Predicción de ataques al corazón : usando Machine Learning

Jesús Emmanuel Ramos Dávila

***Resumen*—Enfermedades cardiovasculares son una de las principales causas de muerte global, con millones de fallecimientos año tras año, Las enfermedades cardiovasculares son un grupo de desórdenes del corazón y los vasos sanguíneos no existe causa exacta para este tipo de enfermedades. Existen algunos patrones de síntomas los cuales están muy asociados a tener este tipo de enfermedades. Actualmente no existen demasiadas investigaciones con respecto a el análisis de grupos y algoritmos de asociación para este tipo de enfermedades. En esta sección se realizara un estudio usando el algoritmo FCM (Fuzzy C Means) Clustering, para determinar el riesgo de un ataque al corazón, en este estudio se realizara un prueba usando 303 observaciones se revisara el performance y la precisión con respecto a otro algoritmo ya conocido.**

I. Introducción

Enfermedades cardiovasculares son una especie que se están presentando con mayor frecuencia y estas frecuentemente suceden en fallecimientos. La Organización mundial de la salud ha estimado alrededor de 12 millones de muertes alrededor del mundo anualmente, debido a este numero avances en la medicina en las últimas décadas habilito la identificación de factores de riesgo que podrían contribuir en este tipo de enfermedades cardiovasculares. La causa más común en este tipo de enfermedades es el estrechamiento o bloqueo de las arterias coronarias. Los vasos que transportan sangre al corazón mismo, Este es llamado enfermedad arteriopatía coronaria y esta sucede comúnmente con el paso del tiempo. Esta es una de las principales causas por las cuales las personas sufren ataques al corazón, Es por eso que un bloqueo que no es tratado dentro de las primeras horas causa que el musculo del corazón muera. Diagnósticos médicos son una importante, pero a la vez una tarea complicada y su automatización podría ser muy útil. Desafortunadamente no todos los doctores están especializados en esta área y no se tiene en algunos casos y no se tienen los mismos recursos médicos. Es por eso que para utilizar el conocimiento de diferentes especialistas y los datos clínicos de pacientes para facilitar el proceso de diagnóstico es considerado muy valioso ya que su integración en tomas de decisiones medicas podría reducir los errores médicos , mejorar la seguridad del paciente y reducir practicas no deseadas.

II. Metodología

II-A. Fuzzy C Means Clustering

El principal objetivo de esta búsqueda es implementar el algoritmo de Fuzzy C Means Clustering usando nuestros datos de pacientes, esto para poder usarse en el soporte de toma de decisiones, por lo tanto, se desarrolla un modelo Fuzzy C Means para predecir los ataques al corazón. El algoritmo

1. Initialize the matrix of μ_{ij}, U
2. Update c_i:

$$c_i = \frac{\sum_{j=1}^{N} \mu_{ij}^{m} x_j}{\sum_{j=1}^{N} \mu_{ij}^{m}}$$

3. Update μ_{ij}:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_j - c_i\|}{\|x_j - c_k\|}\right)^{\frac{2}{m-1}}}$$

4. If the change of U between two iterations is very small (predefined cutoff), then stop; otherwise, go to step #2.

Algorithm of FCM (Image by author)

Figura 1: Imagen de pasos algoritmo FCM [1]

Fuzzy C-Means Clustering fue desarrollado en 1981 este es un extendido del algoritmo K-Means Clustering, FCM (Fuzzy C-Means) es un algoritmo no supervisado que es aplicado hacia un rango muy amplio de problemas conectados con análisis de características , clustering y clasificación. FCM también es usado en otros campos además de la medicina, tales como agricultura, ingeniería, astronomía, química, análisis de imágenes. FCM es una técnica de clustering en la cual un conjunto de datos es agrupado en n-clústeres en la cual cada punto de datos esta relacionado a un clúster el cual tendrá un alto grado de pertenencia hacia este punto siendo así que los puntos de datos que tengan un bajo grado de pertenencia hacia este clúster estarán más alejados de este.

II-B. Linear Regression

Para este modelo con datos de enfermedades del corazón, la regresión lineal intentara crear una relación entre dos o mas variables al ajustar una ecuación lineal hacia los datos observados. Antes de intentar ajustar un modelo lineal a nuestros datos, se debe de identificar cuales variable tienen relación con nuestra variable de respuesta. Se usará la tabla de datos de el trabajo realizado [2] por de selección de características para trabajar con solamente con un subconjunto de datos que se relacionen con nuestra variable de respuesta .

Modelo	Variable	Variables presentes en Análisis clinic
$AnovaF-value$	$exgn, cp, thalla, oldpeak$	si
$Umbraldevarianza$	$age, cp, oldpeak, slp$	si
RFE	$sex, exng, slp, thall$	si
$Informaciónmutua$	$thall, oldpeak, slp, age$	si
$Seleccióndecaracteristicasexhaustiva$	$cp, slp, caa, thall$	si
PCA	$PCA_1 = 0.7550 * exng + 0.4198 * sex + 0.1460 * oldpeak + 0.1247 * caa$	si

Figura 2: Selección de características

La regresión lineal se puede representar mediante una ecuación lineal de la forma.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\epsilon}_i \quad (1)$$

Dado que nuestra variable de respuesta es una variable categórica se pretende que la línea forme un segmento discriminante en el cual una parte de los datos estén segmentado 2 por regiones las cuales se visualizaran por la línea ajustada.

Siguiendo el ejemplo de ecuacion lineal anterior y basandonos en la figura 2 tenemos nuestra ecuacion lineal.

II-B1. Comparacion con otros modelos: El modelo ya mencionado se puede comparar con otros modelos que desempeñan otro tipo de características como Lasso Regression, este modelo usa un técnica de contracción, Contracción es donde los datos se encogen hacia un punto central de la media, Este tipo de modelos son muy utilizados cuando se muestran en nuestros datos un alto nivel de multicolinealidad o cuando la varianza en nuestros datos es muy grande. Otra de las ventajas de este modelo es su capacidad para la selección y eliminación de variables que aportan o no a nuestro modelo. Lasso Regression usa un tipo de técnica llamada regularización la cual es implementada al agregar un término llamada penalización la cual busca lograr la menor varianza en nuestro conjunto de datos de prueba.

II-B2. Métrica de evaluación: Métrica de evaluación (Mean absolute error) En el contexto de una regresión lineal, el error absoluto se refiere a la magnitud de la diferencia entre la predicción de una observación y el verdadero valor de la observación. MAE toma el promedio de errores absolutos de un grupo de predicciones y observaciones como una medida de la magnitud de errores del grupo entero. Una de las propiedades que tiene esta métrica de evaluación es que es fácil de interpretar ya que el valor esta en la misma escala al valor que estamos realizando predicción. Interpretación de MAE: su interpretación es simple ya que si su unidad esta mas cerca de el valor 0 mas preciso es el modelo.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 exng + \hat{\beta}_2 cp + \hat{\beta}_3 thalla + \hat{\beta}_4 caa + \hat{\beta}_5 thall + \hat{\epsilon}_i \quad (2)$$

III. Adquisición de datos

Dentro de este análisis el conjunto de datos es obtenido de UC Irvine [3] , Los datos han sido recolectados de 303 pacientes de un subconjunto conjunto de datos que contiene 13 columnas y su variable de respuesta la cual es si el si el paciente presenta enfermedad del corazón o no. vease cuadro: I.

IV. Resultados

IV-A. Fuzzy C Means Clustering

Se aplico el modelo FCM usando las 303 observaciones a fin de encontrar los mejores clústeres que representen nuestros datos.

Los dados obtenidos en la métrica de evaluación fueron **2** clústeres como los óptimos en una iteración de 1 hasta 9 clústeres, representado claramente en con nuestra grafica de codo. Figura 3 La métrica de evaluación para este modelo fue su **coeficiente de particiones de fuzzy (FCP)** la cual marco un notable porcentaje en 2 clusters.

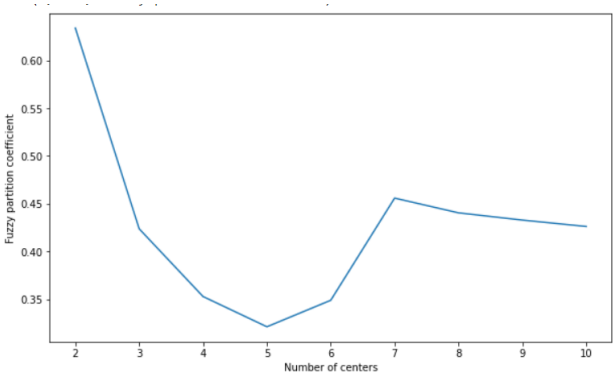

Figura 3: Grafica creada en el analisis de nuestro

IV-B. Regresión Lineal

Se aplico un modelo de regresión lineal el cual obtuvimos una regresión moderada usando nuestros datos, se considero aplicar un modelo de regresión simple ya que solo seleccionamos un conjunto de variables consideradas en el trabajo de selección de características 2, con respecto a los resultados de nuestra métrica de evaluación (MAE) se consideró un valor muy bueno de **0.29** lo cual nos habla de posibles iteraciones en modelos específicos a fin de incrementar nuestro valor de métrica usado.

V. Discusión

V-A. Fuzzy C Means Clustering

Se observo un buen desempeño con respecto a el calculo del mejor número de clústeres siendo de 300ms aproximadamente sin la funcionalidad del multithreading. Los valores de las métricas que usa el modelo FCM nos permitió claramente ver el número de clústeres. Una de las comparaciones a realizar a futuro seria comparar con un algoritmo tradicional tal como K-Means o K-Mediods a fin de comparar tiempo y si es tan notable el encuentro del mejor clúster.

V-B. Linear Regression

Se observo un buen desempeño usando un modelo de regresión simple usando ciertas características de nuestro conjunto de datos, la única parte que seria una parte de análisis es saber si usando mas características con modelos que discriminan y realizar una selección de características obtenemos mejores resultados o al menos saber si nuestro subconjunto de datos es el optimo para nuestro modelo que se ejecutó.

[4] [5] [6] [**?**] [7]

Referencias

[1] Yufeng, "Fuzzy c-means clustering with python," Medium, 11 2021. [Online]. Available: https://towardsdatascience.com/fuzzy-c-means-clustering-with-python-f4908c714081

[2] E. Ramos, "Emmanuelramos143/aa," GitHub, 01 2023. [Online]. Available: https://github.com/EmmanuelRamos143/AA/blob/main/Tarea4-FeatureSelection.ipynb

[3] A. Janosi, "UCI machine learning repository," 2017. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[4] G. Banu, M. Phil, J. Bousal, and J. Mca, "Perdicting heart attack using fuzzy c means clustering algorithm," International Journal of latest Trends in Engineering and Technology, vol. 5, 05 2015.

Cuadro I: Descripción de las 13 variables dentro del dataset usado en los modelos

Proposal	Condition	Identification
Sex	Gender	Male/Female
Age	Age of pacient	20-90
cp	Tipo de dolor de pecho	0 : Angina tipica 1: Angina atipica 2: non-anginal pain 3: asintomático
trtbps	Azucar en ayunas >120	290
chol	Colesterol en mg dl	valor numerico
rest ecg	Resultado electrocardiograma en reposo	0 : normal 1:ST-T wave abnormality 2: ventricular hypertrophy by Estes criteria
thalach	maximum heart rate achieved	valor numerico
exang	Angina de Pecho Inducida	1:Yes 0: No
old peak	ST depression induced by exercise	valor numerico
slp	slope peak ST segment	0 = unsloping 1 = flat 2 = downsloping
caa	number of major vessels	0-3
thall	thalassemia	0 = null 1 = fixed defect 2 = normal 3 = reversable defect

†

[5] Z. Rustamov, Clustering and Association Rule Mining of Cardiovascular Disease Risk Factors. www.researchgate.net, 01 2023, pp. 389–396.

[6] C. Fuster-Barceló, C. Cámara, and P. Peris-López, "Unleashing the power of electrocardiograms: A novel approach for patient identification in healthcare systems with ecg signals," arXiv:2302.06529 [cs], vol. 2302.06529, 02 2023. [Online]. Available: https://arxiv.org/abs/2302.06529

[7] S. Allwright, "How to interpret mae (simply explained)," Dec 2022. [Online]. Available: https://stephenallwright.com/interpret-mae/