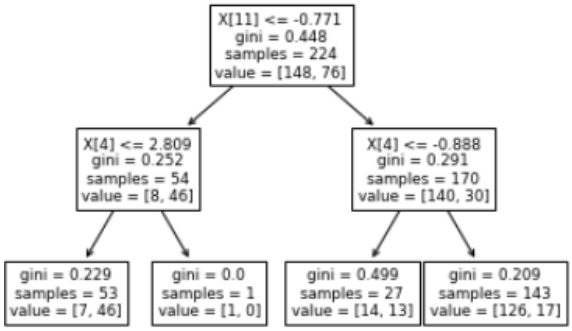

Figura 03. Árbol de decisión del modelo implementado.

Por otra parte, al igual que en el resto de los modelos mostrados anteriormente, se realizó la evaluación de este obteniendo las métricas mostradas en la tabla 06.

MCC	Accuracy	F1 score	Auc-Roc
0.68	0.88	0.75	0.82

Tabla 06. Métricas de evaluación para Decision Tree.

- **Classification Learner**

Finalmente, con el objetivo de poder realizar en la sección posterior una comparación y evaluación final de todos los modelos desarrollados, a continuación se presentan algunos de los resultados obtenidos con la herramienta *Classification Learner* de Matlab, en la cual se pueden estimar a priori los mejores modelos a implementar y se puede tener una noción de a partir de los resultados obtenidos que oportunidades de mejora se tiene en los modelos de aprendizaje. La tabla 07 a continuación.

Método		Accuracy (%)
Tree	Fine	79.7
	Medium	79.7
	Coarse	75.7
Naive Bayes	Gaussian	73.0
	Kernel	73.0
SVM	Linear	79.7
	Quadratic	71.6
	Cubic	74.3
Gaussian SVM	Fine	67.6
	Medium	75.7
	Coarse	73.0
KNN	Fine	67.6
	Medium	70.3
	Coarse	67.6
Ensemble	Random Forest	82.4
LG	Logistic Regression	81.1

Tabla 07. Accuracy con Clasification Learner.

Tal y como se puede observar en la tabla 06, el modelo que implementa arboles de decisión tiene el mayor Accuracy seguido por la regresión logística por un poco, resultados que también fueron dados en los apartados anteriores con un 0.88 y 0.86 respectivamente.

V. EVALUATION

Una vez implementados todos los modelos de aprendizaje con sus métricas de evaluación respectivas, se puede llevar a cabo una comparación para estimar cuál de ellos es el más adecuada para el problema en cuestión, con lo cual a continuación la tabla 08 presenta un resumen de las métricas de evaluación obtenidas para cada modelo.

Modelo	MCC	Accuracy	F1 score	Auc-Roc
Regresión logística	0.66	0.86	0.75	0.90
KNN	0.38	0.79	0.38	0.86
Naive Bayes	0.53	0.83	0.75	0.74
Árboles de decisión	0.68	0.88	0.75	0.82

Tabla 08. Comparación de los modelos.

Teniendo en cuenta los resultados mostrados en la tabla 08 se puede reafirmar el resultado dado mediante la herramienta de Matlab en la tabla 07, donde se obtuvo que la precisión más alta está en los árboles de decisión seguido por la regresión logística, posteriormente por el modelo gaussiano de Naive bayes y finalmente los de KNN, con lo cual se contrasta el resultado coherente de los resultados. Por otra parte, si bien el Accuracy permite estimar a priori los modelos más precisos no es la mejor métrica de evaluación para estos casos ya que depende del balanceo del dataset, por lo que considerando los resultados del MCC y la curva Roc se estima que el mejor modelo a implementar para resolver el problema es la regresión logística ya que con esta se obtienen de manera general las mejores métricas de evaluación, además de presentar la ventaja de que al someterse a un análisis de componentes principales bajan en una pequeña proporción las métricas, tal y como se pudo observar en la tabla 03 del presente documento.

VI. REFERENCIAS

[1] S.N, (2020). *Heart Failure Prediction* [Dataset]. Disponible en: https://www.kaggle.com/andrewmvd/heart-failure-clinical-data