PROYECTO FINAL MODELOS COMPUTACIONALES

Lasso Mora, Daniel Felipe.* Morales Ceballos, Jacobo.*
Orozco Orrego, Sofía.* Peñaloza López, Laura.*
MODELOS COMPUTACIONALES
Noviembre de 2022

I. INTRODUCCIÓN

El aparato cardiovascular se compone de sangre, vasos sanguíneos y el corazón que contribuye a la homeostasis mediante el bombeo de sangre a través del organismo, lo cual permite oxigenar, nutrir y eliminar dióxido de carbono de las células de un organismo que no pueden moverse para realizar estos procesos. El corazón es un órgano relativamente pequeño, se compone de dos bombas musculares que, aunque adyacentes, actúan en serie y tiene cuatro cavidades: atrios (aurículas) derecho e izquierdo y ventrículos derecho e izquierdo. Los atrios son las cavidades receptoras que bombean sangre hacia los ventrículos (las cavidades de eyección), las acciones sincrónicas de bombeo de las dos bombas atrioventriculares (AV) (cavidades derechas e izquierdas) constituyen el ciclo cardíaco, el ciclo empieza con un período de elongación y llenado ventricular (diástole) y finaliza con un período de acortamiento y vaciado ventricular (sístole). [1]

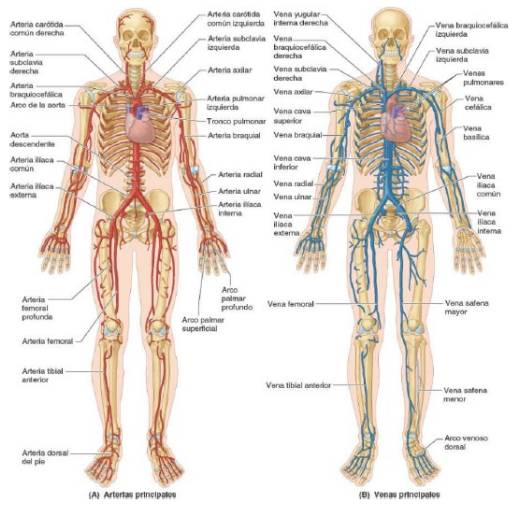

Fig. 1. Diagrama circulación

El ventrículo derecho del corazón impulsa la sangre pobre en oxígeno que procede de la circulación sistémica y la lleva a los pulmones a través de las arterias pulmonares, el dióxido de carbono se intercambia por oxígeno en los capilares pulmonares, y luego la sangre rica en oxígeno vuelve por las venas pulmonares al atrio (aurícula) izquierdo del corazón. Este circuito, desde el ventrículo derecho a través de los pulmones hasta el atrio izquierdo, es la circulación pulmonar.[1][2]

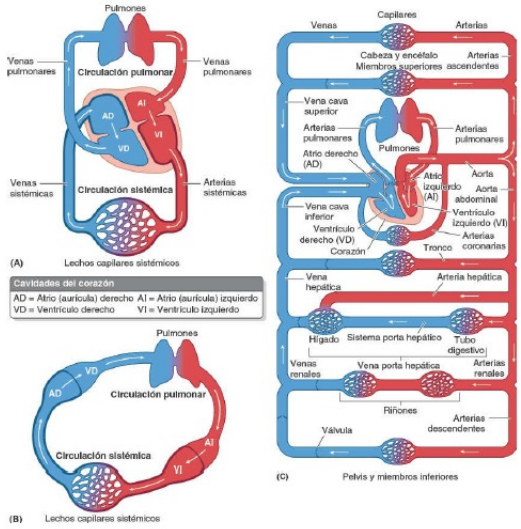

Fig. 2. Diagrama circulación

El ventrículo izquierdo impulsa la sangre rica en oxígeno, que vuelve al corazón desde la circulación pulmonar, a través del sistema arterial (la aorta y sus ramas), con intercambio de oxígeno y nutrientes por dióxido de carbono en los capilares del resto del cuerpo. La sangre pobre en oxígeno vuelve al atrio derecho del corazón por las venas sistémicas (tributarias de las venas cavas superior e inferior). Este circuito desde el ventrículo izquierdo al atrio derecho es la circulación sistémica que consiste en realidad en muchos circuitos en paralelo que sirven a las distintas regiones y/o sistemas orgánicos del cuerpo.[3]

*Estudiantes de Ingeniería Biomédica Universidad Autonóma de Manizales®

Desde los componentes que transporta la sangre tenemos el colesterol, una sustancia grasa natural presente en todas las células del cuerpo humano necesaria para el normal funcionamiento del organismo, la mayor parte del colesterol se produce en el hígado, aunque también se obtiene a través de algunos alimentos, entre sus funciones interviene en la formación de ácidos biliares, vitales para la digestión de las grasas, los rayos solares lo transforman en vitamina D para proteger la piel de agentes químicos y evitar la deshidratación, a partir de él se forman ciertas hormonas, como las sexuales y las tiroideas. [2][3]

La sangre conduce el colesterol desde el intestino o el hígado hasta los órganos que lo necesitan y lo hace uniéndose a partículas llamadas lipoproteínas, existiendo dos tipos de lipoproteínas las de baja densidad (LDL) o colesterol malo se encargan de transportar nuevo colesterol desde el hígado a todas la células de nuestro organismo y de alta densidad (HDL) o colesterol bueno que recogen el colesterol no utilizado y lo devuelve al hígado para su almacenamiento o excreción al exterior a través de la bilis. [2][3]

Así mismo como la acumulación de colesterol se da por un alto consumo en la dieta diaria, puede haber un aumento de los niveles de alcohol en la sangre, por un consumo mantenido de cantidades excesivas de bebidas alcohólicas se asocia al desarrollo de un síndrome de dependencia al alcohol (alcoholismo), pero también a múltiples enfermedades crónicas que eventualmente conducen a la muerte.

Por lo tanto, deben señalarse los efectos del alcohol sobre el hígado (cirrosis hepática y hepatitis alcohólica aguda), páncreas (pancreatitis), sistema nervioso (encefalopatías, polineuritis) y aparato locomotor (miopatía, osteoporosis), así como los deterioros psico orgánicos (amnesias lacunares, demencia alcohólica), trastornos psicóticos (alucinosis y celotipia alcohólica) y otros trastornos psiquiátricos asociados (síndromes ansioso-depresivos).[1][3]

De igual forma, consumido a dosis altas, el etanol es, indudablemente, un tóxico para todo el sistema cardiovascular, ya que daña tanto el miocardio que es una gruesa capa media helicoidal, formada por músculo cardíaco. como los propios vasos sanguíneos. El consumo crónico de alcohol causa, inicialmente, una disfunción ventricular, que puede ser sistólica y/o diastólica (miocardiopatía alcohólica subclínica) y en un porcentaje más reducido de pacientes puede inducir el desarrollo de una miocardiopatía congestiva cuyas manifestaciones clínicas y funcionales son similares a las de la miocardiopatía dilatada idiopática (miocardiopatía alcohólica clínica).

II. Planteamiento del problema

Desde lo mencionado en la introducción hay diferentes factores que afectan el correcto funcionamiento del sistema cardiovascular, siendo desde la dieta diaria hasta el consumo de sustancias nocivas como el alcohol o el cigarrillo, aspectos significantes para contraer algún tipo de complicación o enfermedad, debido a esto se pueden manipular estas variables desde el ML, aplicando diferentes filtros y modelos para hacer predicciones frente a las características descritas a continuación en contraer una enfermedad cardiovascular. Las variables que se utilizaron para la realización del modelo fueron de tres tipos, inicialmente objetivas pues era información factual, posterior a esto de examinación que son resultados de la evaluación médica y por último subjetiva la cual proporcionaba el paciente.

Desde los factores a considerar para la presencia de una enfermedad cardiovascular se obtienen:

- Age / Objective Feature — height — int (cm)
- Height / Objective Feature / height / int (cm)
- Weight / Objective Feature / weight / float (kg)
- Gender / Objective Feature / gender / categorical code
- Systolic blood pressure / Examination Feature / int
- Diastolic blood pressure / Examination Feature / int
- Cholesterol / Examination Feature / cholesterol / siendo en este caso los indicativos de 1: normal, 2: above normal, 3: well above normal
- Glucose / Examination Feature / gluc / para este caso siendo los indicativos de 1: normal, 2: above normal, 3: well above normal
- Smoking / Subjective Feature / smoke / binary
- Alcohol intake / Subjective Feature / alco / binary
- Physical activity / Subjective Feature / active / binary
- Presence or absence of cardiovascular disease / Target Variable / cardio / binary

III. Marco experimental

Las pruebas fueron realizadas con los modelos Random forest classifier (RFC), Support vector classifier (SVC), Linear SVC, Polynomial kernel SVC, Stochastic gradient descent classifier (SGD) y K-Nearest-Neighbor classifier (KNN), se tomaron en cuenta estos modelos debido a que el problema tratado es de clasificación.

A. *Random forest classifier (RFC)*

La importancia de cada característica en un árbol de decisión se calcula entonces como:

$$fi_i = \frac{\sum_{j:nodejsplitsonfeaturei} n_i j}{\sum_{k \epsilon allnodes} n_i k}$$

- fi sub(i)= la importancia de la característica i
- ni sub(j)= la importancia del nodo j

A continuación, se pueden normalizar a un valor entre 0 y 1 dividiendo por la suma de todos los valores de importancia de las características:

$$normfi_i = \frac{fi_i}{\sum_{j \epsilon allfeatures} f_i j}$$

La importancia final de la característica, en el nivel del bosque aleatorio, es la media de todos los árboles. La suma del valor de la importancia de la característica en cada árbol se calcula y se divide por el número total de árboles:

$$RFfi_i = \frac{\sum_{j \epsilon alltrees} normfi_{ij}}{T}$$

- RFfi sub(i)= la importancia de la característica i calculada a partir de todos los árboles del modelo Random Forest
- normfi sub(ij)= la importancia normalizada de la característica i en el árbol j
- T = número total de árboles

B. *Support vector classifier (SVC)*

La SVM es uno de los algoritmos de aprendizaje automático supervisado más populares y versátiles. Se utiliza tanto para tareas de clasificación como de regresión, pero en este caso hablaremos de las tareas de clasificación. Normalmente se prefiere para conjuntos de datos de tamaño medio y pequeño.

- Vectores
  Los vectores son cantidades matemáticas que tienen magnitud y dirección. Un punto en el plano 2D puede representarse como un vector entre el origen y el punto.

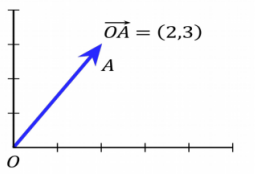

Fig. 3. OA es un vector y la longitud entre O y A es la magnitud

- Longitud del vector
  La longitud de los vectores también se denomina norma. Indica la distancia de los vectores al origen.

Length of vector $x(x_1, x_2, x_3)$ *is calculated as :*

$$\|x\| = \sqrt{x_1^2 + x_2^2 + x_3^2}$$

Fig. 4. Longitud de un vector

- Dirección del vector

Direction of vector $x(x_1, x_2, x_3)$ *is calculated as:*

$$\left\{ \frac{x_1}{\|x\|}, \frac{x_2}{\|x\|}, \frac{x_3}{\|x\|} \right\}$$

Fig. 5. Dirección de un vector

- Producto punto
  El producto escalar entre dos vectores es una cantidad escalar. Indica cómo se relacionan los vectores.

Two vectors u and v and their dot product is calculated as:

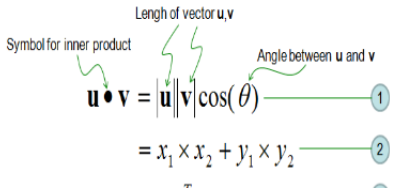

Fig. 6. Producto punto de vectores

C. *Linear SVC*

El objetivo de un SVC lineal (Support Vector Classifier) es adecuarse a los datos que se proporcionan, devolviendo un hiperplano "ideal" que divide o categoriza los datos. Desde allí, después de obtener el hiperplano, se puede entonces alimentar algunas características al clasificador para ver lo que es la clase "predecida".
Es útil cuando se trata de grandes vectores de datos dispersos. Se utiliza a menudo en la categorización de textos. El kernel splines también funciona bien en problemas de regresión. La ecuación es:

$$k(x,y) = 1 + xy + xymin(x,y) - \frac{x+y}{2} min(x,y)^2 + \frac{1}{3} min(x,y)^3$$

D. *Polynomial kernel SVC*

Es muy popular en el procesamiento de imágenes.

$$k(x_i, x_j) = (x_i \cdot x_j + 1)^d$$

Donde d es el grado del polinomio.

E. *Stochastic gradient descent classifier (SGD)*

El SGD trata de encontrar la mejor w, minimizando Q. Generalmente, Q es una función de error; así, siguiendo la dirección del gradiente en el espacio de valores de w, nos dirigimos hacia la W que minimiza el error. Más concretamente, si consideramos la regresión lineal o la clasificación por perceptrón, w especifica los parámetros de peso del modelo y Q(w) es una medida del error que el modelo comete sobre los datos.
El descenso de gradiente normal se escribe:

$$w \leftarrow w - \eta \bigtriangledown Q(w)$$

Donde se escribe el objetivo de error (con su gradiente):

$$Q(w) = \frac{1}{n}\sum_{1} Q_i(w) \Rightarrow \bigtriangledown Q(w) = \frac{1}{n}\sum_{1} \bigtriangledown Q_i(w)$$

En la ecuación w es el modelo de aprendizaje automático (es decir, sus parámetros, que determinan su comportamiento). Así que, dado algún wt, se le añade algún pequeño vector, para movernos en la dirección del w óptimo (que especifica el modelo con la menor Q). El tamaño del paso lo determina el parámetro N e R. Entonces, la dirección se determina por el gradiente de Q. (Cuando el gradiente es cero, todas las derivadas parciales de Q con respecto a w han desaparecido, lo que significa que se ha alcanzado un mínimo).
La solución habitual es limitarse a estimar:

$$\bigtriangledown Q,$$

en lugar de calcularlo. El método más sencillo es elegir una j al azar y utilizar:

$$\bigtriangledown Q \approx \bigtriangledown Qj$$

o (quizás mejor) un minibloque, nótese incluso que:

$$[\bigtriangledown Qj] = \bigtriangledown Q).$$

Esto tiene otro beneficio, más allá de la eficiencia computacional, en el sentido de que ayuda a evitar los mínimos locales en la función de error, en los que el descenso de gradiente "real" podría atascarse.

F. *K-Nearest-Neighbor classifier (KNN)*

KNN entra en los algoritmos de aprendizaje supervisado. Esto significa que tenemos un conjunto de datos con etiquetas de medidas de entrenamiento (x,y) y queremos encontrar el vínculo entre x e y. Nuestro objetivo es descubrir una función h:X→Y de modo que teniendo una observación desconocida x, h(x) pueda predecir positivamente la salida idéntica y. Siendo (x) una característica, (y) para denotar el objetivo.
En el problema de clasificación, el algoritmo K-próximo dice esencialmente que para un valor dado de K el algoritmo encontrará los K vecinos más cercanos del punto de datos no visto y entonces asignará la clase al punto de datos no visto teniendo la clase que tiene el mayor número de puntos de datos de todas las clases de K vecinos.
Para la métrica de distancia, utilizaremos la métrica euclidiana:

$$d\left(x, x^{'}\right) = \sqrt{(x_1 - x_1^{'})^2 + ... + (x_n - x_n^{'})^2}$$

Finalmente, la entrada x se asigna a la clase con mayor probabilidad.

$$P(y = j|X = x) = \frac{1}{K}\sum_{i\epsilon A} I(y^i = j)$$

IV. RESULTADOS

Inicialmente se prepararon los datos para poder ser analizados y procesados, aplicando ingenieria de caracteristicas se generaron 2 cracteristicas nuevas, el indice de masa corporal (IMC) y la presion arterial media (apmid), ademas a la caracteristica "age" se le aplico una conversion para pasar de días a años.

$$IMC = \frac{weigth}{(heigth/100)^2} \tag{1}$$

$$ap_{mid} = \frac{ap_{hi} + (2 * ap_{lo})}{3} \tag{2}$$

Al observar las estadisticas descriptivas de los datos iniciales se identificaron valores atipicos en las caracteristicas de presiones arteriales, edad y el indice de masa corporal. Por lo tanto se aplico la primera etapa de filtrado, donde los valores de presion arterial sistolica mayores a 210 o menores a 30 fueron eliminados, de igual manera, para la presion diastolica los datos mayores a 130 o menores a 40, estos valores fueron determinados teniendo en cuenta los rangos de normalidad para estas presiones con un grado de tolerancia, de manera que el filtrado no sea tan ajustado y solo se eliminen los valores realmente anormales.

Para la edad se identifico que habian muy pocos datos menores a 35 y esto generaria un sesgo en el analisis por lo que se determino eliminarlos. Y finalmente en el filtro correspondiente al IMC se eliminaron los valores mayores a 80 debido a que no se encuntran dentro del rango de normalidad para esta caracteristica.

En la segunda etapa de filtrado, se identificaron valores atipicos en las caracteristicas de peso y altura, que terminarian sesgando los resultados, por lo tanto se eliminaron los valores de peso mayores a 175 y de altura mayores a 200.

TABLA I
PRINCIPALES ESTADÍSTICAS DESCRIPTIVAS

Variable	mean	std	min	max
age	52.828	6.767	39	64
gender	1.348	0.477	1	2
height	164.393	7.969	91	198
weight	74.082	14.221	11	172
ap_hi	126.568	16.585	60	210
ap_lo	81.307	9.423	40	130
cholesterol	1.364	0.678	1	3
gluc	1.227	0.571	1	3
smoke	0.087	0.283	0	1
alco	0.053	0.225	0	1
active	0.803	0.397	0	1
cardio	0.494	0.499	0	1
IMC	27.456	5.255	3.471	74.380
ap_mid	96.394	10.936	46.667	156.666

Con respecto a la variable objetivo las correlaciones obtenidas en la matriz de correlación son:

TABLA II
CORRELACIÓN DE LA VARIABLE ”CARDIO”

Variable	correlación
Cardio	1.000
ap_hi	0.428
ap_mid	0.411
ap_lo	0.338
age	0.239
cholesterol	0.221
IMC	0.188
weight	0.179
gluc	0.089
gender	0.007
alco	−0.008
height	−0.011
smoke	−0.016
active	−0.037

Las variables cómo la presion arterial (alta, media y baja), la edad y el colesterol son las que presentan mayor relación con la ocurrencia de enfermedades cardiovasculares de una forma directa, es decir entre mayor sea el valor de la variable mayor es la probabilidad de una

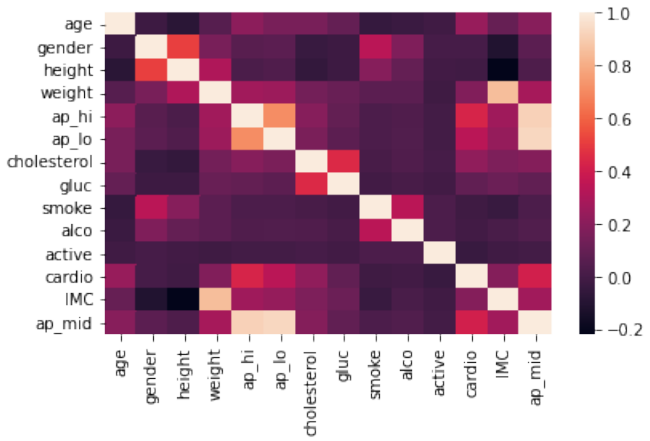

Fig. 7. Matriz de correlación

enfermedad cardiovascular. Mientras que variables como ”smoke”o ”active” presentan una correlación inversa, es decir que si un paciente no fuma y es activo correrá menos riesgo de sufrir una enfermedad cardiovascular.

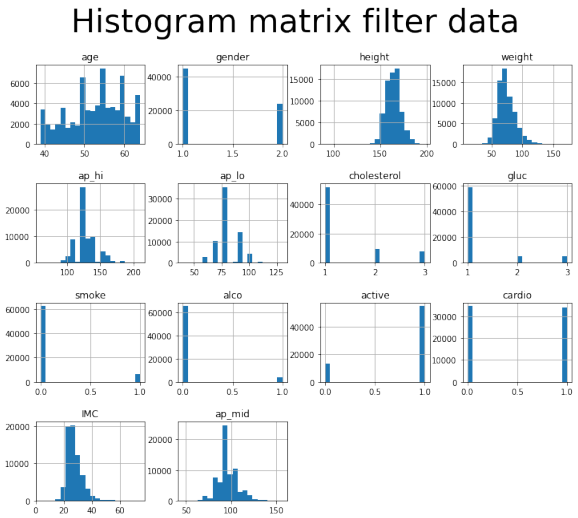

Fig. 8. Matriz de histogramas

Se puede notar que varias de las carateristicas son de tipo booleano, donde se evidencia que la mayoria de los pacientes son de genero masculino, ademas de que tienen habitos saludables (no fuman, no consumen alcohol y realizan actividad fisica). Ademas de que tienen niveles de colesterol y glucosa normales, solo una pequeña parte de los pacientes tienen niveles anormales o altos pero dentro del rango de normalidad.

Las edades de los pacientes estan bien distibuidas entre 39 y 64 años, y para las caracteristicas de altura, peso, presiones arteriales e indice de masa corporal se presenta una distibucion gaussiana tipificada.

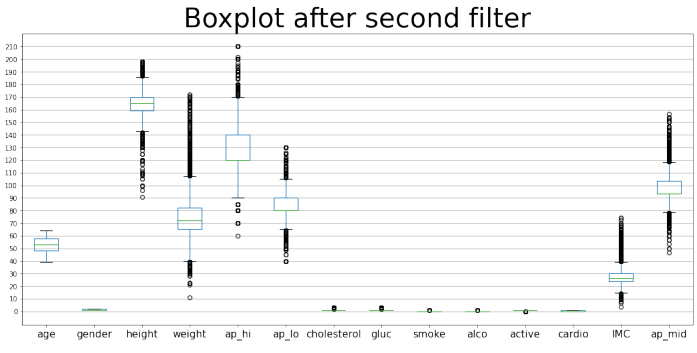

Fig. 9. Diagrama de cajas

En este diagrama tambien se puede observar la distribucion de los datos, aunque fue empleado principalmente para evaluar el comportamiento de los filtros aplicados.

Con el objetivo de comparar las clasificaciones realizadas mediante diferentes modelos de machine learning se presenta la clasificación obtenida de los datos originales.

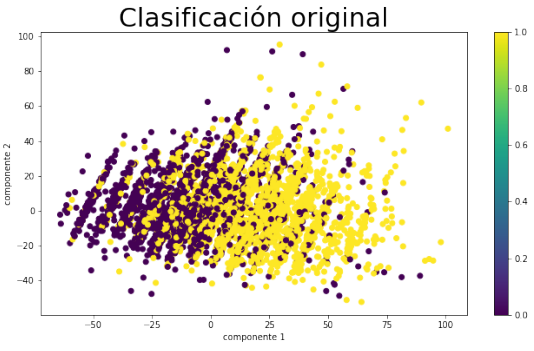

Fig. 10. Clasificación original

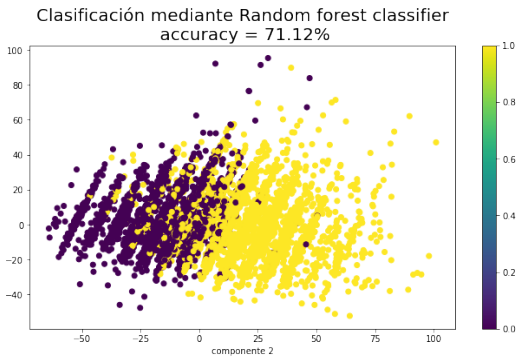

Fig. 11. Clasificación RFC

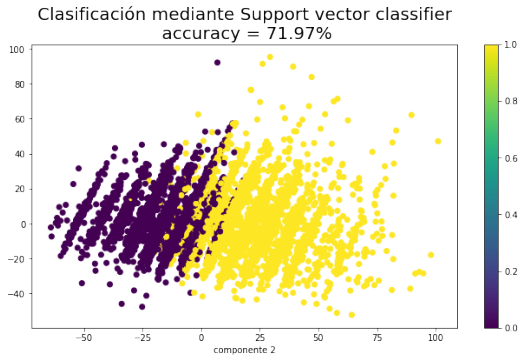

Fig. 12. Clasificación SVC

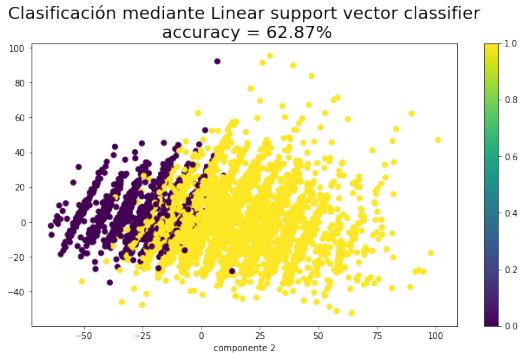

Fig. 13. Clasificación linear SVC

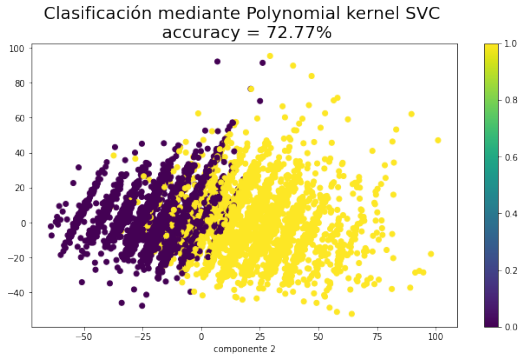

Fig. 14. Clasificación kernel polinomial SVC

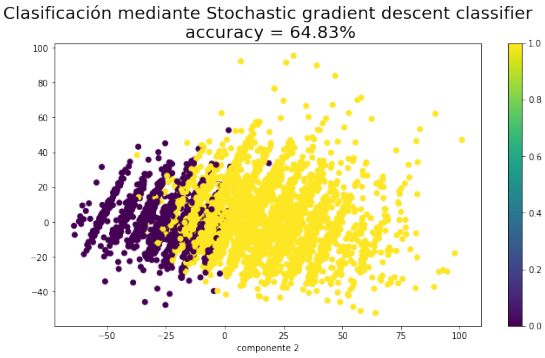

Fig. 15. Clasificación SGD

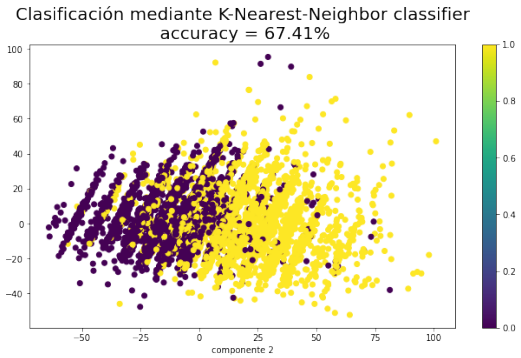

Fig. 16. Clasificación KNN

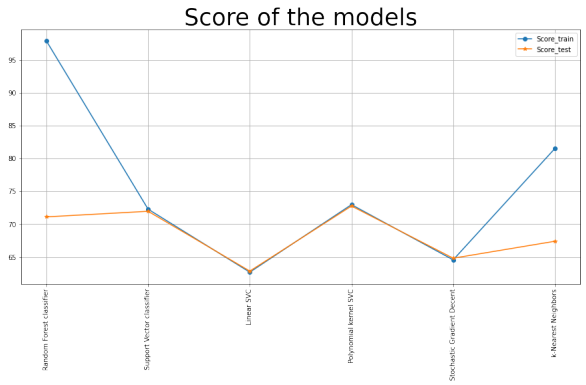

Fig. 17. Comparacion scores test-train modelos

A. *Comparación*

Cuando de observan los graficos de clasificacion de cada modelo presentados anteriormente y se comparan con el grafico de clasificacion orginal, se podria plantear que visualmente el grafico con mayor similitud al original podria ser el de la clasificacion KNN, ya que a diferencia de los demas la separacion de las clases se ve mas distribuida en el espacio, caso contrario el de la clasificacion SVC en el que se puede notar un frontera de caracter mas lineal que nos separa las clases. En los casos de las clasificaciones SGD y SVC linear, tenemos un comportamiento muy similar de ambos modelos en donde la frontera conserva una tendencia similar al modelo SVC pero un poco mas desplazada hacia la izquierda y finalmente tenemos los modelos RFC y kernel polinomial SVC, en ambos de estos modelos el comportamiento del grafico es similar con fronteras ya no tan lineales que asemejan un posible mejor desempeño en la clasificacion.

Si se contrasta el analisis grafico realizado anteriormente con los scores obtenidos de cada modelos (Tabla III) y el grafico de estos, se puede establecer que efectivamente el modelo que presenta una mejor calificacion en el SVC kernel polinomial, para el caso de los modelos RFC y KNN a pesar de que graficamnete aparentan tener una prediccion adecuada y presentan un score en el entrenamiento mas alto, en los datos predichos su score es considerablemente mas bajo lo que nos indica que en ambos modelos se tiene un sobre entrenamiento.

Para los modelos aplicados se obtuvieron aciertos de un nivel alto tanto para el entrenamiento como para las predicciones, sin embargo no son los mejores dado que se encuentran en valores aproximados de un 70% a excepción de el entrenamiento que tuvo el random forest classifier que fue muy bueno con un valor de 97.88% de acierto lo cual indica en relación con el score de los datos predichos que se presentó overfitting porque la diferencia con respecto a la predicción es mayor al 26%.

TABLA III
RESULTADOS OBTENIDOS

Modelo	Scores		
	Train	Test	Diff
Random forest	97.88	71.12	26.76
Support vector	72.29	71.97	0.32
Linear SVC	62.70	62.87	0.17
Kernel polinomial SVC	72.98	72.77	0.21
Stochastic gradien decent	64.58	64.83	0.25
K-nearest neighbors	81.52	67.41	14.11

Por otra parte, el método de kernel polinomial SVC tuvo el mejor comportamiento a la hora de predecir las clases de los datos de test siendo coherente con el acierto de entrenamiento dado que la diferencia es de 0.17% entre el entrenamiento y las pruebas. Por lo tanto, la varianza de los datos genera que se implementen métodos que permiten generar fronteras de decisión menos sesgadas.

V. CONCLUSIÓN

La clasificación realizada mediante diferentes métodos de machine learning indicó que dadas las condiciones de vida que tienen los sujetos del estudio se esperará en mayor cantidad la presencia de una enfermedad cardiovascular, sin embargo se conoce que cierto grupo de personas con hábitos saludables se encuentran en riesgo mientras que otras con hábitos poco saludables gozan de una buena salud. Si bien se recopilaron una gran cantidad de variables asociadas a enfermedades cardiovasculares no se tienen en cuenta diversos factores como los antecedentes familiares, demás afecciones en la salud de los pacientes que puedan conllevar a una enfermedades del corazón, por lo tanto se presentan limitaciones en los modelos realizados en la presente investigación.

BIBLIOGRAFÍA

[1] E. A. Pró. Anatomía Clínica. 2a edición. Buenos Aires, Argentina: Panamericana, 2016.
[2] K. L. Moore. Anatomía con orientación clínica. 8a edición. Barcelona, España: Wolters Kluwer, 2017.
[3] G. J. Tortora. Principios de Anatomía y Fisiología. 15a edición. Bogotá, Colombia: Panamericana, 2018.