Implementación de algoritmos de inteligencia artificial para la identificación de pacientes diabéticos utilizando los niveles de lípidos en sangre

M. Hazael Guerrero-Flores[1], Carlos E. Galván-Tejada1*, Nubia M. Chávez-Lamas[2], Jorge I. Galván-Tejada[1], Hamurabi Gamboa-Rosales[1], José M. Celaya-Padilla[1] y Alejandra García-Hernández[1], Adan Valladares-Salgado[3] and Miguel Cruz[3]

[1] Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Jardín Juárez 147, Centro, Zacatecas, Zac. México; hazaelgf@uaz.edu.mx, ericgalvan@uaz.edu.mx, gatejo@uaz.edu.mx, hamurabigr@uaz.edu.mx, jose.celaya@uaz.edu.mx

[2] Unidad Académica de Odontología, Universidad Autónoma de Zacatecas, Jardín Juárez 147, Centro, Zacatecas, Zac. México; nubiachavez@uaz.edu.mx

[3] Unidad de Investigación Médica en Bioquímica, Hospital de Especialidades, Centro Médico Nacional Siglo XXI, Instituto Mexicano del Seguro Social, Av. Cuauhtémoc 330, Col. Doctores, Del. Cuauhtémoc, Ciudad de México CP 06720, México

Resumen: En los últimos años la principal causa de muerte en México está relacionada con enfermedades multifactoriales, de las cuales, la diabetes ocupa el segundo lugar, solo por debajo de enfermedades de corazón, ambas relacionadas con altos niveles de colesterol y triglicéridos en sangre. **Objetivo:** Clasificar pacientes con diabetes utilizando algoritmos de inteligencia artificial entrenados previamente con los niveles de colesterol total, HDL, LDH y triglicéridos. **Materiales y métodos:** Los descriptores relacionados con los lípidos en sangre pertenecen el Centro Médico Siglo XXI, compuesta por una muestra de 1019. Se consideran: Niveles de colesterol total, HDL, LDH y triglicéridos. La metodología propuesta consiste en dos etapas principales: entrenamiento de algoritmos de inteligencia artificial, en la cual se desarrollan modelos de caja negra para buscar la relación de los determinantes mencionados y el padecimiento de diabetes en los sujetos (padecimiento = 1, ausencia = 0), y una segunda etapa para la validación de los algoritmos, utilizando como métrica la sensitividad y especificidad de los mismos mediante la curva ROC y el área bajo la curva (AUC). **Resultados:** los modelos de regresión logística, árboles de decisión y máquina de soporte vectorial, adquieren un valor de 0.613 hasta 0.727 de AUC, siendo estadísticamente significativos para la detección automática de pacientes diabéticos. **Conclusiones:** La

implementación de algoritmos de Inteligencia artificial, permiten la identificación de pacientes con diabetes utilizando las métricas de lípidos en sangre, para un diagnóstico asistido por computadora.

Palabras clave: *Algoritmos de Inteligencia artificial, Diabetes, Colesterol, Triglicéridos.*

Abstract: In recent years the leading cause of death in Mexico is linked to multifactorial diseases, of which diabetes ranks second, only below heart disease, both related to high cholesterol levels and triglycerides in blood. **Objective:** Classify patients with diabetes using artificial intelligence algorithms previously trained with total cholesterol, HDL, LDL and triglyceride levels. **Materials and methods:** Descriptors related to blood lipids belong to the Centro Médico Siglo XXI, composed of a sample of 1019. They are considered: Total Cholesterol Levels, HDL, LDH and Triglycerides. The proposed methodology consists of two main stages: training of artificial intelligence algorithms, in which black box models are developed to look for the relationship of the determinants mentioned and the suffering of diabetes in the subjects (presence = 1, absence = 0), and a second stage for the validation of the algorithms, using as a metric the sensitivity and specificity of the algorithms by means of the ROC curve and the area under the curve (AUC). **Results:** Logistic regression models, decision trees and support vector machine, acquire a value of 0.613 to 0.727 of AUC, being statistically significant for the automatic detection of diabetic patients. **Conclusions:** The implementation of Artificial Intelligence algorithms, allow the identification of patients with diabetes using blood lipid metrics, for a computer-aided diagnosis.

Keywords: *Artificial Intelligence Algorithms, Diabetes, Cholesterol, Triglycerides.*

1. Introducción

Actualmente en México como en el mundo, la principal causa de muerte está relacionada con enfermedades multifactoriales, de las cuales, en México la diabetes ocupa el segundo lugar, solo por debajo de enfermedades de corazón. De acuerdo a los datos recopilados por el Instituto Nacional de Estadística y Geografía (INEGI), del total de muertes registradas en el país al año 2017, 88.6% que corresponden a 622 647 muertes se debieron a enfermedades y/o problemas relacionados con la salud, de las cuales el 20.1% y 15.2% fueron por

enfermedades del corazón y la diabetes mellitus respectivamente [1]. Las 5 principales causas de muerte en el año 2017 se muestran en la figura 1.

Rango	Total	Hombres	Mujeres
1	Enfermedades del corazón 141 619	Enfermedades del corazón 75 256	Enfermedades del corazón 66 337
2	Diabetes mellitus 106 525	Diabetes mellitus 52 309	Diabetes mellitus 54 216
3	Tumores malignos 84 142	Tumores malignos 41 088	Tumores malignos 43 053
4	Enfermedades del hígado 38 833	Agresiones (homicidios) 28 522	Enfermedades cerebrovasculares 17 881
5	Accidentes 36 215	Enfermedades del hígado 28 400	Enfermedades pulmonares obstructivas crónicas 11 140

Figura 1. Principales causas de muerte en México [1]

Se le llama enfermedades cardiovasculares, a las que se relacionan con desordenes del corazón y los vasos sanguíneos, estos problemas surgen cuando grasas y otras sustancias se acumulan en las paredes de las arterias, lo que provoca que estas se estrechen y con el tiempo puede quedar obstruida completamente, ocasionando un ataque cardiaco o un accidente cerebrovascular, los cuales se presentan con mayor facilidad si existe una combinación de diversos factores de riesgo como obesidad, hipertensión arterial, la diabetes, etc. [2].

La diabetes se define como una enfermedad crónico degenerativa que se presenta cuando el páncreas no produce la suficiente insulina, hormona encargada de regular el azúcar en la sangre y puede llegar a provocar daños en el organismo e incluso la muerte [3], también puede presentarse cuando la resistencia a la insulina es crónica o mantenida, es decir, cuando la insulina va perdiendo su capacidad de ejercer su función [4], y de acuerdo un experimento realizado, se observó que existe una relación significativa entre la resistencia a la insulina y niveles altos de algunos lípidos en la sangre [5].

Como ya se mencionó, los lípidos en la sangre pueden ser un factor importante para el desarrollo de enfermedades, un tipo de lípidos es el colesterol, que es producido por el cuerpo y puede encontrarse en algunos alimentos, el colesterol incluye 2 componentes, la lipoproteína de baja densidad (LDL por sus siglas en ingles) y lipoproteína de alta densidad (HDL por sus siglas en ingles), estas están encargadas de transportar el colesterol a través de la sangre, de

las cuales LDL se considera como perjudicial, ya que provoca la acumulación del colesterol en las arterias, y HDL considerada buena, ya que es la encargada de ayudar a que el cuerpo elimine el colesterol. Otro tipo de lípidos son los triglicéridos, que a niveles elevados son malos para la salud. Los niveles de lípidos recomendados se muestran en la figura 2 [6].

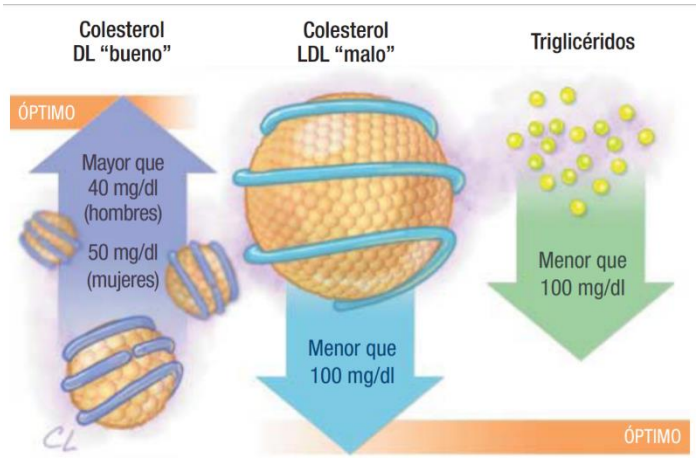

Figura 2. Niveles recomendados de lipoproteínas y triglicéridos [6]

Gracias al avance tecnológico ahora es posible el desarrollar herramientas enfocadas a la medicina, ya sea para el cuidado de la salud o para el diagnóstico de enfermedades en etapas tempranas, tal es el caso de la inteligencia artificial, una rama de la informática, la cual es capaz de analizar datos médicos complejos, para buscar una relación significativa entre los datos y con ello poder diagnosticar o predecir ciertos escenarios, así que se considera como una herramienta vital que ayudara a brindar atención medica de manera eficiente [7].

De acuerdo a lo anterior, el objetivo de este proyecto es la implementación de algoritmos de inteligencia artificial para la clasificación de pacientes con diabetes, dichos algoritmos se entrenarán previamente con el nivel de colesterol total, LDL, HDL y triglicéridos.

2. Materiales y métodos

En este apartado se muestra el procedimiento usado para la clasificación de pacientes que padecen diabetes, así como la descripción del conjunto de datos empleado, la metodología empleada que consiste en 2 principales etapas: la primera, consta el entrenamiento de los algoritmos de inteligencia artificial, los cuales se desarrollan como modelos de caja negra para identificar la relación de los determinantes ya mencionados, la segunda, comprende la validación de los algoritmos, utilizando como métrica la sensitividad y especificidad de los

mismos mediante la curva ROC y el área bajo la curva (AUC). Además de agregar la descripción del software y hardware usado para llevar a cabo los experimentos.

2.1 Descripción del conjunto de datos

El conjunto de datos pertenece al Centro Médico Siglo XXI, el cual contiene los datos de 1019 pacientes mexicanos, comprendidos entre los 35 y 65 años, de los cuales 517 son hombres y 502 mujeres. Se tienen los datos de 520 pacientes que padecen diabetes y de 499 que no padecen esta enfermedad. Además, se tienen los niveles de lípidos en la sangre como el colesterol total, LDH, HDL y triglicéridos.

En la tabla 1 se muestra la descripción de las características pertenecientes al conjunto de datos, de las cuales, los lípidos se usaron en el entrenamiento de los algoritmos de inteligencia artificial.

Tabla I. Descripción de las características

Característica	Descripción
Edad	Edad del paciente
Genero	Genero del paciente (Masculino = 0, Femenino = 1)
Colesterol	Nivel de colesterol total
LDL	Nivel de lipoproteína de baja densidad
HDL	Nivel lipoproteína de alta densidad
TG	Nivel de triglicéridos
Diabetes	Condición del paciente (No diabético = 0, Diabético = 1)

Para una mejor visualización de los niveles de lípidos contenidos en el conjunto de datos se crearon histogramas con los niveles de colesterol total, LDL, HDL y triglicéridos, los que se muestran en las figuras 3, 4, 5 y 6 respectivamente.

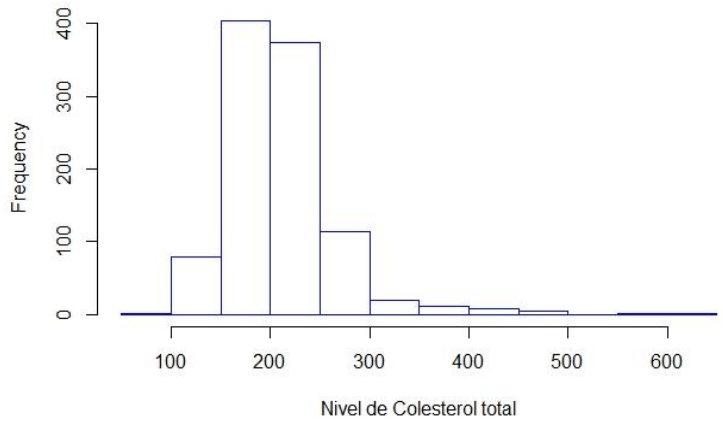

Figura 3. Histograma del nivel de colesterol total.

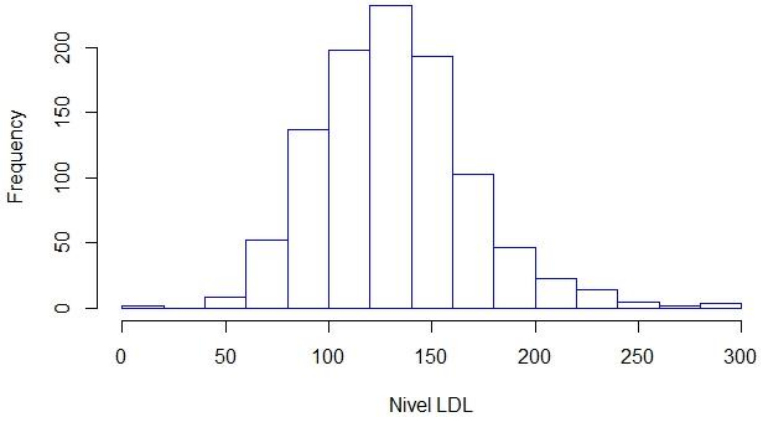

Figura 4. Histograma del nivel de LDL.

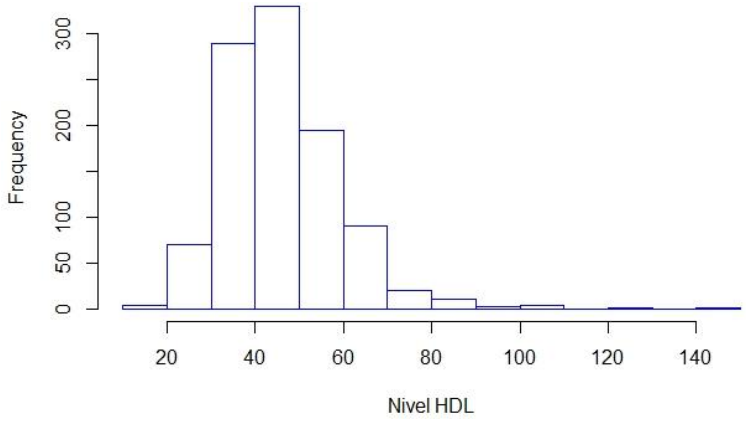

Figura 5. Histograma del nivel de HDL.

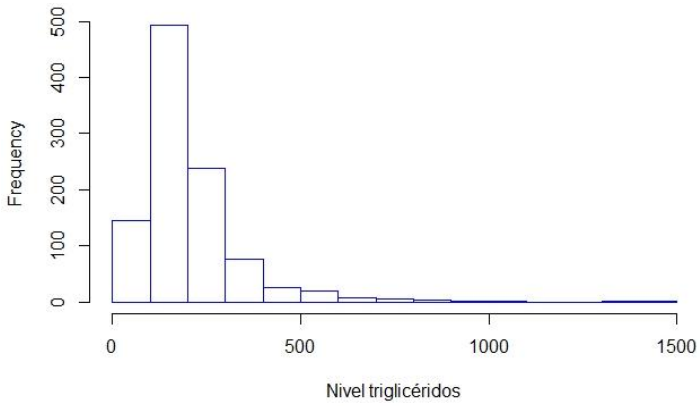

Figura 6. Histograma del nivel de triglicéridos.

2.2 Entrenamiento de los algoritmos de inteligencia artificial

Se implementaron varios algoritmos de inteligencia artificial, como regresión logística, arboles de decisión, K-ésimo vecino más cercano (KNN, por sus siglas en ingles) y máquina de soporte vectorial (SVM, por sus siglas en ingles).

La regresión logística ya sea simple o múltiple es un método de regresión que nos permite encontrar la probabilidad de una variable cualitativa binaria, en función de otra cuantitativa, es muy utilizada cuando es necesario una clasificación binaria, esto ya que se modela el logaritmo de la probabilidad de pertenencia a cada grupo. En la figura 7 se muestra un ejemplo del modelo de regresión logística [8].

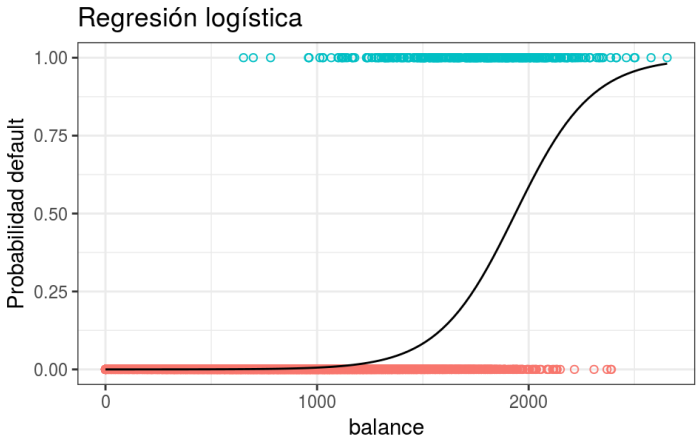

Figura 7. Modelo de regresión logística [8]

El modelo de árboles de decisión se basa principalmente en el aprendizaje inductivo a partir de las observaciones y construcciones lógicas, este modelo es llamado así, ya que el conocimiento obtenido se representa mediante un árbol, el cual está compuesto por un

conjunto de nodos, hojas y ramas. El punto donde comienza el proceso de clasificación se le llama nodo principal o raíz, en los nodos secundarios se realizan preguntas a los atributos, donde las posibles respuestas se representan con nodos hijos, conectados mediante ramas, los nodos hoja o finales son los que representan una decisión, correspondiente a una clase de la variable de salida. La estructura de un modelo de árbol de decisión se muestra en la figura 8 [9].

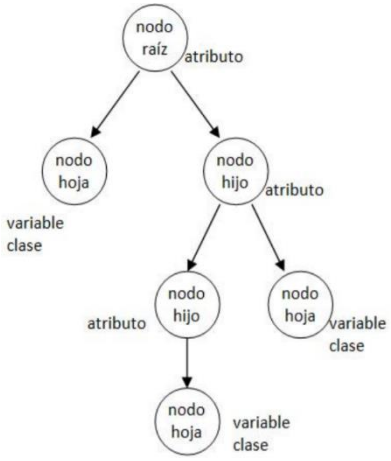

Figura 8. Estructura de un árbol de decisión [9]

Otro de los algoritmos es el de SVM, el cual comienza mapeando los datos de entrada a una dimensión mayor, es decir, si se encuentran en segunda dimensión los mapea a tercera dimensión, con el objetivo de poder encontrar hiperplanos que separe las clases, después elije al que maximice más el margen de separación entre clases, como se observa en la figura 9 [10].

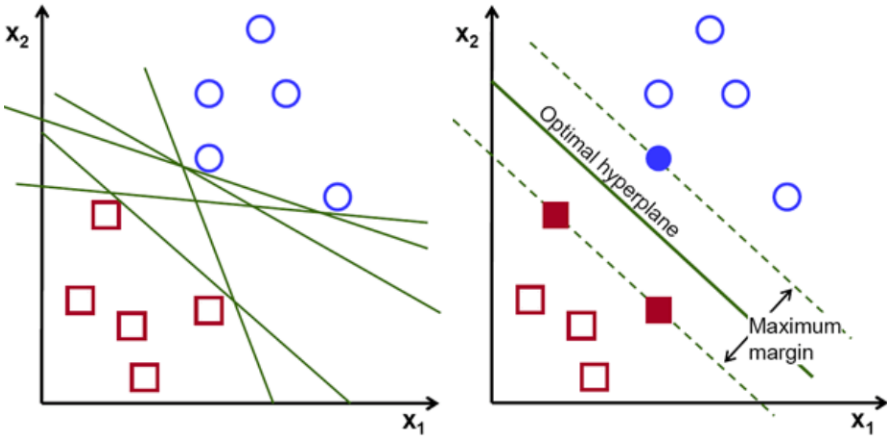

Figura 9. Elección del hiperplano que maximiza más el margen de separación entre clases [11]

Por ultimo cabe mencionar al modelo de K-ésimo vecino más cercano, el cual se basa en clasificar a los nuevos datos, dependiendo de la clase más frecuente entre sus k vecinos más cercanos [12], en este algoritmo existen 2 principales factores que afectan su desempeño, el primero es el tipo de distancia a calcular, ya sea distancia Euclidiana, Mahalanobis, Minkowsky o cualquiera de sus variantes. La segunda es la elección del valor de k, siendo la cantidad de elementos cercanos a tomar, para definir la clase más frecuente entre ellos [13]. La figura 10 puede observarse un ejemplo de este algoritmo, en donde se cuenta con 2 clases, el valor de k es 3, por lo que el nuevo elemento se clasifica mediante los 3 elementos más cercanos a él.

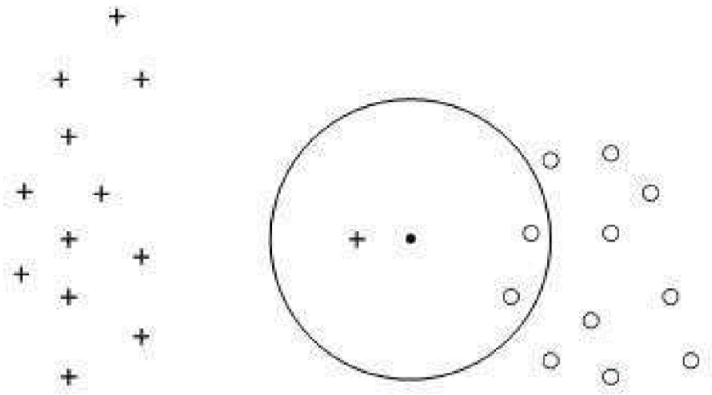

Figura 10. Ejemplo del algoritmo KNN [12]

2.3 Validación de los algoritmos

Para realizar la evaluación de los algoritmos descritos en el apartado anterior, se utilizó como métrica la sensitividad y especificidad de los mismos mediante la curva ROC y el área bajo la curva (AUC).

Para describir el significado cada uno de los conceptos usados para la evaluación de los algoritmos, es necesario para la clasificación de un paciente diabético, especificar un punto de corte, en el cual, los que se encuentren debajo de este valor, pertenezcan a una clase y los que lo superen a otra. Después se realiza la clasificación de los datos de prueba, para luego poder comparar el resultado arrojado por el algoritmo, con el valor verdadero. Así surgen 4 posibles grupos de la comparación [14].

1) Los pacientes que el algoritmo clasificó como diabéticos y que en realidad lo son, se les denomina verdaderos positivos (VP)

2) Los pacientes que el algoritmo clasificó como no diabéticos, pero que en realidad si padecen la enfermedad, a estos se les llama falsos negativos (FN)

3) Los pacientes que el algoritmo clasificó como diabéticos, pero que en realidad no padecen la enfermedad, a estos se les nombra falsos positivos (FP)

4) Los pacientes que el algoritmo clasificó como no diabéticos y que en realidad no padecen la enfermedad, se led denomina verdaderos negativos (VN)

Al conocer que cantidad de pacientes pertenece a cada uno de los 4 grupos, es posible calcular la sensitividad y especificidad del punto de corte empleado. Indicándonos la probabilidad de clasificar correctamente a los pacientes diabéticos, y correctamente a los que no la padecen la enfermedad, respectivamente. Estas métricas son calculadas mediante las siguientes formulas [14].

$$Sensitividad = \frac{VP}{VP + FN} \qquad Especificidad = \frac{VN}{VN + FP}$$

Se entiende ahora que el valor de la sensitividad y de la especificidad varían dependiendo del punto de corte elegido, por lo que es necesario, ir variando este punto de corte para encontrar en qué valor se obtiene mayor sensitividad y especificidad. La curva ROC es graficada mediante los valores de estas dos métricas en cada uno de los puntos de corte, en donde el eje x corresponde a 1- especificidad y el eje y corresponde a la sensitividad, como

ya se mencionó, al ser una probabilidad estos valores van de 0 a 1 para ambas métricas. En la figura 11 puede observarse un ejemplo del grafico de una curva ROC [14].

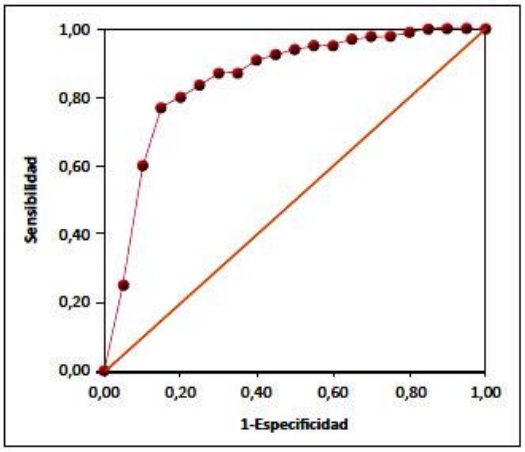

Figura 11. Gráfico de una curva ROC, donde cada punto negro representa un punto de corte [14]

El área bajo la curva (AUC por sus siglas en ingles), nos permite determinar qué tan bueno es el algoritmo, es decir, que capacidad tiene para distinguir entre pacientes que padecen diabetes y los que no la padecen, ya que los valores de la sensitividad y especificidad solo toman valores de 0 a 1, la curva ROC queda limitada a un área máxima de 1, sí se llegase a obtener un valor de 1 para el AUC, este nos indicaría que el algoritmo clasifica perfectamente a todos los pacientes. Cuando el valor de la curva ROC es de 0.5 se forma una línea recta desde el punto (0,0) al punto (1,1) indica que el algoritmo es incapaz de distinguir o discriminar entre las clases, por lo que entre más cerca se encuentre el valor del AUC al 1, el algoritmo es mejor [14].

2.4 Software y Hardware empleados

Para la creación de los histogramas como para el entrenamiento de los modelos y la validación se empleó el lenguaje de programación R en su versión 3.5, instalando varias librerías específicas para cada algoritmo de inteligencia artificial y para la validación.

Esto fue desarrollado en una laptop Dell G7 con procesador Intel(R) Core(TM) i7-8750H CPU 2.20GHz, con 16 GB de memoria RAM, 128 GB SSD, con sistema operativo Windows 10.

3. Resultados

En esta sección, se muestran los resultados obtenidos al evaluar cada uno de los algoritmos mencionados anteriormente, obteniendo para cada uno de ellos el grafico de la curva ROC y el AUC.

Para llevar a cabo el entrenamiento y la evaluación de cada uno de los algoritmos, inicialmente los datos se separaron en 2 subconjuntos, los cuales corresponden al 70% y 30% de los datos, que fueron empleados para la fase de entrenamiento y para la evaluación respectivamente.

3.1 Evaluación del modelo de regresión logística

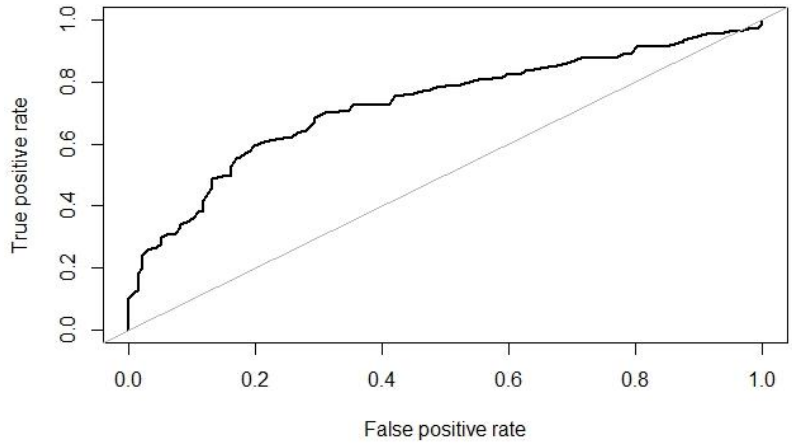

Figura 12. Grafico de la curva ROC con el modelo de regresión logística

Para el modelo de regresión logística se obtuvo un valor de AUC de 0.727.

3.2 Evaluación del modelo de árbol de decisión

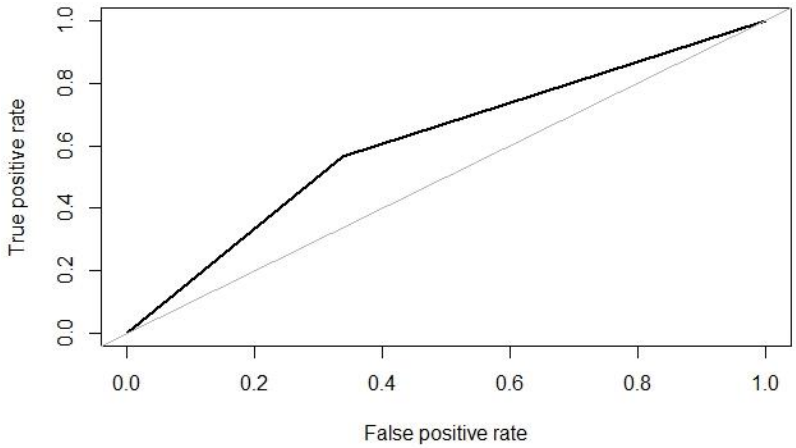

Figura 13. Grafico de la curva ROC con el modelo de árbol de decisión

Para el modelo de árbol de decisión se obtuvo un valor de AUC de 0.613.

3.3 Evaluación del modelo de máquina de soporte vectorial (SVM)

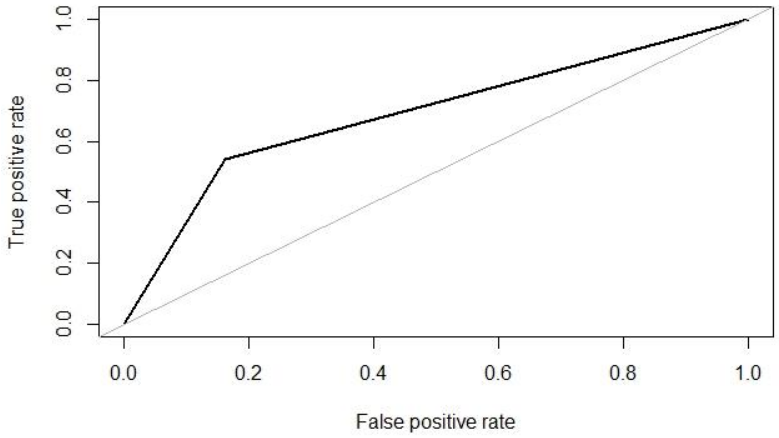

Figura 14. Grafico de la curva ROC con el modelo SVM

Para el modelo de SVM se obtuvo un valor de AUC de 0.69.

3.4 Evaluación del modelo de K- ésimo vecino más cercano (KNN)

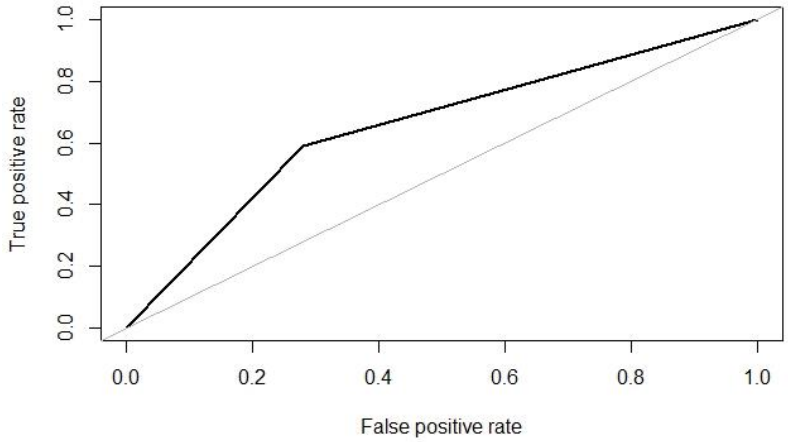

Figura 15. Grafico de la curva ROC con el modelo KNN

Para el modelo de KNN, como se mencionó, un factor importante es la elección del valor de K, por lo que se realizó una búsqueda exhaustiva para encontrar el valor adecuado que resultara en el mayor valor para el AUC, obteniendo K=9 con el que se obtuvo un valor de AUC de 0.654.

4. Conclusiones

La implementación de algoritmos de Inteligencia artificial, permiten la identificación de pacientes con diabetes utilizando las métricas de lípidos en sangre, para un diagnóstico asistido por computadora.

Los 4 algoritmos evaluados arrojaron valores de AUC estadísticamente significativos, en un rango de 0.613 hasta 0.727. Lo que muestra que el algoritmo que mostro un mejor desempeño fue el modelo de regresión logística.

El modelo obtenido puede ser usado para posteriores aplicaciones, tales como herramientas que sean de utilidad para área médica, dando soporte a especialistas de la salud en realizar diagnósticos de forma rápida en etapas tempranas, con el objetivo de brindar una mejorar asistencia médica y así prevenir futuras complicaciones, además de reducir los costos para el paciente o en su caso del servicio de salud pública en el tratamiento de las enfermedades en etapas más avanzadas.

Referencias

[1] «CARACTERÍSTICAS DE LAS DEFUNCIONES REGISTRADAS EN MÉXICO DURANTE 2017,» 31 OCTUBRE 2018. [En línea]. Available: https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2018/EstSociodemo/DEFUNCIONES2017.pdf. [Último acceso: septiembre 2019].

[2] «Enfermedades cardiovasculares,» 17 mayo 2017. [En línea]. Available: https://www.who.int/es/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). [Último acceso: septiembre 2019].

[3] A. Loke, «Diabetes,» 30 octubre 2018. [En línea]. Available: https://www.who.int/es/news-room/fact-sheets/detail/diabetes. [Último acceso: septiembre 2019].

[4] M. Serrano Ríos, «Resistencia a la insulina y su implicación en múltiples factores de riesgo asociados a diabetes tipo 2,» 2002. [En línea]. Available: https://www.sciencedirect.com/science/article/pii/S0025775302734556. [Último acceso: septiembre 2019].

[5] A. González Chávez, L. E. Simental Mendía y S. Elizondo Argueta, «Relación triglicéridos/colesterol-HDL elevada y resistencia a la insulina,» 6 12 2010. [En línea]. Available: https://www.medigraphic.com/cgi-bin/new/resumenI.cgi?IDARTICULO=29336. [Último acceso: septiembre 2019].

[6] D. Tolmach Sugerman, «Lípidos en la sangre,» 30 octubre 2013. [En línea]. Available: https://sites.jamanetwork.com/spanish-patient-pages/2013/hoja-para-el-paciente-de-jama-131023.pdf. [Último acceso: septiembre 2019].

[7] A. N. Ramesh, C. Kambhampati, J. Monson y P. Drew, «Artificial intelligence in medicine.,» septiembre 2004. [En línea]. Available:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1964229/. [Último acceso: septiembre 2019].

[8] J. Amat Rodrigo, «Regresión logística simple y múltiple,» agosto 2016. [En línea]. Available: https://rpubs.com/Joaquin_AR/229736. [Último acceso: septiembre 2019].

[9] R. E. Barrientos Martínez, N. Cruz Ramírez, H. G. Acosta Mesa, I. Rabatte Suárez, M. d. C. Gogeascoechea Trejo, P. Pavón León y S. L. Blázquez Morales, «Árboles de decisión como herramienta en el diagnóstico médico,» septiembre 2009. [En línea]. Available: https://www.medigraphic.com/cgi-bin/new/resumen.cgi?IDARTICULO=27872. [Último acceso: septiembre 2019].

[10] G. A. BETANCOURT, «LAS MÁQUINAS DE SOPORTE VECTORIAL (SVMs),» abril 2005. [En línea]. Available: http://revistas.utp.edu.co/index.php/revistaciencia/article/view/6895. [Último acceso: septiembre 2019].

[11] P. P. Ippolito, «SVM: Feature Selection and Kernels,» Junio 2003. [En línea]. Available: https://towardsdatascience.com/svm-feature-selection-and-kernels-840781cc1a6c. [Último acceso: septiembre 2019].

[12] A. Moujahid, . I. Inza y P. Larrañaga, «Clasificadores K-NN,» [En línea]. Available: http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t9knn.pdf. [Último acceso: septiembre 2019].

[13] S. ZHANG, X. LI, M. ZONG, X. ZHU y D. CHENG, «Learning k for kNN Classification,» abril 2017. [En línea]. Available: https://dl.acm.org/citation.cfm?id=2990508. [Último acceso: septiembre 2019].

[14] J. Cerda y L. Cifuentes, «Uso de curvas ROC en investigación clínica. Aspectos teórico-prácticos,» 2012 abril. [En línea]. Available: https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0716-10182012000200003. [Último acceso: septiembre 2019].