ISSN 1870-4069

Predicción de enfermedades cardíacas derivadas de diabetes, mediante algoritmos genéticos: caso de estudio

Isamar Aparicio-Montelongo[1], José M. Celaya-Padilla[2], Huizilopoztli Luna-García[2], Carlos E. Galván-Tejada[2], Jorge I. Galván-Tejada[2], Hamurabi Gamboa-Rosales[2]

[1] Consejo Nacional de Ciencia y Tecnología,
Unidad Académica de Ingeniería Eléctrica,
México

[2] Universidad Autónoma de Zacatecas,
México

{20204469, jose.celaya, hlugar, ericgalvan,
gatejo, hamurabigr}@uaz.edu.mx

Resumen. En México, según datos del INEGI las principales causas de muerte son Enfermedades Cardíacas (EC) y Diabetes Mellitus (DM). En el primer semestre de 2021 se registraron 579,586 decesos, ocupando los primeros lugares solo después de COVID-19. Correspondiendo a las EC, como la segunda más importante con un 19.7%. Gracias al avance tecnológico, en la actualidad, es posible crear modelos para el diagnóstico oportuno de patologías, mediante técnicas de Inteligencia Artificial (IA). El objetivo de esta investigación es implementar algoritmos genéticos (AG) como método de selección de características posteriormente, generar un modelo multivariado con Regresión Logística para conocer si es posible considerarlo como una herramienta eficaz en la detección de pacientes propensos a sufrir un episodio cardíaco derivado de DM.

Palabras clave: Enfermedades cardíacas, inteligencia artificial, predicción, diabetes, algoritmos genéticos, selección de características.

Prediction of Diabetes-Related Heart Disease Using Genetic Algorithms: A Case Study

Abstract. In Mexico, according to INEGI data, the main causes of death are Heart Disease (CD) and Diabetes Mellitus (DM). In the first semester of 2021, 579,586 deaths were registered, occupying the first places only after COVID-19. Corresponding to CD, as the second most important with 19.7%. Thanks to

technological advances, it is now possible to create models for the timely diagnosis of pathologies, using Artificial Intelligence (AI) techniques. The objective of this research is to implement genetic algorithms (GA) as a feature selection method and subsequently generate a multivariate model with Logistic Regression to find out if it is possible to consider it as an effective tool in the detection of patients prone to suffer a cardiac episode derived from DM.

Keywords: Heart disease, artificial intelligence, prediction, diabetes, genetic algorithms, feature selection.

1. Introducción

Según la Organización Mundial de la Salud (OMS), la Diabetes es una enfermedad metabólica crónica caracterizada por niveles elevados de glucosa en sangre (o azúcar en sangre), que con el tiempo es una causa importante de complicaciones cardíacas, retinopatía, nefropatía, accidente cerebrovascular y amputación de miembros inferiores [1]. Cada año mueren más personas por Enfermedades Cardíacas (EC) que por cualquier otra causa. Más de tres cuartas partes de los fallecimientos relacionados con cardiopatías ocurren en países en proceso de desarrollo [2].

Lamentablemente, en México, los datos del INEGI muestran que tan solo en 2019 fallecieron 156,041 personas, por distintas enfermedades del corazón. Así mismo, en 2020 el número fue de 218,704 individuos, es decir, hubo un incremento del 40% aproximadamente de personas fallecidas entre ambos años [3] y para el primer semestre de 2021 ya existían 113,899 muertes [4]. La probabilidad de padecer una Enfermedad Cardiovascular (ECV) aumenta conforme se desarrolla la obesidad y sobrepeso inducido por malos hábitos alimenticios, inactividad física, consumo nocivo de alcohol, sal, azúcares, grasas, entre otros [1].

Hoy en día, la Ciencia y la Tecnología han concentrado sus esfuerzos para tratar de mitigar el aumento de decesos por estas enfermedades, por tal motivo, en el presente estudio, se realizó una aproximación precisa para determinar las principales causas de complicaciones cardíacas, derivadas de Diabetes y presentes en grupos de edad (Jóvenes, Adultos y Adultos Mayores) proponiendo a algoritmos genéticos como método de selección de características, dado que están inspirados en la selección natural, simulando el proceso evolutivo de los organismos vivos para resolver problemas de optimización y búsqueda. A continuación, se describen algunas investigaciones, en donde se apreciarán sus contribuciones, en el ámbito detección de estos padecimientos por medio de técnicas de Inteligencia Artificial.

2. Trabajos relacionados

El uso de técnicas de Inteligencia Artificial (IA) se refiere a la combinación de algoritmos para crear máquinas o sistemas que imitan la inteligencia humana siendo capaces de analizar gran cantidad de datos, identificar patrones, formular predicciones

de manera automática y precisa, automatizar actividades como la toma de decisiones y resolución de problemas. En esta sección abordaremos el aporte de este ámbito desde un aspecto médico y científico analizando las contribuciones en la detección o diagnósticos de las enfermedades de interés de este estudio.

La comunidad científica ha tratado de incorporar estas técnicas de IA en el proceso de diagnóstico temprano de Enfermedades del Corazón, por ejemplo en 2019 González-Cedillo et al [5], desarrolló un sistema predictivo para detectar pacientes propensos a sufrir alguna EC, a través del uso de Naive Bayes, el autor reporta una exactitud en el modelo de 86.81%, sin embargo, no fueron exploradas otras opciones de clasificación, así mismo, el modelo propuesto incluía 75 atributos, de los cuales, 14 fueron seleccionados en base a investigaciones efectuadas en otros estudios. Además, la cantidad de instancias analizadas (303), limita la efectividad del resultado.

Posteriormente en 2020, Chicco et al [6], realizaron una comparativa de diez clasificadores de aprendizaje automático (Random Forest, Gradient boosting, Decision tree, Regresión Lineal, Naive Bayes, SVM, Redes Neuronales y KNN, por mencionar algunos) para predecir la supervivencia de pacientes (299) con Insuficiencia Cardíaca y categorizar las características clínicas correspondientes a los factores de riesgo más importantes, comparando los resultados de las predicciones a través de índices comunes de la matriz de confusión, como el coeficiente de correlación de Matthews (MCC), área bajo la curva (AUC), accuracy, entre otros.

Obteniendo el 74% de exactitud mediante Random Forest. Así mismo, Alí et al [7] en 2020, proponen un sistema sanitario inteligente para la predicción de Enfermedades Cardíacas utilizando enfoques de fusión de características y aprendizaje profundo. Llevaron a cabo una comparativa del modelo propuesto, contra seis clasificadores existentes (Naive Bayes, Redes Neuronales, Decision tree, Random Forest, Regresión Logística y SVM).

El dataset utilizado fue una combinación de dos conjuntos diferentes sobre enfermedades del corazón, obteniendo un total de 597 pacientes y 90 variables, realizan la eliminación de características ruidosas mediante el método de la ganancia de información y entropía, reduciendo a 14 atributos más significativos. Manejan métricas de evaluación como accuracy, recall, error cuadrático medio (por mencionar algunas).

Los autores informan que su modelo propuesto muestra una exactitud del 83.5%, siendo mayor a otros algoritmos convencionales. Por otro lado, en 2021 Gallego Valcárcel & Lucas Monsalve et al [8], implementaron modelos de clasificación utilizando técnicas de aprendizaje automático (Redes Neuronales, Máquina de soporte de vectores y Random Forest), para predecir el riesgo de fallecer por Insuficiencia Cardíaca a partir de datos clínicos de pacientes recopilados (299). Apoyándose de técnicas de reducción de variables (análisis de componentes principales y eliminación hacia atrás), se obtienen 17 características más relevantes.

Realizaron la evaluación de los modelos por medio de validación cruzada, siendo las Redes Neuronales, el mejor algoritmo, con una exactitud del 82.63%. Mientras tanto, Faiayaz Waris & Koteeswaran et al [9] en 2021, proponen mejorar el algoritmo de vecinos cercanos (KNN, por sus siglas en inglés) y, con ello, corroborar que es más preciso que el KNN normal, en predicción temprana de Enfermedades Cardíacas.

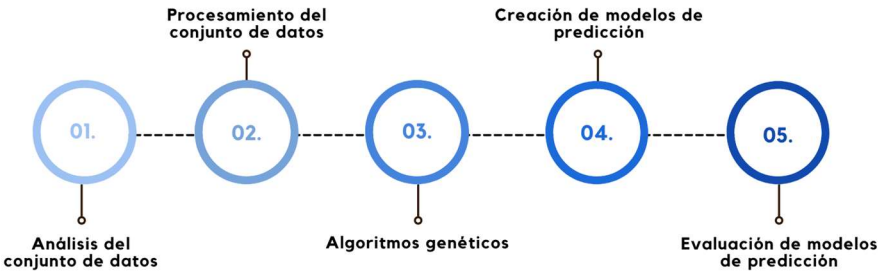

Fig. 1. Proceso general de ejecución de estudio.

El conjunto de datos analizado consta de 303 pacientes y 14 características, mismas que fueron incluidas en los modelos planteados. Alcanzando un 93% de exactitud mediante KNN modificado. Como se puede observar, la mayoría de las propuestas utilizan una cantidad muy pequeña de datos y no todas realizan una selección de características, por lo que se requiere un análisis más profundo.

Es por ello, que en esta investigación se plantea trabajar con un conjunto de datos de mayor dimensión considerando información de personas que radican en Estados Unidos, dado que la mayoría de sus pacientes tiene etnicidad similar a la de los mexicanos por lo que nos permitiría obtener conocimiento más amplio en la detección de enfermedades, además de que la infraestructura y políticas de aquel país permiten tener acceso a este tipo de datos de manera abierta para la comunidad científica y académica. De acuerdo con lo anterior, se optó por realizar un primer acercamiento a la predicción de Enfermedades Cardíacas con datos ya existentes en la comunidad médica de aquel país lo suficientemente amplio para generar un modelo preciso y eficaz y poder ser implementado en un contexto como el de México.

Explorando la selección de las variables más significativas, por medio de algoritmos genéticos (basados en la teoría evolutiva funcionan con una población de cromosomas que por sus diversas características abarcan un gran campo de distintas soluciones al mismo tiempo, por el contrario, los algoritmos tradicionales manipulan un solo punto de búsqueda) [10], para después generar modelos predictivos con Regresión Logística, realizar una evaluación mediante validación cruzada y así evitar el sobreajuste.

3. Metodología

Por medio de un modelo se busca determinar cuáles son los factores de riesgo para clasificar pacientes diabéticos propensos a sufrir una Enfermedad Cardíaca. La metodología propuesta se muestra en la Fig. 1.

En primera instancia, el conjunto de datos es separado en distintas clases (sujetos con Diabetes, Pre-diabetes, No Diabéticos) así como por grupos de edad (Jóvenes, Adultos, Adultos mayores), posteriormente, se realiza la selección de características mediante un algoritmo genético, para encontrar el mejor subconjunto de variables que permita detectar una EC, por último, este modelo es evaluado mediante validación cruzada.

Tabla 1. Datos demográficos del conjunto de datos original.

Sector poblacional			Enfermos	Sanos
Hombres	Jóvenes	6736	10,205	131,769
	Adultos	66,125		
	Adultos Mayores	69,113		
Mujeres	Jóvenes	6562	13,688	98,018
	Adultos	51,943		
	Adultos Mayores	53,201		

3.1. Datos utilizados en el estudio

Se trabajó en un conjunto de datos disponible de manera abierta obtenidos de la plataforma Kaggle en su repositorio de datasets, que lleva por nombre "Indicadores de salud de Enfermedades Cardíacas (Health Indicators for Heart Disease)" [11], consta de 22 características entre binarias u ordinales de las cuales se tomarán en cuenta los atributos clínicos como edad, sexo, índice de masa corporal (IMC), control de colesterol, presión arterial alta, colesterol alto, diabetes, características sobre hábitos como actividad física, dificultad para caminar, consumo de frutas y verduras, estado de salud física, mental y general, así mismo, vicios tales como consumo excesivo de alcohol y cigarrillos.

Tiene un total de 253,680 pacientes entre sanos (57% hombres y 43% mujeres) y enfermos (predominando las mujeres con el 57%) como se muestra en la Tabla 1, de los cuales, 141,974 son hombres, 111,706 mujeres, con un rango de edad que van desde 18 a 80 años en adelante; sobresaliendo el grupo de 60 a 64 años en ambos géneros con el 13% para cada uno. El 91% de las observaciones son personas sanas (229,787) y 9% son pacientes con alguna Enfermedad Cardíaca (23,893).

3.2. Procesamiento de datos

Con el propósito de probar si hay diferencia entre grupos de edad, se determinó segmentar tres sectores de la población, los cuales, son personas Jóvenes entre 18 y 29 años (13,298), Adultos entre 30 y 59 años (118,068) y Adultos mayores de 60 años (122,314),así mismo, a fin de explorar la posibilidad de padecer alguna Enfermedad Cardíaca en pacientes Diabéticos, se dividió en tres agrupaciones: No diabéticos (213,703), Pre-diabéticos (4631) y Diabéticos (35,346), encontrando los factores de riesgo más significativos entre cada segmento de la población.

3.3. Algoritmos genéticos (AG)

Para la selección de características se implementó un AG, el cuál, es un método de búsqueda de variables que se basa en el principio de la evolución por selección natural. El procedimiento funciona haciendo evolucionar conjuntos de atributos (cromosomas) que se ajustan a determinados criterios a partir de una población aleatoria inicial

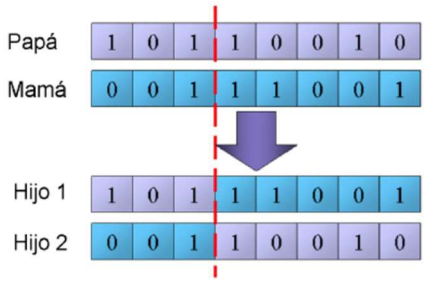

Fig. 2. Intercambio de información genética entre dos individuos.

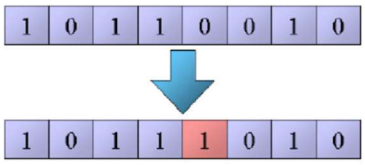

Fig. 2. Una mutación modifica al azar parte del cromosoma de los individuos.

mediante ciclos de replicación diferencial, recombinación y mutación de los cromosomas más aptos [12].

Por imitación de este proceso, los algoritmos genéticos son capaces de ir creando soluciones para problemas del mundo real. El proceso de evolución tiene como único objetivo mejorar la población de soluciones mediante la aplicación repetitiva de las operaciones de cruzamiento (sinónimo de apareamiento entre dos individuos de diferente sexo) mutación (alteraciones ocasionales del cromosoma) y selección (los cromosomas del individuo más fuerte o mejor adaptado son transferidos a su descendencia).

El intercambio de la información genética del par de individuos también se puede llevar a cabo de diferentes formas (una de ellas se ilustra en la Fig. 2), donde aleatoriamente se ha seleccionado un punto de corte común a ambos padres, y que sirve como referencia para intercambiar su información genética para producir dos hijos con características diferentes a los padres, aunque éstos hereden parte de su información genética.

Una vez establecida la frecuencia de mutación, se genera un número entre 0 y 1 de manera aleatoria y si ese número es menor que la frecuencia de mutación se permite que un gen del cromosoma cambie su información; si no, se dejará como está. La mutación modifica al azar parte del cromosoma de los individuos (ver Fig. 3), y permite alcanzar zonas del espacio de búsqueda que no estaban cubiertas por los individuos de la población actual.

Finalmente, una vez aplicados los operadores genéticos, se seleccionan los mejores individuos para conformar la población de la generación siguiente. Este proceso se realiza por medio de la evaluación de cada individuo con la función de aptitud y se reemplaza la población original. El algoritmo genético se deberá detener cuando se

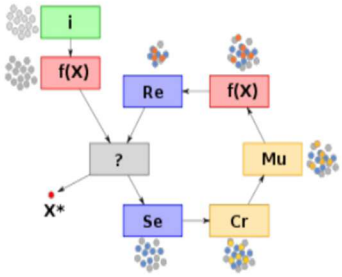

Fig. 3. Ciclo del algoritmo genético.

alcance la solución óptima, por lo general ésta se desconoce, así que se deben utilizar otros criterios de detención.

Normalmente se usan dos criterios: 1) correr el AG un número máximo de iteraciones (generaciones), y 2) detenerlo cuando no haya cambios en la población. Mientras no se cumpla la condición de término se repite el ciclo: gSelección (Se) → Cruzamiento (Cr) → Mutación (Mu) → Evaluación ($f(x)$) → Reemplazo (Re). Ver Fig. 4, donde (?) es la condición de término y $x*$ es la mejor solución [13].

En este estudio se utilizaron 500 soluciones máximas también conocidas como cromosomas, conteniendo 5 genes (chromosomeSize=5) que corresponden a modelos desarrollados utilizando un clasificador del centroide más cercano (classification.method=”nearcent”) con una precisión de clasificación del 100% (goalFitness=1) el cual también se le conoce como función objetivo ya que representa el valor de precisión que desea obtener el cromosoma siendo este el criterio de detención, para la selección de características se empleó el método selección hacia adelante, el cual es iterativo comenzando con un modelo en la que no tiene ninguna variable y en cada iteración se va añadiendo una variable hasta que la adición de nuevas características no mejore el rendimiento del modelo.

3.4. Modelos de predicción y métricas de desempeño

Después de obtener las características más significativas, se procede a la creación de modelos de predicción implementando Regresión Logística (RL) ya que en primera instancia es un algoritmo sencillo, fácil de entrenar sobre gran cantidad de datos y rara vez existe un sobreajuste en comparación con otros algoritmos, por tal motivo se optó por aplicar RL en esta investigación.

Es una técnica de aprendizaje automático que proviene del campo de la estadística, mide la relación entre la probabilidad de la característica dependiente con una o más variables independientes y el conjunto de atributos disponibles para el modelo. Lo que se busca en estos problemas es una clasificación, por lo que se obtiene un resultado binario entre 0 y 1.

Se utiliza un valor umbral para asignar los valores de probabilidad, cuando es mayor a 0.5 el resultado es positivo, de lo contrario, será negativo. A la función que relaciona la variable dependiente con las independientes se le llama función sigmoidea, la cual

Tabla 1. Matriz de confusión.

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	**Verdaderos Positivos (VP)**	**Falsos Negativos (FN)**
	Negativos	**Falsos Positivos (FP)**	**Verdaderos Negativos (VN)**

es una curva en forma de S que puede tomar cualquier valor entre 0 y 1, pero nunca valores fuera de estos límites [14], la Ecuación 1 define esta función:

$$p(x) = \frac{1}{1 + e^{-z}}, \tag{1}$$

donde z es la representación de los coeficientes del modelo de regresión que después de realizar un proceso algebraico sobre la ecuación sigmoide se obtiene la expresión matemática que representa el modelo de Regresión Logística, como se muestra en la ecuación 2:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}. \tag{2}$$

Con ayuda de algunas métricas se evalúan los modelos midiendo el desempeño de las predicciones realizadas. A continuación, se describen las métricas utilizadas en el presente estudio:

- Curva ROC: herramienta estadística utilizada para clasificar a los individuos de una población en dos grupos (uno que represente un evento de interés y otro que no). Dada por el área bajo la curva (AUC), métrica de precisión estándar para modelos de clasificación binaria el cual mide la capacidad de predecir eventos positivos en comparación con negativos, devuelve un valor decimal comprendido entre 0 y 1; los valores cercanos a 1 indican un modelo de aprendizaje automático muy preciso.
- Matriz de confusión. Es una herramienta que permite obtener el desempeño de un algoritmo, se aplica en problemas de clasificación binaria (2 clases). Está compuesta por verdaderos positivos (VP), falsos negativos (FN) es decir, casos que en realidad fueron positivos pero el modelo lo clasificó como negativo, falsos positivos (FP) y verdaderos negativos (VN) los cuales son casos que en realidad fueron negativos pero el modelo lo clasificó como positivo [15]. Como se muestra en la Tabla 2, cada columna de la matriz representa el número de predicciones de cada clase y las filas interpretan los valores reales. Mediante la matriz de confusión se pueden obtener algunas métricas de evaluación como exactitud, sensibilidad y especificidad.

- Exactitud (Accuracy). Métrica que indica el porcentaje de predicciones clasificadas correctamente, tanto como para positivos y negativos [16], dada por la Ecuación 3:

$$\text{Exactitud} = \frac{VP + VN}{VP + FP + FN + VN}. \tag{3}$$

- Sensibilidad. Indica la tasa de clasificación positiva, es decir, la proporción de casos positivos que el modelo predijo correctamente (verdaderos positivos), está dada por la Ecuación 4: (4)

$$\text{Sensibilidad} = \frac{VP}{VP + VN}. \tag{4}$$

- Especificidad. Es la capacidad de un algoritmo para predecir un falso positivo, es decir, el resultado real de la predicción es negativo y el modelo lo clasificó como positivo [17], dada por la Ecuación 5:

$$\text{Especificidad} = \frac{VN}{VN + FP}. \tag{5}$$

3.5. Evaluación de modelos de predicción

Para determinar qué tan eficaces son nuestros modelos se evaluaron el rendimiento de cada uno con Validación cruzada K Fold (K-Fold Cross-Validation), la cual nos permite estimar la capacidad predictiva de estos cuando se utilizan nuevas observaciones diferentes a los usados en el entrenamiento.

Este proceso consiste en dividir al conjunto de datos de forma aleatoria en K particiones aproximadamente del mismo tamaño cada uno, mientras que k-1 folds se usan para entrenar el modelo y una partición se utiliza como prueba.

Este proceso se repite k veces utilizando una partición distinta como validación en cada iteración, generando k estimaciones del error cuyo promedio se emplea como estimación final [18].

4. Diseño experimental

Con la intención de generar una herramienta que contribuya en mejorar las estrategias de detección temprana de Enfermedades y a la necesidad de fortalecer la atención primaria de salud en México por las altas tasas de mortalidad prematura, deterioro en la calidad de vida de los pacientes y los altos costos de atención de sus complicaciones, este estudio busca contribuir con esta finalidad realizando una adaptación de diagnósticos más personalizados.

Para ello fue necesario dividir el conjunto de datos procedente de pacientes de Estados Unidos en diferentes sectores poblacionales por edades y por grupos de casos positivos y grupos de control cómo se menciona en la sección 4.2, seguido de la implementación de algoritmos genéticos (descrito en la sección 4.3) mediante Galgo,

Tabla 2. Características más significativas y porcentaje de desempeño en grupos de Diabetes.

Subconjunto	Características más significativas	Desempeño del modelo mediante AUC
No Diabéticos	Dificultad para caminar, control de colesterol, consumo excesivo de alcohol y sexo	68.64%
Prediabéticos	Salud física, edad, frutas y salud mental	69.76%
Diabéticos	Dificultad para caminar, colesterol alto, sexo, salud física, salud en general, consumo excesivo de alcohol, edad, presión arterial alta, IMC, salud mental y fumador	73.62%

librería del software estadístico R que utiliza AG para resolver problemas de optimización, especialmente en conjuntos de datos con grandes dimensiones [12].

Seguido de la creación de modelos clasificatorios por Regresión Logística, centrándose en métricas de desempeño vistas en la sección 4.4 por último, evaluar las predicciones a través de Cross-Validation (sección 4.5) con k=3, para evitar el sobreajuste en las predicciones, esta etapa es fundamental para obtener un modelo adaptable al contexto de la sociedad mexicana.

5. Resultados

Después de revisar el diseño experimental (descrito en la sección 4), se ejecutaron dos análisis diferentes obteniendo los siguientes resultados:

5.1. Grupos de diabetes

Para el primer análisis se emplearon tres subconjuntos: No diabéticos, Pre-diabéticos y Diabéticos. Los algoritmos genéticos obtuvieron entre 4 y 11 características más significativas para cada uno de los grupos (ver Tabla 2).

La evaluación de los modelos propuestos se realiza con la validación cruzada con la intención de evitar un sesgo en las predicciones y efectúa particionando el conjunto de datos original en subconjuntos aleatorios conformado por el 70% para entrenamiento y 30% para pruebas, realizando este proceso 3 veces (k=3), es decir 3 particiones diferentes.

Se obtuvieron los resultados de las métricas para después conseguir el promedio general de cada una, como se muestra en la Fig. 4. Logrando de esta manera una exactitud (accuracy) del 55.6% para personas No diabéticas, 67.4% en Pre-diabéticos y 65.8% para Diabéticos.

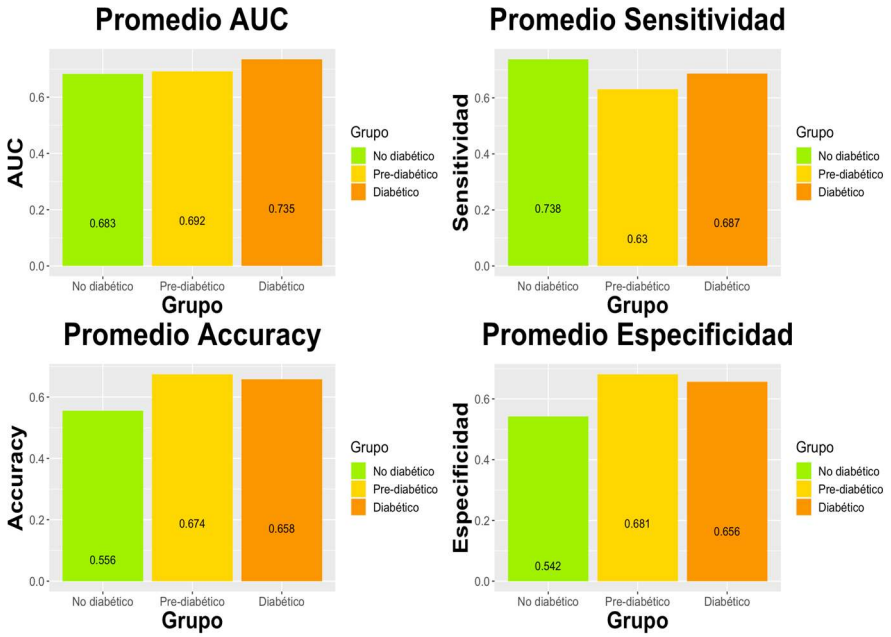

Fig. 4. Promedios de métricas de desempeño, obtenidas en grupos de Diabetes mediante validación cruzada.

Tabla 3. Características más significativas en grupos de Edad.

Subconjunto	Características más significativas	Desempeño del modelo mediante AUC
Jóvenes	Diabetes, control de colesterol, frutas	58.4%
Adultos	Edad, Diabetes	71.87%
Adultos Mayores	Salud física, fumador	63.68%

5.2. Grupos de edad

Para el segundo análisis, las variables más significativas obtenidas por el algoritmo genético se muestran en la Tabla 3, al igual que el desempeño del modelo descrito por AUC de cada modelo multivariado, siendo el grupo de adultos el de mayor precisión en las predicciones con el 71.87%, seguido de adultos mayores con 63.68% y 58.4% para jóvenes.

Un buen modelo de clasificación debe proporcionar predicciones precisas, por lo cual, es necesario aplicar una validación cruzada sobre los datos y de esta manera evitar un sesgo o sobreajuste. El promedio general de las métricas de desempeño utilizadas se muestra en la Fig. 5, abordando solamente exactitud se obtuvo 97.7% para Jóvenes, 68.3% en Adultos y 72.1% en Adultos Mayores.

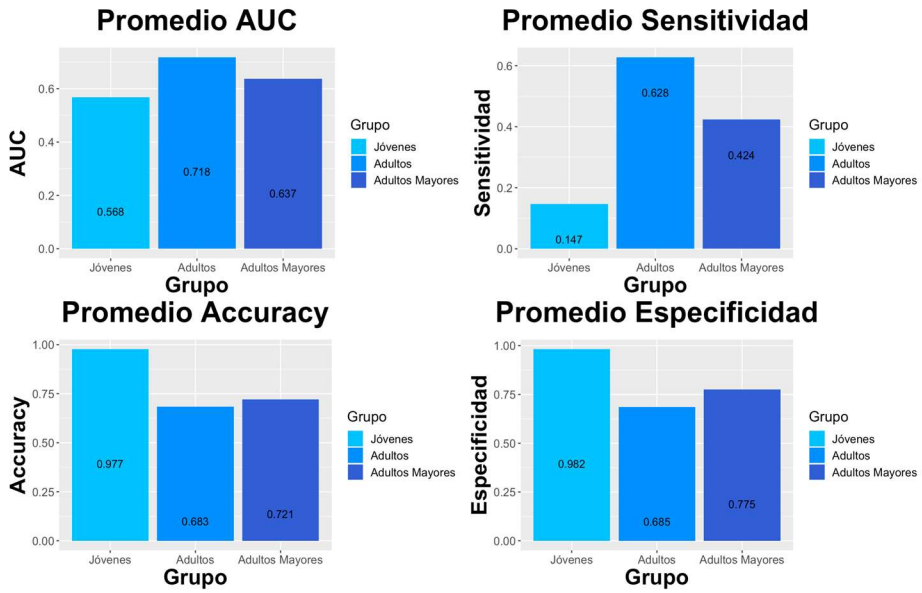

Fig. 5. Promedios de métricas de desempeño, obtenidas en grupos de Edad mediante validación cruzada.

6. Conclusiones y trabajo a futuro

La mayoría de los trabajos que incorporan IA tratan de generar un modelo que permita una correcta clasificación y no consideran las características que cambian en base a la etapa del sujeto, sin embargo, en esta investigación se busca estudiar si las variables que se están encontrando siguen siendo significativas a lo largo de la vida de los pacientes, con el fin de localizar modelos adaptativos dependiendo de su edad. De esta manera, contribuir en el diagnóstico preventivo y certero de padecimientos cardíacos proporcionando un tratamiento personalizado.

Este estudio muestra que la dificultad para caminar, un mal control de colesterol, género y un alto consumo de alcohol, son factores de riesgo que nos pueden ayudar a predecir Enfermedades Cardíacas en personas No Diabéticas, mientras que un bajo consumo de frutas, edad, mal estado de salud física y mental son características importantes para detectar una EC en pacientes Pre-diabéticos y Diabéticos, además, para este último grupo, es importante mencionar que el riesgo aumenta considerando otras variables como la dificultad para caminar, colesterol alto, género, mal estado de salud en general, alto consumo de alcohol, presión alta, índice de masa corporal y haber consumido más de 100 cigarrillos a lo largo de la vida.

En cuanto a los grupos de edad, en específico para Jóvenes y Adultos, es muy fundamental saber si el paciente es Diabético, ya que es una enfermedad clave para que un individuo sea propenso a sufrir un episodio cardíaco, debido a un mal control de

colesterol y malos hábitos alimenticios, por lo tanto, muestra un alto índice de padecer alguna complicación cardíaca.

La edad, también es un atributo importante que hay que tomar en cuenta en los Adultos, mientras que el mal estado de salud física y ser fumador son indicios que se deben considerar en Adultos Mayores. Las técnicas de Inteligencia Artificial y algoritmos de aprendizaje automático permiten predecir con un alto grado de precisión cualquier tipo de enfermedad en etapas tempranas, de manera no invasiva.

Como trabajo a futuro, se propone mejorar el desempeño del modelo de clasificación implementando otro método de selección de características dentro de los algoritmos genéticos, además de realizar un análisis de este mismo conjunto de datos, aplicado en hombres y mujeres, para determinar los factores de riesgo presentes en cada segmento y relacionar sus similitudes siguiendo la misma metodología que en el presente estudio.

Además, sería de gran interés realizar una comparativa de los resultados obtenidos en esta investigación con datos de pacientes mexicanos, como se mencionó en la sección de trabajos relacionados en la actualidad no existe un conjunto de datos abiertos con acceso al público de información de esta índole en México, debido a que nuestro país sigue en proceso de desarrollo.

Agradecimientos. Este artículo fue desarrollado gracias al apoyo de las Becas de Posgrado otorgado por Consejo Nacional de Ciencia y Tecnología (CONACYT) a la alumna Isamar Aparicio-Montelongo con el número de becario 1110281 de la Maestría en Ciencias del Procesamiento de la Información.

Referencias

1. Organización Mundial de la Salud: Diabetes (2021)
2. Organización Mundial de la Salud: Enfermedades cardiovasculares (2017)
3. Instituto Nacional de Estadística Geografía e Informática: Comunicado de prensa núm. 592/21 28 de octubre de 2021 pp. 2/4 (2021)
4. Instituto Nacional de Estadística Geografía e Informática: Estadísticas de defunciones registradas, enero-junio 2021, vol. 2021, pp. 1–40 (2022)
5. González-Cedillo, C. D.: Diagnóstico de enfermedades cardíacas con los algoritmos supervisados Naives Bayesian (2019)
6. Chicco, D., Jurman, G.: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med. Inform. Decis. Mak., vol. 20, no. 1, pp. 1–16 (2020)
7. Ali, F., El-Sappaghb, S., cS.M. Riazul-Islam, S. M., Kwak, D., Ali, A., Imran, M., Kwak, K. S.: A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Inf. Fusion, vol. 63, pp. 208–222 (2020)
8. Gallego-Valcárcel, D., Delly-Fabián, L. M.: Modelos De Aprendizaje Automático Para La Predicción Del Riesgo De Fatalidad Por Insuficiencia Cardiaca Con Datos Clínicos (2021)
9. Faiayaz-Waris, S., Koteeswaran, S.: Heart disease early prediction using a novel machine learning method called improved K-means neighbor classifier in python. Mater. Today Proc., pp. 1–7 (2021)

10. Álvarez, J., Hurtado, S., Trujillo, H.: Algoritmos genéticos (2010)
11. Kaggle, Teboul, A.: Heart Disease Health Indicators Dataset (2021)
12. Trevino, V., Falciani, F.: GALGO An R package for Genetic Algorithm Searches (Customized for Variable Selection in Functional Genomics) (2006)
13. Garduño Juárez, R.: Algoritmos genéticos (2018)
14. Rodríguez, D.: La regresión logística (2018)
15. Barrios, J.: La matriz de confusión y sus métricas (2019)
16. González, L.: Amazon Machine Learning, Guía para desarrolladores. pp. 94–96 (2016)
17. Irizarry, R.: Introducción a la Ciencia de Datos - Análisis de datos y algoritmos de predicción con R. CRC Press, pp. 538–539 (2021)
18. Amat, R.: Validación de modelos predictivos: Cross-validation, OneLeaveOut, Bootstraping (2020)