Análisis de la correlación entre las condiciones de salud previas y ataques al corazón para crear un modelo de predicción

Aldo Jair Valdiviezo Sanchez

Abstract

Utilizando maching learnig para determinar la posibilidad de que una persona tenga un ataque al corazón debido a las condiciones pre-médicas. Esto representa un tema importante para la salud de la población en general motivando que se genere un cambio a la población de riesgo utilizando un repositorio proporcionado por Kaggle entrenamos diferentes modelos de clasificación utilizando las condiciones previas características de los pacientes que ingresan a los hospitales. Con una sola entrada los modelos pueden hacer predicciones que coinciden con las notificadas por el dataset.

1 Introducción

1.1 Datos

Se utilizaron datos sobre condiciones médicas previas en la población proporcionados por el Sitio web Kaggle. Los datos están distribuidos en 14 condiciones medicas previas, al igual que la edad, las cuales van desde un valor de 29 a los 77 años de edad 6, hasta llegar al nuestra variable objetivo la cual es, si el paciente presento o no presento un ataque al corazón.

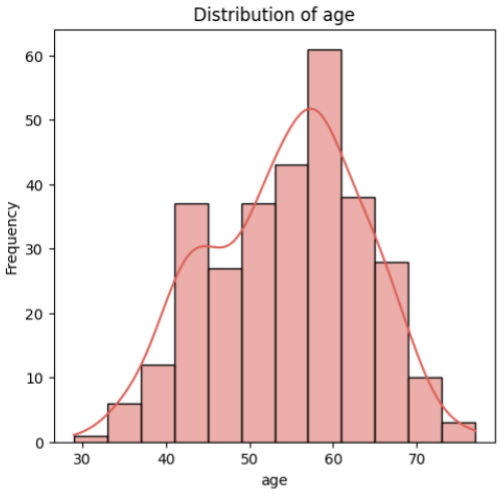

Figure 1: Distribución de edades que van de los 29 a los 77 años de edadt

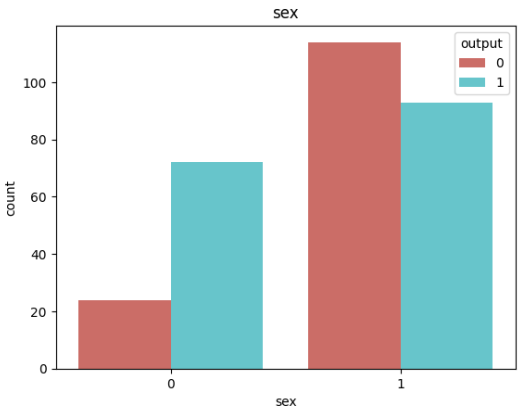

Figure 2: distribución de ataques al corazón entre los dos hombres y mujeres

2 Modelo

Nuestro proceso general consistió en extraer características de los pacientes y sus afecciones médicas previas para entrenar y utilizar de forma iterativa diferentes modelos de predicción, como **redes neuronales**, **Random Forest**, **SVM** (Suport Vector Machine), **KNN**, **Decision Tree**, **Logistic Regression** y **Naive Bayes**. Utilizamos la búsqueda en cuadrícula para obtener los mejores hiperparámetros y la selección del modelo.

2.1 Caracteristicas

- age: Edad
- sex: Sexo 1 es para hombre 0 es para mujer
- cp: dolor de pecho
  - Valor 1: angina típica
  - Valor 2: angina atípica
  - Valor 3: dolor no anginoso
  - Valor 4: Asintomatico
- trtbps: Presión arterial en reposo
- chol: colesterol
- fbs: glucemia en ayunas ($>$ 120 mg/dl, 1 = verdadero; 0 = falso)
- restecg: resultados electrocardiogramas en reposo
  - Valor 0: normal
  - Valor 1: con anomalía de la onda ST-T (inversión de la onda T y/o elevación o depresión del ST $>$ 0,05 mV)
  - Valor 2: hipertrofia ventricular izquierda probable o definida según los criterios de Romhilt-Estes
- thalachh: Frecuencia cardíaca máxima alcanzada
- exng: Angina inducida por el ejercicio (1 = si; 0 = no)
- oldpeak: depresión del ST causada por la actividad en comparación con el reposo
- slp: la pendiente del pico del segmento ST de ejercicio
  - 0: Pendiente descendente
  - 1: Plano
  - 2: Pendiente ascendente
- caa: Número de vasos sanguíneos mayores (0–3)
- thall: Un trastorno sanguíneo llamado talasemia
  - Valor 0: NULL (eliminado previamente del conjunto de datos)
  - Valor 1: Defecto fijo (ausencia de flujo sanguíneo en alguna parte del corazón)
  - Valor 2: Flujo sanguíneo normal
  - Valor 3: Defecto reversible (se observa un flujo sanguíneo pero no es normal)
- output: Cardiopatía (1 = no, 0 = sí)

2.2 División de datos para el entrenamiento

La base de datos utilizada para entrenar la red neuronal artificial se dividió en un 70% para el conjunto de entrenamiento y un 30% para el conjunto de prueba. el conjunto de entrenamiento y el 30% para el conjunto de prueba. Se buscó que el número de etiquetas entre las personas que tuvieron una cariopatia y las que no se distribuyera por igual en ambos conjuntos, como puede verse en la figura 3.

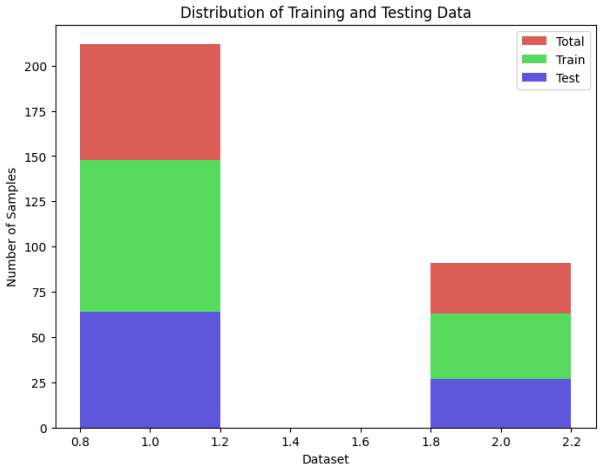

Figure 3: Distribución de los datos de validación y entrenamiento del conjunto total de datos.

3 Resultados

Utilizan los modelos **'Logistic Regression'**, **'Decision Tree'**, **'Random Forest'**, **'SVM'**, **'KNN'**, **'Gradient Boosting'**, **'XGBoost'**, **'AdaBoost'**, **'Naive Bayes'**, **'MLP Neural Network'**, para clasificar eficientemente el estado final del paciente. utilizaremos 3 métodos de para encontrar las mejores características para nuestros modelos, utilizando 3 metodologías, la primera es utilizando un atributo de los modelos de árboles de decisión y ensambles de árboles (como Random Forest) en scikit-learn, que mide la importancia de cada característica. La siguiente es utilizar el método de información mutua utilizando solo las 7 mejores relaciones entre nuestras variables, al igual que se utilizan las 7 mejores características que se midieron con la importancia, esto para comparar como reacciona nuestros modelos si solo utilizamos 7 características respecto a todas las características. utilizando la medición de la importancia de los modelos de árboles de decisión y de información mutua (tabla 1) obtenemos que el top 7 para ambas clasificaciones son diferentes, aun que comparten ciertas características.

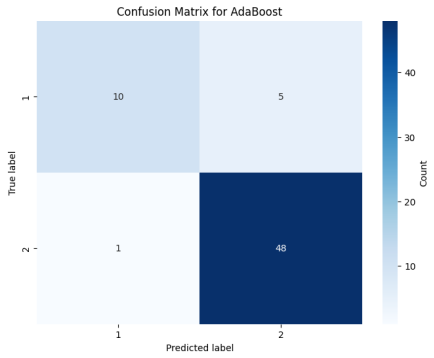

Figure 4: Matriz de confusion para el modelo AdaBoost

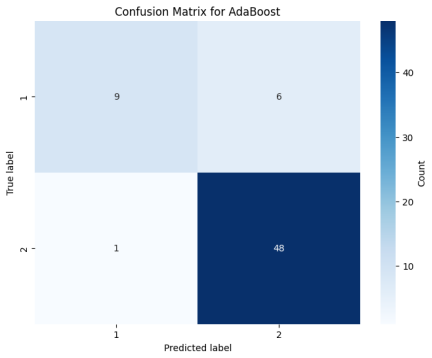

Figure 5: Matriz de confusion para el modelo XGBoost

Después de entrenar los modelos, los resultados obtenidos se muestran en la figura **??**. En ella, se puede observar que los dos mejores modelos son XGBoost y AdaBoost. Aunque ambos presentan un rendimiento destacado, XGBoost sobresale al alcanzar un 91% de precisión, lo cual es notable considerando que únicamente se utilizaron 7 características seleccionadas a través del método de información mutua. Por otro lado, AdaBoost también obtiene una precisión competitiva, pero lo hace empleando la totalidad de las características disponibles. Este comportamiento sugiere que XGBoost es más eficiente al manejar conjuntos de datos reducidos, seleccionando solo las variables más relevantes, mientras que AdaBoost parece beneficiarse del uso de un mayor número de características para alcanzar resultados comparables.

Resultados			
Característica	Importancia	Característica	IM
oldpeak	0.154751	thall	0.157433
thall	0.124284	cp	0.140007
cp	0.119302	caa	0.109471
thalachh	0.110617	oldpeak	0.102685
caa	0.107291	exng	0.088278
age	0.086428	slp	0.086248
chol	0.073123	chol	0.080257
exng	0.063940	thalachh	0.073537
trtbps	0.059306	sex	0.069749
sex	0.035273	fbs	0.047493
slp	0.034932	trtbps	0.030601
restecg	0.021596	restecg	0.019951
fbs	0.009156	age	0.000000

Table 1: Coches disponibles

4 Discusión

Los resultados obtenidos muestran como nuestras variables thall, cp, caa, son factores de riesgo mas altos para poder padecer un ataque al corazón como lo muestran las investigaciones consultadas en los artículos [1,2]. El entrenamiento de nuestros modelos muestra ademas una significativa precisión con la cual podemos apoyar a un mejor diagnostico de ataques al corazón dada las condiciones pre medicas de los pacientes, y poder comprender mejor cuales son los cambios pertinentes que se pueden llegar a dar en sus estilos de vida.

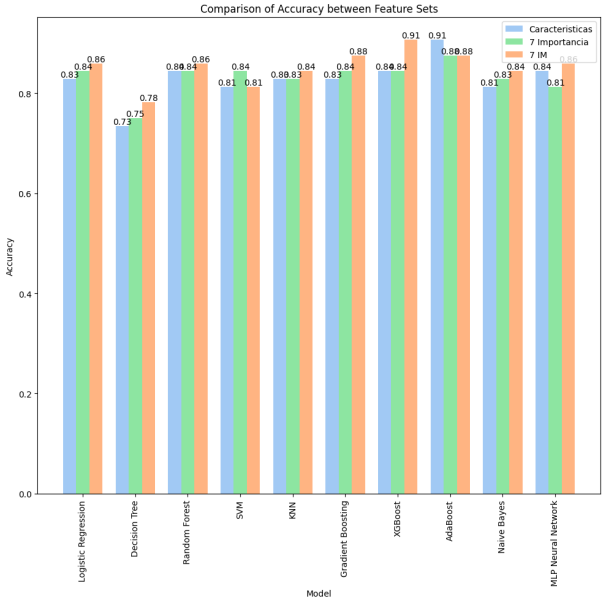

Figure 6: Resultados

5 Conclusiones

Utilizando modelos de regresión, arboles de decisión, métodos para encontrar las mejores características para entrenar nuestros modelos y así conseguimos clasificar las posibilidades de tener un ataque al corazón. Nuestra investigación concluye que para ciertas características como los son, los dolores al pecho, su intensidad, la gravedad de la talasemia (si la padece), ademas de los número de vasos sanguíneos mayores, nuestro modelo puede predecir con alta precisión si puedes o no, llegar a padecer de un paro cardíaco.

References

[1] Cohen, A. R., Galanello, R., Pennell, D. J., Cunningham, M. J., & Vichinsky, E. (2004). Thalassemia. ASH Education Program Book, 2004(1), 14-34.

[2] Vilchis, J. F. R., & Vásquez-Sánchez, L. Angina de pecho: diagnóstico y manejo inicial.

[3] GitHub