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 desempeño 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 [WHO(2022)], 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 Fuzzy C-Means Clustering fue desarrollado en 1981 este es un extendido del algoritmo *K-Means Clustering* [Banu et al.(2015)Banu, Phil, Bousal, and Mca]. FCM (*Fuzzy C-Means*) es un algoritmo no supervisado que es aplicado hacia un rango muy amplio de problemas conectados con análisis de características , agrupamiento 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 agrupamiento en la cual un conjunto de datos es agrupado en n-clústeres en la cual cada punto de datos esta relacionado a un grupo 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 grupo estarán más alejados de este.

Pasos de algoritmo FCM

1. Inicializar la matriz de $\mu_{i_j} U$.
2. Actualizar C_i.
   $C_i = \frac{\sum_{i=1}^{N} \mu_{i_j}^m X_j}{\sum_{j=1}^{N} \mu_{i_j}^m X_j}$
3. Actualizar μ_{i_j}.
   $\mu_{i_j} = \frac{1}{\sum_{k=1}^{C} (\frac{||X_j - C_i||}{||X_j - C_k||})} \frac{2}{m-1}$
4. Si el cambio de U entre las dos iteraciones es muy bajo, entonces se detiene en otro caso sigue el paso 2.

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 [Ramos(2023)] 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 1: 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 ecuación lineal anterior y basándonos en la figura 1 tenemos nuestra ecuación 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.

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

II-C. Clasificación

Los modelos de clasificación buscan predecir categorías a las que pertenece cierta etiqueta dados un conjunto de variables dependientes. Para nuestro conjunto de datos los modelos de clasificación son la mejor opción ya que trataremos de predecir si cierto paciente presenta una enfermedad del corazón, Los modelos evaluados en esta sección fueron.

1. SVM Support Vector Machine
   SVM ha sido ampliamente utilizado como un poderoso método de machine learning en diferentes problemas de clasificación incluyendo bioinformatica. El modelo trata de buscar un óptimo hiper plano el cual maximice la distancia de el punto de datos de entrenamiento mas cercano. El hiper plano de el modelo SVM maximiza el margen mientras que minimiza el error de clasificación. El margen es computado como la suma de las distancias a unas de las mas cercanas positivas, sobre este modelo se encontró una optimización de un modelo para predecir ataques al corazón usando como base SVM [Ali and Niamat(2019)].
2. Random Forest classifier
   El método Random Forest es un estimador que ajusta diferentes arboles de clasificación en diferentes subconjuntos de el dataset y con el uso de la media para mejorar la predicción y controlar el sobre entrenamiento del modelo. este método fue elegido por trabajos previos con un buen despeño del modelo [Pal and Parija(2021)].

III. Adquisición de datos

Dentro de este análisis el conjunto de datos es obtenido de UC Irvine [Janosi(2017)], 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. véase cuadro: II.

IV. Diseño de Experimentos

El Diseño de Experimentos, es una de las tecnicas estadisticas que se basa en estudiar el efecto de uno o mas factores sobre la media de una variable continua, Las suposiciones que se plantean es:

$$H_0 : \mu_1 = \mu_2 ... = \mu_i \quad (3)$$

$$H_1 : \mu_i \neq \mu_j \quad (4)$$

En este caso se planteará la hipótesis nula de que las medias son iguales, mientras que en la alternativa se plantea que al menos una media dentro del grupo es diferente. El estatistico de prueba a usar sera un ANOVA el cual tiene que cumplir ciertos criterios para aplicarse, Los cuales son:

1. **Independencia** Se asume que las muestras son independientes y aleatorias.
2. **Distribución normal** Indica los datos siguen una distribución de tipo normal
3. **Homoscedasticidad** Misma varianza entre grupos

Dado que nuestra variable de respuesta es de tipo binomial , en este caso no cumplimos el criterio de distribución normal de cualquier manera se realizara la prueba para **3** modelos de clasificación ya vistos en la sección II Clasificación.

IV-A. Resultados

Se observo que el desempeño de nuestros modelos no fue significativo aunque el modelo tenia una ligero menor desempeño que el primer modelo, **ANOVA** no arrojo un p-valor de **0.6616** mayor que alfa **0.05**, con lo cual no procedemos a rechazar **Ho** por lo tanto no existe diferencia entre el desempeño de modelos.

V. Resultados

En esta sección se incluyeron los resultados del compendio de modelos utilizados tanto **Agrupamiento , Clasificación y Regresión**, En el cual la mayoría de los modelos utilizados se encontraron artículos relacionados a este tipo de predicción de ataques al corazón, También se incluyeron notas para futuras revisiones para nuestro modelo a fin de mejorar nuestra predicción. En resumen obtuvimos desempeños significativos en los tres tipos de algoritmos antes mencionados.

V-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.

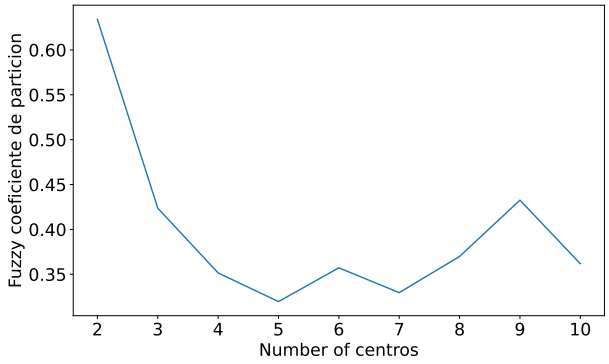

Figura 2: Grafica creada en el analisis de nuestro Fuzzy C Means

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 2 La métrica de evaluación para este modelo fue su **coeficiente de particiones de fuzzy (FCP)** la cual marcó un notable porcentaje en 2 grupos.

V-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 1, 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-C. *Clasificación*

Se aplicaron dos modelos de clasificación de los cuales se encontró trabajos relacionados con predicción de ataques al corazón [Pal and Parija(2021)] [Ali and Niamat(2019)], para nuestro conjunto de datos se aplico un preprocesa-miento usando PCA tomando solo los tres principales componentes y aplicándolos a cada uno de los modelos. Nuestro modelo tuvo un **accurary** del 83%, el modelo con mejor desempeño fue **Support Vector Classifier (SVC)**.

Los resultados de los modelos fueron.

Cuadro I: Modelo usado y porcentaje de precisión.

Modelo	Precisión
SVM classifier	83.60%
Random Forest classifier	80.32%

Dada la precisión de los modelo podemos observar que el modelo SVM tuvo una mejor precisión.

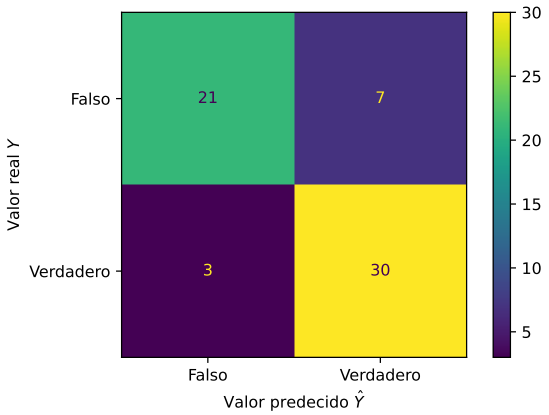

Figura 3: Matriz de confusión para nuestro modelo SVC

VI. Discusión

VI-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.

VI-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 óptimo para nuestro modelo que se ejecutó.

VI-C. *Clasificación*

En esta sección se aplicaron 2 modelos de clasificación basados en investigaciones realizadas y cada uno de ellos obtuvo un buen desempeño, para nuestro caso aplicamos una parte de pre-procesamiento con PCA el cual no arrojo un mejor modelo de 83.6%, una de las partes que se puede mejorar en este modelo es agregar nuevas variables al modelo con tal de mejorar el porcentaje de precisión. Una de las partes que no se comento es que computacional mente este modelo con mejor desempeño es mas robusto, siendo así podríamos realizar una nueva iteración teniendo en cuenta el modelo *Random Forest* como opción. Dado que la predicción de ataques al corazón debe de ser sensible a los errores de **tipo 2**, este modelo redujo estos tipo de errores al solo clasificar incorrectamente tres observaciones. Un tipo de mejora al desempeño de este modelo podría ser incluir mas componentes de PCA a fin de tener mas varianza explicada y aumentar nuestro porcentaje de desempeño.

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

Variable	Condición	Identificación
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

†

REFERENCIAS

[WHO(2022)] WHO, "Cardiovascular diseases," World Health Organization, 2022. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[Banu et al.(2015)Banu, Phil, Bousal, and Mca] 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.

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

[Ali and Niamat(2019)] L. Ali and Niamat, "An optimized stacked support vector machines based expert system for the effective prediction of heart failure," IEEE Access, vol. 7, pp. 54 007–54 014, 2019.

[Pal and Parija(2021)] M. Pal and S. Parija, "Prediction of heart diseases using random forest," Journal of Physics: Conference Series, vol. 1817, no. 1, p. 012009, mar 2021. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/1817/1/012009

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

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

[Fuster-Barceló et al.(2023)Fuster-Barceló, Cámara, and Peris-López] 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

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