Algoritmos de clasificación para predecir ataques al corazón

Modelos y Simulación de Sistemas II

Arley Giovany Berrio Arroyave
Deyber Sepulveda Tuberquia
Jhonatan Alejandro Torres Vasco

Prof. María Bernarda Salazar Sánchez
Departamento de Ingeniería de Sistemas
Universidad de Antioquia, Colombia
bernarda.salazar@udea.edu.co

Resumen:
En este documento se encuentra un primer acercamiento al análisis de la predicción de enfermedades cardiovasculares, en el cual se encuentra una identificación de las características que se van a utilizar y una revisión de algunos antecedentes a casos similares, es realizado a partir de algoritmos de clasificación para encontrar el mejor modelo de predicción.

Palabras clave: machine learning, clasificación, aprendizaje automatizado, algoritmo, validación .

I.

A. Parte I: Comprensión del problema de aprendizaje automático

1.

Las enfermedades cardiovasculares se encuentran entre las enfermedades crónicas más prevalentes a nivel mundial, siendo la causa principal de muertes con un aproximado de 17.9 millones de vidas cada año. En los Estados Unidos afecta a millones de estadounidenses cobrando aproximadamente 647.000 vidas cada año. Si bien existen diferentes tipos de enfermedades cardiovasculares, la mayoría de las personas se percatan de que tienen la enfermedad después de desarrollar síntomas como dolor en el pecho o un paro cardíaco repentino. Este hecho destaca la importancia de las medidas preventivas y las pruebas que puedan predecir con precisión esta condición en la población antes de que se produzcan resultados negativos como infarto o muertes.

De acuerdo con Los Centros para el Control y la Prevención de Enfermedades, padecer de presión arterial y colesterol alto, así como también tener hábitos nocivos para la salud como fumar son factores de riesgo clave para el desarrollo de enfermedades cardíacas. Además, El Instituto Nacional del Corazón, los Pulmones y la Sangre destaca una gama más amplia de factores como la edad, el medio ambiente, la ocupación, los antecedentes familiares, la genética, los hábitos de estilo de vida, otras afecciones médicas, la raza o la etnia y el sexo que podrían ser influyentes en el diagnóstico de enfermedades cardiovasculares. Por tanto, se busca brindar una herramienta que pueda ayudar a predecir el padecimiento de esta enfermedad en base a los factores anteriormente mencionados. Para esto se desea trabajar en un **modelo de clasificación** que relacione las variables de entrada (como por ejemplo: nivel de presión arterial, nivel de colesterol, estilo de vida y hábitos, género, edad y presencia de otras enfermedades) para brindar una salida bastante aproximada a la realidad y con la que se pueda realizar inferencias sobre el estado de salud de la persona y prevenir a tiempo esta enfermedad y sus impactos negativos sobre una persona.

Esta clasificación será de gran ayuda en el campo de la medicina, pues se podría actuar más rápido en caso de que una persona marque positiva a padecer una enfermedad cardiaca

2.

Dataset

El dataset utilizado para este proyecto puede ser encontrado en [1]. Contiene 22 columnas y un total de 253680 registros.

Variables

En el dataset se pueden encontrar las siguientes variables:

- **HeartDiseaseorAttack**: es la variable de interés en este modelo, indica si una persona ya ha padecido algún ataque cardíaco o presenta alguna enfermedad cardiovascular.
- **HighBP:** indica si una persona sufre de presión alta en la sangre.
- **HighChol:** indica si una persona sufre de colesterol alto.
- **CholCheck:** indica si una persona se ha hecho control o examen de colesterol en los últimos 5 años.
- **BMI:** índice de masa corporal.
- **Smoker:** indica si una persona ha fumado al menos 100 cigarrillos (5 paquetes) en su vida.
- **Stroke:** indica si una persona ha sufrido un derrame cerebral.
- **Diabetes:** indica si una persona padece diabetes o prediabetes.
- **PhysActivity:** indica si una persona se ha ejercitado o ha realizado alguna actividad física en los últimos 30 días.
- **Fruits:** indica si una persona consume frutas una o más veces por día.
- **Veggies:** indica si una persona consume vegetales una o más veces por día.
- **HvyAlcoholConsump:** indica si una persona es bebedora compulsiva (para los hombres: 14 bebidas por semana y para las mujeres: 7 bebidas por semana).
- **AnyHealthCare:** indica si una persona tiene seguro médico.
- **NoDocbcCost:** indica si en los últimos 12 meses una persona no pudo asistir al doctor por falta de dinero.
- **GenHlth:** indica cómo una persona considera que se encuentra de salud en un nivel del 1 al 5.
- **MentHlth:** considerando factores como estrés, depresión y desorden emocional, indica cuántos días de los últimos 30 días una persona considera que su salud mental no ha sido óptima.
- **PhysHlth:** indica cuántos días de los últimos 30 días una persona considera que su salud física no ha sido óptima.
- **DiffWalk:** indica si una persona tiene dificultades para caminar o subir escaleras.
- **Sex:** género de la persona.
- **Age:** categoría de edad de 14 niveles a la que pertenece la persona.
- **Education:** indica la nota máxima que ha tenido una persona o en su defecto, el último año escolar completado.
- **Income:** ingreso anual.

Datos nulos o faltantes

El dataset no contiene datos nulos o faltantes y todos los datos son de tipo flotante, es por esto que no se llenó ningún vacío.

Datos duplicados

Se evidenció que el dataset contenía 23899 datos duplicados.

3.

Primer articulo[2]

1.Las técnicas usadas fueron:

- J48
- Naïve Bayes
- REPTREE
- CART
- Bayes Net

2.La metodología de validación que usaron fue:

Los algoritmos se aplican en el conjunto de datos utilizando Stratified Cross-Validation 10 veces para evaluar el desempeño de técnicas de clasificación para predecir una clase.

3.Resultados obtenidos:

La investigación muestra que los resultados no presentan una diferencia muy grande en la predicción cuando se utilizan diferentes algoritmos de clasificación.
La precisión predictiva determinada por los algoritmos J48, REPTREE y SIMPLE CART mostraron
que los parámetros utilizados son indicadores confiables para predecir la presencia de enfermedades del corazón.

Segundo artículo[3]

1.Las técnicas usadas fueron:

- J48
- Decision Tree
- K Nearest Neighbors(KNN)
- Naive Bayes(NB)
- SMO

2.La metodología de validación que usaron fue:

> Este estudio empleó una validación cruzada de 10 veces en la construcción del modelo de clasificación y eficiencia de evaluación. Este método aumenta la validación de clasificación y previene los resultados aleatorios e inválidos.

3.Resultados obtenidos:

> Los resultados de la comparación mostraron que el árbol de decisión j48 alcanzó el valor más alto en precisión, sensibilidad, especificidad, medida F y las medidas de rendimiento de precisión.
> El modelo predictor óptimo de enfermedades cardíacas obtenido en este estudio, adoptó el árbol de decisión j48 como algoritmo de clasificación.

Tercer artículo[4]

Técnicas de aprendizaje

- Árboles de decisión
- Regresión logística
- Clasificación Naive Bayes
- Random Forest
- Nearest centroid

Metodología de validación

Para evitar el sobreajuste cómo metodología de validación utilizaron Validación Cruzada.

Resultados

Los algoritmos con mayor F1-score fueron regresión logística y random forest con un puntaje de 85% y tienen una exactitud del 90%.

La prioridad del modelo era el valor F porque su interés es evitar los falsos negativos sin perder precisión, en general todos los algoritmos alcanzan buenos resultados.

- Naive Bayes: ha obtenido 8 falsos positivos.
- Árboles de decisión y Nearest centroid: han obtenido 7 falsos positivos.
- Regresión logística y random forest: han obtenido 3 falsos positivos.

Por lo que Regresión Logística y Random Forests son los dos algoritmos mejores para entrenar el modelo que ellos están utilizando.

Cuarto artículo[5]

La técnica que usan son redes bayesianas y árboles de decisión.

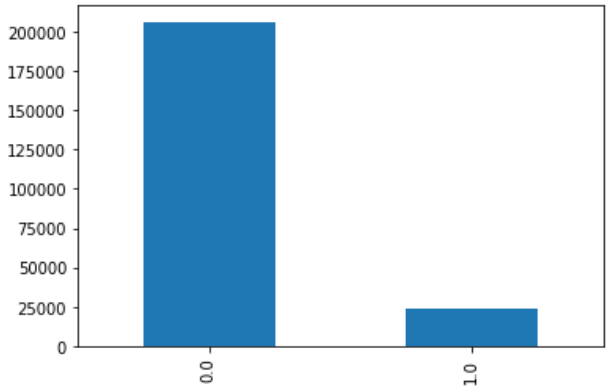
Metodología: si se tiene un grupo de datos S donde contenga valores positivos o negativos sobre un concepto dicotómico en el estudio, para calcular la entropía de S relativa, su clasificación booleana se debe definir Pn=1- Pp
Pp. es la probabilidad de que las respuestas sean positivas según el conjunto S.
Pn. es la probabilidad de que las respuestas sean negativas según el conjunto S.

Resultados: Se demostró que es posible determinar la administración, o no, de un procedimiento clínico a un paciente con síntomas de enfermedad cardiovascular, usando las variables tales como: presión arterial, gota, Hipotiroidismo; mediante la utilización del modelo híbrido, utilizando las ventajas de los árboles de decisión en la clasificación de los datos y las probabilidades en las redes bayesianas.

4.

La base de datos que se usará es Heart Disease Health Indicators Dataset[1] que cuenta con 253680 muestras y 22 variables dentro de las cuales se encuentra nuestra variable de interés que es HeartDiseaseorAttack la cual será nuestra variable a predecir, en la cual se implementara una metodología de validación cruzada (Cross-validation).

Como la base de datos presentaba un desbalance como se ve a continuación:



Se utilizó el método UnderSampling viendo el siguiente resultado:

```
Before UnderSampling, counts of label '1': 23717
Before UnderSampling, counts of label '0': 206064

After UnderSampling, the shape of train_X: (229781, 21)
After UnderSampling, the shape of train_y: (229781,)

After UnderSampling, counts of label '1': 23717
After UnderSampling, counts of label '0': 23717
```

Se usó las siguientes medidas de desempeño:

- Matriz de confusión
- F1-Score
- ROC Curve

5. Se realiza la evaluación de los siguientes modelos de predicción, y se implementan sus debidas métricas de evaluación, los resultados fueron los siguientes:

• Análisis discriminante Cuadrático[6]:

```
Model Score: 82.55 %
Precision: 0.28866505163840367
F1 score: 0.36085431941345236
ROC-AUC score: 0.6729741477513653
Confusion Matrix:
 [[35673  5579]
 [ 2441  2264]]

{'Algorithm': 'Análisis discriminante Cuadrático',
 'Model Score': '82.55%',
 'Precision': 0.29,
 'F1 score': 0.36,
 'ROC-AUC score': 0.67}
```

• Gradient Boosting Tree[6] :

```
Model Score: 90.06 %
Precision: 0.5811688311688312
F1 score: 0.19035802906770652
ROC-AUC score: 0.5522175471695148
Confusion Matrix:
 [[40852   387]
 [ 4181   537]]

{'Algorithm': 'Gradiente Boosting Tree',
 'Model Score': '90.06%',
 'Precision': 0.58,
 'F1 score': 0.19,
 'ROC-AUC score': 0.55}
```

• Redes Neuronales Artificiales[6]:

```
Model Score: 90.1 %
Precision: 0.5650723025583982
F1 score: 0.1825696316262354
ROC-AUC score: 0.5497016601031955
Confusion Matrix:
 [[40900   391]
 [ 4158   508]]

{'Algorithm': 'Redes Neuronales Artificiales',
 'Model Score': '90.1%',
 'Precision': 0.57,
 'F1 score': 0.18,
 'ROC-AUC score': 0.55}
```

• Máquinas de Soporte Vectorial[6]:

```
Model Score: 89.92 %
Precision: 0.5894736842105263
F1 score: 0.08818897637795275
ROC-AUC score: 0.5219391989702328
Confusion Matrix:
 [[41101   156]
 [ 4476   224]]

{'Algorithm': 'Máquinas de Soporte Vectorial',
 'Model Score': '89.92%',
 'Precision': 0.59,
 'F1 score': 0.09,
 'ROC-AUC score': 0.52}
```

A continuación se realiza una tabla comparativa de cada modelo evaluado, estas se encuentran ordenadas de mejor a peor modelo.

En la comparativa[7] se puede evidenciar cada puntaje arrojado por las métricas de evaluación.

Por tales valores se decide optar por el modelo de gradient boosting tree, al tener un mejor puntaje del

modelo y en cada medida de desempeño. También se evidencia que el true positive y el false positive en la matriz de confusión correspondiente a este modelo fueron los más altos, es decir, se logra una mejor predicción y menos tasa de errores.

6. Para el análisis de cada una de las características se empezó determinando el nombre de cada una como se ve a continuación :

```
['HeartDiseaseorAttack', 'HighBP', 'HighChol', 'CholCheck', 'BMI',
 'Smoker', 'Stroke', 'Diabetes', 'PhysActivity', 'Fruits',
 'Veggies', 'HvyAlcoholConsump', 'AnyHealthcare', 'NoDocbcCost',
 'GenHlth', 'MentHlth', 'PhysHlth', 'DiffWalk', 'Sex', 'Age',
 'Education', 'Income'], dtype=object)
```

Verificamos la cantidad de registros por variable, la media y la desviación estándar:

	HeartDiseaseorAttack	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke
count	128681.000000	128681.000000	128681.000000	128681.000000	128681.000000	128681.000000	128681.000000
mean	0.093510	0.429776	0.426543	0.963522	28.443166	0.447929	0.041335
std	0.291147	0.495046	0.494577	0.187477	7.010712	0.497283	0.199064
min	0.000000	0.000000	0.000000	0.000000	12.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	1.000000	24.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	1.000000	27.000000	0.000000	0.000000
75%	0.000000	1.000000	1.000000	1.000000	31.000000	1.000000	0.000000
max	1.000000	1.000000	1.000000	1.000000	98.000000	1.000000	1.000000

rows × 22 columns

Diabetes	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
128681.000000	128681.000000	128681.000000	...	128681.000000	128681.000000	128681.000000	128681.000000	128681.000000	128681.000000	128681.000000	128680.000000	128680.000000	128680.000000
0.296058	0.760633	0.638627	...	0.951065	0.081931	2.503934	3.183482	4.222628	0.166155	0.459171	8.051910	5.078691	6.101041
0.697781	0.426699	0.480573	...	0.214051	0.274261	1.067940	7.346822	8.687006	0.372221	0.498288	3.048847	0.978253	2.057898
0.000000	0.000000	0.000000	...	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000
0.000000	1.000000	0.000000	...	1.000000	0.000000	2.000000	0.000000	0.000000	0.000000	0.000000	6.000000	4.000000	5.000000
0.000000	1.000000	1.000000	...	1.000000	0.000000	2.000000	0.000000	0.000000	0.000000	0.000000	8.000000	5.000000	7.000000
0.000000	1.000000	1.000000	...	1.000000	0.000000	3.000000	2.000000	3.000000	0.000000	1.000000	10.000000	6.000000	8.000000
2.000000	1.000000	1.000000	...	1.000000	1.000000	5.000000	30.000000	30.000000	1.000000	1.000000	13.000000	6.000000	8.000000

Adicionalmente se puede evidenciar que no se cuenta con datos nulos por característica:

```
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   HeartDiseaseorAttack  253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   Diabetes              253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
```

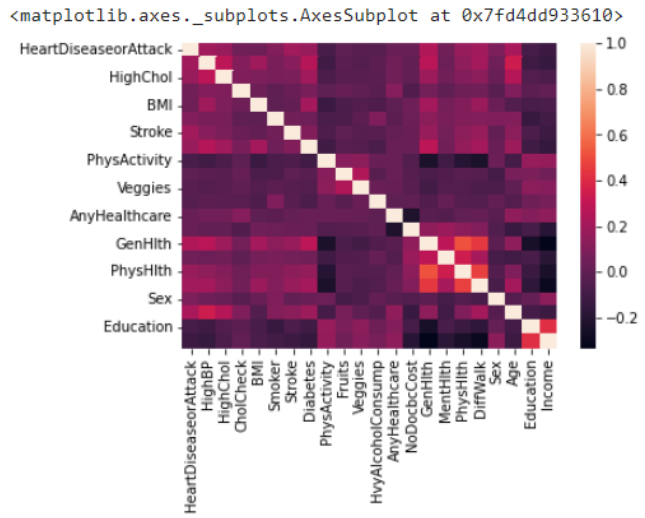
También se verifica que hay 23899 datos duplicados y se procede a eliminarlos:

	HeartDiseaseorAttack	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	Diabetes	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	Physl
1242	0.0	1.0	1.0	1.0	27.0	1.0	0.0	2.0	0.0	0.0	...	1.0	0.0	5.0	0.0	
1563	0.0	0.0	0.0	1.0	21.0	1.0	0.0	0.0	1.0	1.0	...	1.0	0.0	1.0	0.0	
2700	0.0	0.0	0.0	1.0	32.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	0.0	
3160	0.0	0.0	0.0	1.0	21.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	1.0	0.0	
3332	0.0	0.0	0.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	1.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
253492	0.0	1.0	1.0	1.0	33.0	0.0	0.0	2.0	1.0	1.0	...	1.0	0.0	3.0	0.0	
253650	0.0	0.0	0.0	1.0	25.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	1.0	0.0	
253653	0.0	0.0	1.0	1.0	24.0	1.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	0.0	
253697	0.0	0.0	0.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	0.0	
253638	0.0	0.0	0.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	0.0	

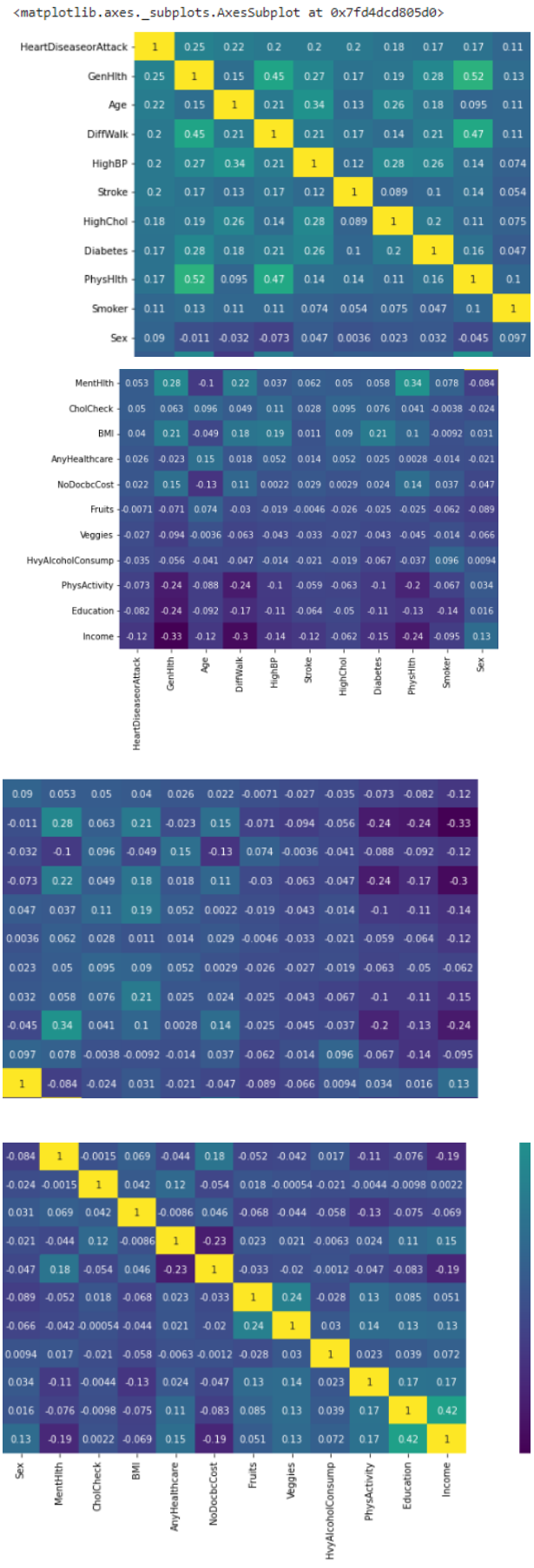
23899 rows × 22 columns

quedando con 229781 registros.

Finalmente se implementó la matriz de correlación con las 22 características que tiene nuestro caso de estudio, dando como resultado la siguiente matriz:



Con la finalidad de poder observar de una mejor forma la información se construyó una segunda matriz

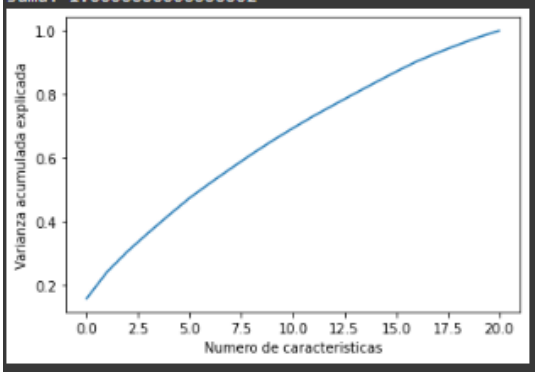


A partir de los datos obtenidos se pudo evidenciar que las variables GenHlth, Age, DiffWalk, Stroke, HighBP, PhysHlth, HighChol y Diabetes son las que tienen una mayor correlación positiva con la variable de interés, mientras que las variables Income, Education, PhysActivity, Veggies, HvyAlcoholConsump y Fruits son las que tienen una mayor correlación negativa con la variable de interés y por lo tanto son candidatas a ser eliminadas.

7. Se realiza la extracción de características por el método PCA que nos arroja los siguientes resultados:

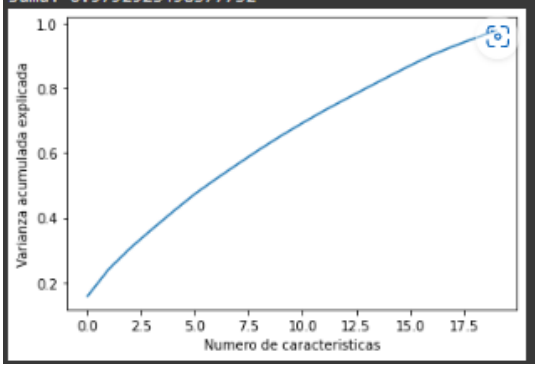
Con las 21 características se observa lo que se espera que el modelo representa el 100%

```
shape of X_pca (229781, 21)
[0.15763868 0.08432812 0.06489052 0.05774794 0.05577682 0.05379117
 0.04710458 0.04557982 0.0444467  0.04226643 0.03964826 0.03817319
 0.03567269 0.03518068 0.03464135 0.03392342 0.03227366 0.02706621
 0.02514719 0.0239949  0.02070765]
suma: 1.0000000000000002
```



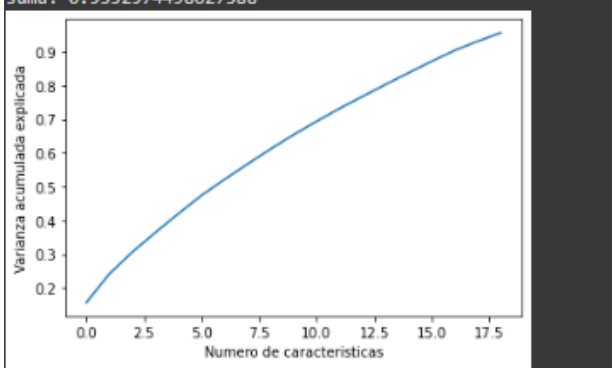
En la siguiente gráfica de variabilidad explicada acumulada, vemos que tomando los primeros 20 componentes llegamos al 97.9%:

```
shape of X_pca (229781, 20)
[0.15763868 0.08432812 0.06489052 0.05774794 0.05577682 0.05379117
 0.04710458 0.04557982 0.0444467  0.04226643 0.03964826 0.03817319
 0.03567269 0.03518068 0.03464135 0.03392342 0.03227366 0.02706621
 0.02514719 0.0239949 ]
suma: 0.9792923498377732
```



Ahora tomamos 19 características y llegamos a un 95.5%:

```
shape of X_pca (229781, 19)
[0.15763868 0.08432812 0.06489052 0.05774794 0.05577682 0.
 0.04710458 0.04557982 0.0444467  0.04226643 0.03964826 0.
 0.03567269 0.03518068 0.03464135 0.03392342 0.03227366 0.
 0.02514719]
suma: 0.9552974498627588
```



Así se continúa el proceso llegando a los resultados observados en la **siguiente tabla:**

Número de características	**Porcentaje explicado según tabla de variabilidad**
21	100%
20	97.9%
19	95.5
18	93%
17	90%

8. Al tener en cuenta la precisión y el puntaje del modelo se puede generar una discusión entorno a cuál de los algoritmos representa mejor la solución del problema, aunque el puntaje del modelo para algunos algoritmo no cambio casi, a la hora de comparar la precisión si se abría más la brecha entre estos, dando luces de cuál sería el mejor, sin embargo fue la matriz de confusión la que permitiría abrir más la brecha y dar la respuesta positiva para el algoritmo Gradient Boosting Tree, también es rescatable lo que se puede deducir a partir de la métricas de validación que aunque solo fueron 3 su resultado fue significativo para cada una de las opciones, pues como ya se observó, determinaron el que finalmente sería el algoritmo óptimo para predecir si alguien es propenso a sufrir de un ataque al corazón.

Ahora si comparamos los resultados que se obtuvieron y los comparamos con los artículos 1[2] y 2[3] se observa que la validación que se hizo, varía significativamente debido a los algoritmos usados, pues mientras en el primero se usa J48, Naïve Bayes, REPTREE,CART, Bayes Net y en el segundo J48, Decision Tree, K Nearest , Neighbors(KNN), Naive Bayes(NB) , SMO los que se realizaron en este documento fueron Análisis discriminante Cuadrático, Gradient Boosting Tree, Redes Neuronales Artificiales y Máquinas de Soporte Vectorial esto lleva a un análisis distinto, empezando por la medidas de desempeño pues en este artículo se implementó la Matriz de confusión, F1-Score, ROC Curve a diferencia de la implementada en los artículos mencionados la cual fue el cross validation el cual se implementó con 10 repeticiones.

El modelo Gradient boosting tree, cuenta con un Model Score de 90.06% y una de las precisiones más altas de los cuatro modelos utilizados con 0.5811, además de que similar cómo en el tercer artículo[4] es uno de los que tiene más puntaje en el F1 score y de los que menos falsos positivos entrega.

En el artículo 4[5] se utilizan árboles de decisión y el método ID3 de los mismos, utilizando una propiedad estadísticas llamada ganancia de información, que mide cómo clasificar cada atributo, es decir, elige el nodo del árbol que tenga mayor ganancia de información y luego expande sus ramas, buscando establecer las características de mayor relevancia para el modelo, en nuestro caso buscamos establecer las características de menor importancia para el modelo con la finalidad de reducir el ruido que esas puedan estar generando.

9. Sustentación[8]

Bibliografía:

[1] Taboul, A. (s.f.). Heart Disease Health Indicators Dataset. Recuperado de: Heart Disease Health Indicators Dataset

[2]Hlaudi Daniel Masethe, Mosima Anna Masethe (2014, 22 Octubre). Prediction of Heart Disease using Classification Algorithm
(10) Prediction of Heart Disease using Classification Algorithms | Dan Masethe - Academia.edu

[3]Boshra Bahrami, Mirsaeid Hosseini Shirvani, Prediction and Diagnosis of Heart Disease by Data Mining Techniques. JMEST.

https://www.jmest.org/wp-content/uploads/JMESTN42350475.pdf

[4]Moreno Sánchez, Juan Sebastián (2021, 30 julio). Predicción de ataques cardiacos mediante técnicas de Machine Learning. Proyecto Fin de Carrera / Trabajo Fin de Grado, E.T.S.I. de Sistemas Informáticos (UPM), Madrid. https://oa.upm.es/68821/1/TFG_JUAN_SEBASTIAN_MORENO_SANCHEZ.pdf

[5] Martínez, S. G. R. (2011, 22 julio). *Modelo híbrido para el diagnóstico de enfermedades cardiovasculares basado en inteligencia artificial*. Modelo híbrido para el diagnóstico de enfermedades cardiovasculares basado en inteligencia artificial. Recuperado 29 de agosto de 2022, de Modelo hibrido para el diagnóstico de enfermedades cardiovasculares basado en inteligencia artificial

[6] Enlace del GitHub

[7] Tabla comparativa de cada modelo

[8] Enlace del video de sustentación