Análisis de Algoritmo SVM para Predecir Enfermedades Cardiovasculares en Tiempos de COVID-19

Alfaro Zapana Estefany Reyna[1]
Escuela Profesional Ingeniería de Sistemas
Universidad Nacional de San Agustín
ealfaroz@unsa.edu.pe

Vera Mamani José Miguel[2]
Escuela Profesional Ingeniería de Sistemas
Universidad Nacional de San Agustín
jverama@unsa.edu.pe

***Resumen*—El padecimiento de una enfermedad al corazón puede llegar a ser muy trágica en la vida de las personas hoy en día, ya que están clasificadas como personas vulnerables ante el COVID 19. Las causas pueden ser variables desde la forma hereditaria hasta una vida no saludable.**

***Keywords*— enfermedades corazón, COVID-19, daño cardiaco**

I. Introducción

La infección por el síndrome respiratorio severo del adulto (SARS CoV-2) ha ocasionado un incremento de personas afectadas y más de un millón de fallecidos en todo el mundo. En esta última semana solo en Perú hay más de 33 millones de muertes. A pesar de todos los recursos para combatir la pandemia, en muchos países se ha comenzado a observar nuevas variantes del coronavirus que está causando temor en la población.

La enfermedad puede variar en gravedad desde pacientes asintomáticos, a ser tan leve como un resfriado común o conducir a un proceso de neumonía y síndrome de distrés respiratorio del adulto, insuficiencia respiratoria y falla multiorgánica.Del compromiso no respiratorio relacionado a COVID-19, se destaca el daño miocárdico que puede alterar el curso de la enfermedad y puede ser un indicador del pronóstico. [1]

Los mecanismos del compromiso cardíaco relacionado a COVID-19, pronostican un daño miocárdico y puede producir complicaciones cardiacas indirectas de la pandemia en la población con enfermedades cardiovasculares.

En esta investigación hablaremos sobre las enfermedades del corazón comúnmente vistas y las causadas por COVID-19, algunos datos médicos de padecimiento, algoritmo a emplear, data que será cargada para entrenar a nuestro algoritmo, cuán importante es esta investigación en el mundo actual, resultados del proyecto y trabajos a futuro para una mejora.

II. Trabajos Relacionados

Jose y Cyntia [2] usaron dos métodos para implementar redes neuronales y una base de datos que tiene diferentes características de pacientes, algunos tienen algún tipo de enfermedad cardíaca y otros no. Se desarrolla una primera parte de red neuronal supervisada, específicamente se implementa un perceptrón multicapa, la segunda parte presenta redes no supervisadas, implementando una red ART2 (Teoría de resonancia adaptativa), para la implementación de redes neuronales para la detección de la presencia de enfermedades en el corazón.

En este trabajo se desarrolla un algoritmo de clasificación de sonidos cardíacos utilizando redes neuronales convolucionales (CNN) para la determinación de valvulopatías. Se realizó una etapa de pre procesamiento de la señal de fonocardiograma (PCG).

III. Marco Teórico

A. Enfermedades del Corazón

Desde la descripción de los casos iniciales de COVID-19 en la provincia de Wuhan en China, se descartaron los reportes que describen a los pacientes con enfermedades cardíacas subyacentes (hipertensión sistémica. insuficiencia cardiaca, enfermedad coronaria)[1]

1) Hipertensión sistémica

Es la presión alta en las arterias sistémicas - vasos sanguíneos que llevan sangre del corazón a

los tejidos del cuerpo. La presión sanguínea sistémica alta es causada usualmente por la constricción de las arterias pequeñas, es una enfermedad crónica, controlable.

2) Insuficiencia Cardiaca

Trastornos cardiovasculares y no cardiovasculares, así como factores relacionados con el paciente y factores iatrogénicos, que pueden desencadenar una progresión rápida o un agravamiento de los signos y síntomas de insuficiencia cardiaca, lo que conduce a un episodio de insuficiencia cardiaca aguda que suele requerir el ingreso hospitalario del paciente. [3]

La prevención primaria de la insuficiencia cardíaca aguda se centra principalmente en la prevención, el diagnóstico precoz y el tratamiento de los factores de riesgo cardiovascular y la cardiopatía. [3]

El COVID-19 puede generar daño miocárdico a lo largo de la enfermedad; causada por una elevación de troponina tiene mecanismos que han sido ampliamente estudiados

El SARS CoV-2 se liga al receptor de la enzima convertidora de angiotensina, la que tiene una distribución amplia por el organismo incluyendo en las cédulas miocárdicas. Se ha reportado casos de invasión del virus SARS CoV-2 de las células miocárdicas que pueden llevar a un cuadro de miocarditis, lo que produce daño miocárdico y elevación del nivel de troponina cardiaca.[1]

B. **Datos Médicos de Padecimiento**

Las características que viven las personas que padecen este mal, se ve reflejado en las siguientes características, los cuales en un excesivo consumo o alto riesgo, pueden contribuir a desarrollar en mayor proporción las enfermedades al corazón:

a. Presion Arterial Sistolica

Según los autores Blanco, Macias y Lopez, propusieron en su trabajo que la presión arterial sistólica se define como el valor máximo de la presión arterial cuando el corazón se contrae (sístole). Es la presión de la sangre que expulsa el corazón sobre la pared de los vasos [4].

Por otro lado los autores Valero y Garcia, la definen en su investigación como el acto donde la presión sistólica es mayor de 150 mmHg y la diastólica es menor de 90 mmHg, definiéndose la presión arterial sistólica como la fuerza ejercida por la sangre contra la pared arterial cuando el ventrículo se contrae [5].

b. Consumo de Tabaco

El autor, la define como la planta que se cultiva por sus hojas, las cuales se secan y fermentan y luego se usan en varios productos. Contiene nicotina, un ingrediente que puede conducir a la adicción, lo que explica por qué a muchas personas que consumen tabaco les resulta difícil dejar de consumirlo [6].

c. Proteína de Baja Densidad del Colesterol

1. ***Proteínas:*** Según lo mencionado por los autores Gonzales, Tellez y Sanpedro en su investigación [7], el nombre proteína fue sugerida por Berzelius para llamar así, al material que describiera el químico holandés Mulder en 1838 como “sustancia compleja” en cuya composición intervenía el nitrógeno (N), y la cual, era sin duda la más importante de todas las sustancias conocidas en el “reino orgánico”, sin la cual no parecía posible la vida sobre nuestro planeta.

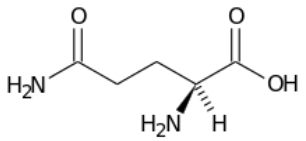

Figura 1. Estrutura de la proteina

2. ***Lipoproteínas****:* El concepto mencionado por Marassi y Diaz en su trabajo [8] , enfatizan que estos constituyen un sistema polidisperso y heterogéneo de partículas de morfología casi esférica, que tienen un núcleo o core hidrófobo formado por lípidos no polares, es decir, colesterol esterificado y triglicéridos (TAG), y por una capa superficial hidrófila que contiene colesterol no esterificado, fosfolípidos (FL) y unas proteínas específicas denominadas apoproteínas (Apo). Las Apo no solamente cumplen un papel estructural en las partículas lipoproteicas, además intervienen en el metabolismo de las mismas, en el que ejercen distintas funciones.

d. Adiposity

Lo mencionado en el trabajo de Kissebah y Krakower [9], mencionan que la adiposidad localizada, es la ubicación del tejido graso en determinadas zonas del

cuerpo que provocan una alteración estética del contorno corporal. Esto puede ir acompañado o no de sobrepeso.

e. Antecedentes Familiares

Las enfermedades del corazón suelen ser hereditarias. Por ejemplo, si los padres o hermanos padecieron de un problema cardíaco o circulatorio antes de los 55 años de edad, la persona tiene un mayor riesgo cardiovascular que alguien que no tiene esos antecedentes familiares [10].

1. ***Hipertensión:*** Existe cierta predisposición familiar a padecerla. Además, se ha demostrado que la hipertensión es un factor de riesgo muy importante de infarto de miocardio, insuficiencia coronaria, angina de pecho y arritmias.
2. ***Hipercolesterolemia familiar***: Uno de los factores hereditarios de mayor riesgo cardiovascular. Si tenemos antecedentes familiares de colesterol elevado debemos someternos a analíticas desde una edad temprana para que se pueda hacer un diagnóstico precoz.
3. ***Diabetes tipo 2:*** También tiene un componente genético. Si uno de los padres tiene diabetes, el hijo contempla un alto riesgo de desarrollarla. Por ello, se aconseja controlar los factores de riesgos.

f. Obesidad

En el trabajo dado por Garcia [11], él menciona que la obesidad es una enfermedad compleja, por ello su definición ha sido un proceso difícil. Su evidente relación con el consumo de alimentos determinó que durante mucho tiempo fuera vista como un trastorno de conducta, existiendo una gran resistencia a considerar las múltiples alteraciones que le dan el carácter de enfermedad.

g. Consumo de Alcohol

El consumo mantenido y excesivo puede dañar el corazón porque el alcohol es un tóxico para el músculo cardíaco, puede llegar a debilitar el corazón y causar una enfermedad denominada miocardiopatía dilatada (el corazón se dilata y disminuye la fuerza de "bombeo"), provocando en el paciente síntomas de insuficiencia cardiaca.

IV. Algoritmo de Clasificación

El algoritmo implementado en la herramienta JupyterLab, fue el de Support Vector Machine (SVM), el Support Vector Classifier (SVC), junto con el Kernel Radial Basis Function (RBF).

A. Support Vector Machine (SVM)

Las máquinas de vectores de soporte o máquinas de vector soporte (del inglés *Support Vector Machines*, SVM) son un conjunto de algoritmos de aprendizaje supervisado desarrollados por Vladimir Vapnik y su equipo en los laboratorios ***AT&T***.

Estos métodos están propiamente relacionados con problemas de clasificación y regresión. Dado un conjunto de ejemplos de entrenamiento (de muestras) podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra. Intuitivamente, una SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases a 2 espacios lo más amplios posibles mediante un hiperplano de separación definido como el vector entre los 2 puntos, de las 2 clases, más cercanos al que se llama vector soporte. Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de los espacios a los que pertenezcan, pueden ser clasificadas a una o la otra clase.

Los objetivos de SVM son separar los datos con hiperplano y extender esto a no lineales

límites usando el truco del kernel. Para calcular la SVM vemos que el objetivo es clasificar correctamente todos los datos [15].

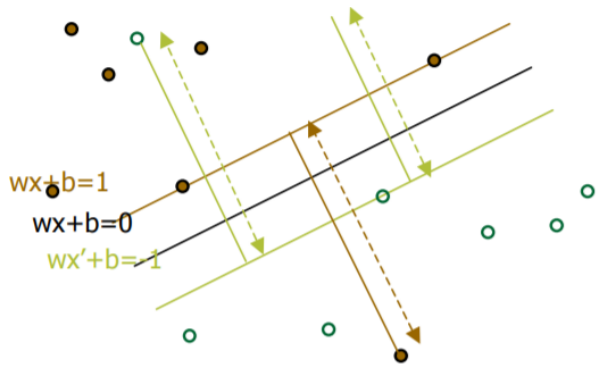

Figura 2. Representación de un Hiperplano

B. Support Vector Classifier (SVC)

La SVC es un clasificador binario que asigna una etiqueta $\boldsymbol{y} \in \{+1, -1\}$ al vector de entrada $\boldsymbol{x}$ conforme al signo de la siguiente expresión:

$$f(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) + b$$

donde

$$\boldsymbol{\phi} : \mathbb{R}^d \mapsto \mathcal{H}$$

Es una transformación del espacio de entrada a un espacio de características de igual o mayor dimensión (incluso infinita), en el que se supone una mayor separación entre las clases. El vector **w** define el hiperplano de decisión en dicho

espacio y *b* representa el sesgo respecto al origen de coordenadas. La máquina de vectores soporte es una generalización no lineal del hiperplano óptimo de decisión para problemas no separables, por lo que la formulación de la SVC parte del funcional.

La SVC aborda el problema de clasificación no separable relajando el concepto de margen, para lo que se introducen unas variables [16].

C. Kernel Radial Basis Function (RBF)

En el machine learning , el kernel de función de base radial ,o kernel RBF , es una función de kernel popular que se utiliza en varios algoritmos de aprendizaje kernelizados. En particular, se usa comúnmente en la clasificación de máquinas de vectores de soporte .

El kernel RBF en dos muestras **x** y **x'** , representadas como vectores de características en algún *espacio de entrada* , este se define como:

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

Ecuación 1:. Fórmula en dos muestras

$\|\mathbf{x} - \mathbf{x}'\|^2$, puede reconocerse como la distancia euclidiana al cuadrado entre los dos vectores de características. Donde omega es un parámetro gratuito. Una definición equivalente involucra un parámetro $\gamma = \frac{1}{2\sigma^2}$:

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2)$$

Ecuación 2:. Fórmula de definición

A. *Normalización*

Para que funcionen los datos de forma correcta en el algoritmo de Machine Learning, se deben normalizar las variables de entrada al algoritmo. Entonces debemos comprimir o extender los valores de la variable para que estén en un rango definido. Sin embargo, una mala aplicación de la normalización, o una elección descuidada del método de normalización puede arruinar tus datos, y con ello tus análisis.

Existen diversos tipos de normalización que se puede aplicar como el escalado de variables (Feature Scaling o MinMax Scaler) o el escalado estándar (Standard Scaler).

B. Escalado de Variables

En este caso, cada entrada se normaliza entre unos límites definidos:

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Ecuación 3: Fórmula normalización

La desventaja de este tipo de normalización, es que comprime los datos de entrada entre unos límites empíricos (máximo y mínimo de la variable). Esto quiere decir que si existe ruido, éste va a ser ampliado.

Después del escalado, los datos se han distorsionado. Lo que era una conexión estable, ahora parece tener muchas variaciones. Esto nos dice que este método de normalización no es adecuado para señales estables.

C. Escalado Estándar

El escalonado estándar, donde a cada dato se le resta la media de la variable y se le divide por la desviación típica).

$$X_{normalized} = \frac{X - X_{mean}}{X_{stddev}}$$

Ecuación 4. Fórmula normalización Estándar

Antes de normalizar, se calcula la media y la desviación típica. Cuando los datos sin anomalías no pasan de valores en torno al 4, se aplica la normalización estándar.

D. Grid Research

En el proyecto se enfrentó a una situación en la que necesitaba probar diferentes clasificadores con diferentes hiper parámetros. El cambiar manualmente los hiper parámetros y ajustarlos a los datos de entrenamiento cada vez se torna una situación hostil.

Si se cambian los parámetros de forma manual llevará mucho tiempo y es difícil hacer un seguimiento de los hiper parámetros que probamos.

Se implementó gracias a una librería *GridSearchCV* que es una función de biblioteca que es miembro del paquete *model_selection* del sklearn. Ayuda a recorrer hiperparametros predefinidos y ajustar su estimador (modelo) en su conjunto de entrenamiento. Selecciona los mejores parámetros de los hiperparametros enumerados.

Puede especificar el número de veces para la validación cruzada para cada conjunto de hiperparametros.

V. Datos Empleados

En esta sección se podrá detallar toda la información recogida de sitios web con respecto a nuestra base de datos y también los datos más importantes que usaremos para tener en cuenta nuestra predicción, nombrando las principales características que usaremos.

D. Base de Datos

La base de datos con la cual se está trabajando está ofrecida en el sitio web OpenML [11], los autores Vanschoren, Bischl, N van Rijn y Torgo [12]en su trabajo manifiestan que, este repositorio es un lugar donde los investigadores de machine learning pueden compartir datos automáticamente con gran detalle y organizarlos, trabajar de forma más eficaz y colaborar a escala global.

Permite a cualquiera desafiar a la comunidad con nuevos datos. para analizar, y todos los que puedan extraer esos datos para compartir su código y resultados (por ejemplo, modelos, predicciones y evaluaciones) .OpenML se asegura de que cada (sub) tarea sea claramente definida, y que todos los resultados compartidos se almacenan y organizan en línea para facilitar el acceso, la reutilización y la discusión.

E. Datos de Entrenamiento

Los datos con los cuales estaremos trabajando será dada por 10 características sacadas de la base de datos OpenMl, estas están ligadas a las principales condiciones que pueden tener anteriormente pacientes que sufren del corazón, como:

1. Presion Arterial Sistolica.
2. Acumulacion de tabaco en kg.
3. Proteína de Baja Densidad de Colesterol.
4. Adiposidad.
5. Antecedentes Familiares.
6. Comportamiento Tipo-A
7. Obesidad.
8. Consumo recurrente de alcohol.
9. Edad del paciente.
10. Enfermedad Cardiaca (respuesta).

Estos datos servirán como características principales para poder definir si una persona puede sufrir o no de una enfermedad cardiovascular. [11]

VI. Importancia de la Investigación

La investigación planteada busca implementar un analizador de SVM para predecir posibles enfermedades al corazón, teniendo en cuenta diversas características, dicha investigación ayudará a médicos o personas naturales que puedan predecir si ellos o sus pacientes podrían sufrir enfermedades del corazón, esto previamente detallando la información de cada persona, como su nivel de adiposidad, si padece o no obesidad, si tienen antecedentes familiares, etc.

Además el proyecto busca tener una efectividad con resultados predictorios buenos, oscilando entre un 70% a 95% de predicción acertada.

VII. Resultados

En esta sección de resultados se obtuvieron diversos resultados, entre los cuales se encuentran el valor obtenido de aproximación, el cual resulta de un valor de 0.73.55 realizando el proceso de aprendizaje, por otro lado se obtuvo una precisión de 0.6551, en el algoritmo.

Por otro lado el uso de los Kernels ayudaron a mejorar el resultado, se usó un Kernel RBF, para que sea más exacta la vista de datos en pantalla y mejore la salida de nuestro algoritmo con una mejor precisión y exactitud.

Por lo tanto se puede afirmar que nuestro modelo se encuentra entrenando de manera efectiva, dando los resultados antes mencionados, los cuales pueden mejorar abordando otro modelo de entrenamiento.

VIII. Trabajos Futuros

Al realizar el desarrollo de este proyecto, se llegó a puntos importantes a tener en cuenta, como realizar en un trabajo futuro la implementación con un nuevo tipo de algoritmo usando Redes Neuronales o CNN, las cuales ayudarían y darían una mejor exactitud para el aprendizaje.

Un trabajo futuro planteado es la implementación de una página web, el cual esté basado primordialmente de nuestra Inteligencia Artificial con SVM, donde el usuario pueda ingresar los datos necesarios y finalmente pueda retornar un diagnóstico acertado.

IX. Conclusiones

Podemos concluir que en el desarrollo de este proyecto se recabó toda la información pertinente al caso de estudio elegido, teniendo como bases al algoritmo de Support Vector Machine (SVM), y la base de datos, donde se pudieron obtener resultados favorables con una precisión de 65% y una exactitud de un valor aproximado de 74% de efectividad , la cual es una buena salida de datos donde el entrenamiento se

dio con éxito, a pesar de estar usando SVM para este caso de estudio, favoreciendo a un aprendizaje mucho más eficaz.

Se concluye además que con dicho proyecto implementado, se podrá tener resultados buenos para el entrenamiento y saber cómo este ha mejorado, tomando como base diferentes algoritmos como CNN, o Redes Neuronales, los cuales implementados conjuntamente podrían tener mejores resultados que los obtenidos actualmente en este trabajo.

Referencias

[1] F. Del-Carpio-Muñoz, "COVID-19 y el corazón," Diagnóstico, vol. 59, no. 3, pp. 133–140, Jan. 2020, doi: 10.33734/DIAGNOSTICO.V59I3.236.

[2] D. Farmakis, J. Parissis, J. Lekakis, and G. Filippatos, "Insuficiencia cardiaca aguda: epidemiología, factores de riesgo y prevención," Rev. Española Cardiol., vol. 68, no. 3, pp. 245–248, Mar. 2015, doi: 10.1016/J.RECESP.2014.11.009.

[3] L. Blanco Cedres, C. Coromoto Macias Tome, and M. López Blanco, "Relación Entre la Maduración Temprana, Índice de Masa Corporal y el Comportamiento Longitudinal de la Presión Arterial Sistólica," *09 Sep. 2000,*

[4] R. Valero and García Soriano, A, "Normas, consejos y clasificaciones sobre hipertensión arterial," Enfermería Global, vol. , no. 15, 2021,

[5] National Institute on Drug Abuse, "Cigarrillos y otros productos con tabaco – DrugFacts | National Institute on Drug Abuse," National Institute on Drug Abuse, Jan. 16, 2020.

[6] L. González-Torres, A. Téllez-Valencia, J. Sampedro, and H. Nájera, "LAS PROTEÍNAS EN LA NUTRICIÓN," 2007. [Online]. Available: https://www.medigraphic.com/pdfs/revsalpubnut/spn-2007/spn072g.pdf.

[7] N. Brandan, C. Llanos, B. I. Barrios, A. P. Escalante, and D. A. Ruiz Diaz, "Lipoproteínas," *2006*. Available: https://med.unne.edu.ar/sitio/multimedia/imagenes/ckfinder/files/files/Carrera-Medicina/BIOQUIMICA/lipoproteinas.pdf.

[8] A. H. Kissebah and G. R. Krakower, "Physiological Reviews," *Department of Medicine, Division of Endocrinology, Metabolism, and Clinical Nutrition, Froedtert Memorial Lutheran Hospital, Milwaukee, Wisconsin, The American Physiological Society-Vol. 74, No. 4, October 1994.*

[9] Administrator, "Antecedentes familiares - Fundación Española del Corazón," Fundaciondelcorazon.com, 2021. https://fundaciondelcorazon.com/prevencion/marcadores-de-riesgo/antecedentes-familiares-historial.html (accessed Jul. 16, 2021).

[10] E. García García, "¿Qué es la obesidad?", Revista *de Endocrinología y Nutrición Vol. 12, No. 4 Supl. 3*. 2004. Available: https://www.medigraphic.com/pdfs/endoc/er-2004/ers043c.pdf

[11] J. Vanschoren, "OpenML," OpenML: exploring machine learning better, together., 2015. https://www.openml.org/d/1498 (accessed Jul. 23, 2021)

[12] C. de I. de la U. D. F. J. de Caldas and C. M. O. Rey, "Revista Ingeniería," Redes Ing., vol. 1, no. 2, pp. 38–46, May 2010, doi: 10.14483/2248762X.7159.

[13] M. Molina and L. Eduardo, "Determinación de valvulopatías cardíacas a través del análisis de sonidos del corazón mediante algoritmos de machine learning."

[14] E. Juliá Martínez, E. L. Rubio, R. María, and M. Quiroga, "HEALTH ENGINEERING MAJOR IN BIOMEDICAL ENGINEERING CARDIOVASCULAR DISEASE PREDICTION THROUGH ARTIFICIAL INTELLIGENCE ALGORITHMS PREDICCIÓN DE ENFERMEDADES CARDIOVASCULARES MEDIANTE ALGORITMOS DE INTELIGENCIA ARTIFICIAL," 2020.

[15] V. Jakkula, "Tutorial on Support Vector Machine (SVM),." *Vikramaditya Jakkula, School of EECS, Washington State University,*.

[16] Vikramaditya Jakkula, School of EECS, Washington State University, Pullman 99164.