Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

Maestría de Ciencias de Datos

Predicción de enfermedades cardíacas utilizando Aprendizaje Automático

---

Profesor: José Alberto Benavides Vázquez

Actuario: Damián Atilano Martínez Alvarado

Marzo 2023

1. Introducción

El conjunto de datos de predicción de enfermedades cardíacas proporciona información vital sobre la relación entre los factores de riesgo y la salud cardíaca. Este conjunto de datos contiene 270 estudios de casos de personas clasificadas con o sin enfermedades cardíacas según los resultados de los cateterismos cardíacos, el estándar de oro en la evaluación de la salud cardíaca. Cada paciente se identifica mediante 7 variables.

- Edad: La edad del paciente.
- Sexo: Género del paciente, si es hombre (0) o si es mujer (1).
- TDP: Tipo de dolor en el pecho calificado del 1 al 4, siendo el 4 mas fuerte.
- Colesterol: Nivel de colesterol del paciente.
- FC: Frecuencia cardíaca.
- EA: Ejercicio de angina, siendo 0 bueno y 1 malo.
- Cardiopatía: Si el paciente cuenta con problemas cardíacos, siendo 0 si y 1 no.

En este proyecto se expondrán los resultados de los modelos aplicados al conjunto de datos para la predicción de enfermedades cardíacas. Para obtener el resultado más cercano se aplicarán aprendizajes no supervisados y supervisados.

El aprendizaje no supervisado es una técnica de aprendizaje automático en la que se utilizan algoritmos para analizar datos sin la necesidad de tener una etiqueta o respuesta previa que se quiera predecir. En este tipo de aprendizaje, el modelo no recibe información específica sobre la salida deseada y es el encargado de encontrar patrones o estructuras en los datos por sí solo.

El aprendizaje supervisado es un subconjunto del machine learning que consiste en la deducción de información a partir de datos de entrenamiento. Estos datos se clasifican en dos secciones: datos de entrenamiento y datos de prueba. Los datos de entrenamiento se utilizan para entrenar a un modelo, y los datos de prueba son los que se usan para determinar la eficacia del modelo creado. El objetivo del aprendizaje supervisado es crear un programa que sea capaz de resolver cualquier variable de entrada luego de ser sometido a un proceso de entrenamiento.

2. Marco Teoríco

El aprendizaje automático es una rama de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos computacionales que permiten

a las máquinas aprender y mejorar su capacidad para realizar tareas específicas a partir de datos.

El aprendizaje automático se basa en la idea de que las máquinas pueden aprender de forma autónoma a través de la experiencia, sin necesidad de ser programadas explícitamente para realizar una tarea. En lugar de eso, se les proporciona un conjunto de datos de entrenamiento y algoritmos que les permiten aprender patrones y relaciones en los datos, y luego aplicar este conocimiento para realizar tareas similares en nuevos datos. Para llevar a cabo el entrenamiento se utilizaron los siguientes modelos:

2.1. Bosque aleatorio

El modelo de bosque aleatorio (Random Forest en inglés) es un algoritmo de aprendizaje automático que se utiliza para la clasificación, la regresión y otras tareas relacionadas con la predicción. Se basa en la idea de combinar múltiples árboles de decisión independientes para construir un modelo más robusto y preciso.

Cada árbol de decisión en un bosque aleatorio se entrena con una submuestra aleatoria de los datos de entrenamiento, y cada nodo en el árbol se divide en función de la característica que proporciona la mayor reducción en la impureza del conjunto de datos. Esto se repite recursivamente hasta que se alcanza un criterio de parada, como la profundidad máxima del árbol o el número mínimo de muestras por hoja.

La salida del modelo se determina por mayoría de votos (en el caso de la clasificación) o por promedio (en el caso de la regresión) de los resultados de todos los árboles de decisión en el bosque.

Matemáticamente, el modelo de bosque aleatorio se puede expresar como:

Para la clasificación:

$$f(x) = \text{argmax}_i \sum_{j=1}^{N} T_j(x) == i \tag{1}$$

donde f(x) es la clase predicha para el vector de características x, i es la clase de interés, N es el número de árboles de decisión en el bosque, y $T_j(x)$ es el resultado del j-ésimo árbol de decisión en el bosque para el vector de características x.

Para la regresión:

$$f(x) = \frac{1}{N} \sum_{j=1}^{N} T_j(x) \tag{2}$$

donde f(x) es el valor predicho para el vector de características x, N es el número de árboles de decisión en el bosque, y T_j(x) es el resultado del j-ésimo árbol de decisión en el bosque para el vector de características x.

En resumen, el modelo de bosque aleatorio es una técnica poderosa y versátil en el aprendizaje automático que utiliza múltiples árboles de decisión independientes para construir un modelo preciso y robusto.

2.2. Árbol de decisión

El modelo de árbol de decisión es un algoritmo de aprendizaje automático que se utiliza para la clasificación y la regresión. Se basa en la idea de dividir el conjunto de datos en subconjuntos más pequeños y homogéneos utilizando una serie de preguntas que se realizan en cada nodo del árbol, hasta que se llega a las hojas del árbol, que representan las clases o valores de salida.

Cada nodo en el árbol de decisión representa una característica de los datos de entrada y cada rama representa una posible respuesta a la pregunta en ese nodo. El objetivo es encontrar la mejor pregunta (característica) que divide los datos de la manera más homogénea posible. Para medir la homogeneidad de los subconjuntos, se utilizan diferentes métricas como la ganancia de información o la impureza de Gini.

Matemáticamente, un árbol de decisión puede representarse como:

Para la clasificación:

$$f(x) = \text{argmax}_i \sum_{j=1}^{M} w_j I(y_j = i) \tag{3}$$

donde f(x) es la clase predicha para el vector de características x, i es la clase de interés, M es el número de hojas en el árbol de decisión, w_j es el peso asociado a la j-ésima hoja y y_j es la clase correspondiente a la j-ésima hoja.

Para la regresión:

$$f(x) = \sum_{j=1}^{M} w_j I(x \in R_j)$$

donde f(x) es el valor predicho para el vector de características x, M es el número de hojas en el árbol de decisión, w_j es el peso asociado a la j-ésima hoja y R_j es el conjunto de datos de entrenamiento que cae en la j-ésima hoja.

2.3. Clasificador bayes ingenuo

El clasificador de Bayes ingenuo es un algoritmo de aprendizaje automático que se basa en el teorema de Bayes para la clasificación de datos. El nombre ïngenuo"se debe a la suposición de que todas las características de los datos de entrada son independientes entre sí, lo que simplifica los cálculos.

El clasificador de Bayes ingenuo asume que todas las características (o variables) del conjunto de datos son independientes entre sí, lo que significa que la presencia o ausencia de una característica no afecta a la presencia o ausencia de otra. El objetivo es calcular la probabilidad de que un ejemplo pertenezca a una

clase determinada, dado que se han observado ciertos valores de características para ese ejemplo.

Matemáticamente, el clasificador de Bayes ingenuo se expresa como:

$$p(C_k|x_1, x_2, ..., x_n) = \frac{p(C_k) \prod_{i=1}^{n} p(x_i|C_k)}{p(x_1, x_2, ..., x_n)} \tag{4}$$

donde $p(C_k|x_1, x_2, ..., x_n)$ es la probabilidad de que el ejemplo pertenezca a la clase C_k dada la observación de los valores de características $x_1, x_2, ..., x_n$. $p(C_k)$ es la probabilidad previa de la clase $C_k, yp(x_i|C_k)$ es la probabilidad condicional de la característica i dado que el ejemplo pertenece a la clase C_k.

La fórmula anterior puede ser reescrita utilizando la regla de Bayes como:

$$p(C_k|x_1, x_2, ..., x_n) = \frac{p(C_k) \prod_{i=1}^{n} p(x_i|C_k)}{\sum_{j=1}^{k} p(C_j) \prod_{i=1}^{n} p(x_i|C_j)} \tag{5}$$

donde la sumatoria en el denominador se extiende a través de todas las clases.

2.4. Regresión logística

La regresión logística es un algoritmo de aprendizaje supervisado utilizado para la clasificación de datos binarios, es decir, para predecir si una observación pertenece a una de dos categorías posibles (por ejemplo, "sí.º "no", "verdadero.º "falso", "1.º "0"). El modelo de regresión logística utiliza una función logística para modelar la probabilidad de que una observación pertenezca a una de las dos categorías.

La función logística se define como:

$$f(z) = \frac{1}{1 + e^{-z}} \tag{6}$$

donde z es una combinación lineal de las características de entrada y los coeficientes del modelo, es decir,

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{7}$$

donde $x_1, x_2, ..., x_n$ son las características de entrada y $\beta_0, \beta_1, \beta_2, ..., \beta_n$ son los coeficientes del modelo.

La función logística mapea los valores de z a un rango de [0,1], lo que se interpreta como la probabilidad de que la observación pertenezca a la categoría "positiva.º "1". La probabilidad de que la observación pertenezca a la categoría "negativa.º "0"se puede calcular como 1 menos la probabilidad de que pertenezca a la categoría "positiva".

El modelo de regresión logística se entrena mediante el ajuste de los coeficientes del modelo para minimizar la función de pérdida logarítmica:

$$L(\beta) = -\frac{1}{m}\sum_{i=1}^{m}[y_i \log(f(z_i)) + (1 - y_i)\log(1 - f(z_i))] \tag{8}$$

donde m es el número de observaciones, y y_i es la etiqueta de clase verdadera (0 o 1) para la i-ésima observación.

3. Metodología

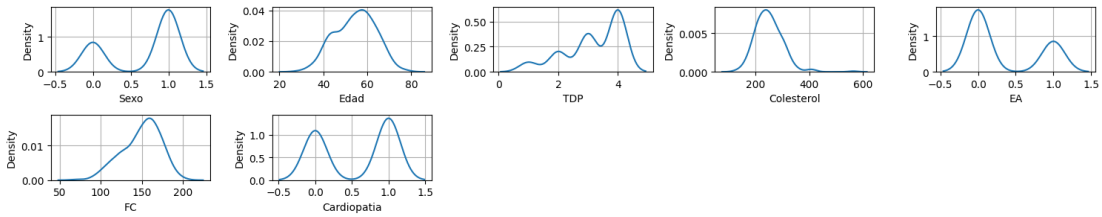
El conjunto de datos se creo por Ïrvine de la Universidad de Californiaz este es un conjunto de datos que se utiliza para predecir enfermedades del corazón. Se clasificó a los pacientes según tuvieran o no enfermedades cardíacas según el cateterismo cardíaco, el estándar de oro. Si tenían más del 50 % de estrechamiento de una arteria coronaria, se los etiquetaba como enfermos del corazón. En esta conjunto hay 270 pacientes y contamos con 6 variables predictivas (Edad, Sexo, TDP, Colesterol, FC, EA) y 1 variable de interés (Cardiopatía).

Cuadro 1: Datos de las primeras filas para detectar Cardiopatía.

Sexo	Edad	TDP	Colesterol	EA	FC	Cardiopatía
0	1	704	322	0	109	0
1	0	673	564	0	160	1
2	1	572	261	0	141	0
3	1	644	263	1	105	1
4	0	742	269	1	121	1

En el Cuadro 1 se muestran los primeros datos con los que se trabajarán, después de esto obtuvimos la densidad de cada variable.

Figura 1: Distribución de las variables.



A simple vista podriamos decir que nuestras variables Edad, Colesterol y FC siguen una distribución normal.

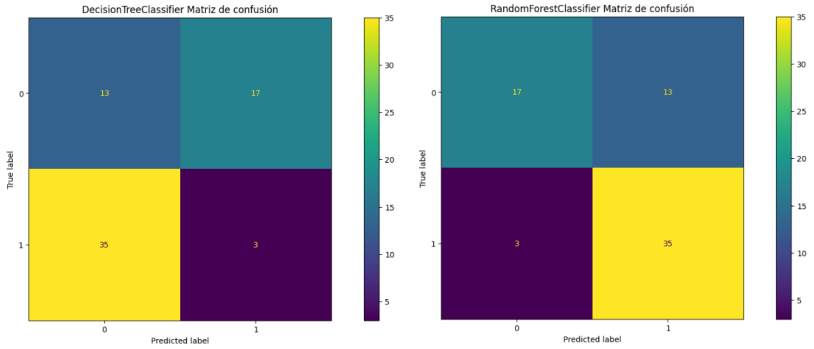
4. Resultados

Para obtener los resultados deseados se forman los modelos a utilizar para llegar a la predicción mas cercana, estás son las antes ya mencionadas que son bosque aleatorio, árbol de decision, clasificador bayes ingenuo y regresión logística. Procedemos al entrenamiento ya haciendo un ajuste en el modelo y continuamos con la prueba de los datos para predecir la cardiopatía en base a las variables predictivas.

Cuadro 2: Precisión de modelos.

Modelo	Puntuación
Bosque aleatorio	100.00
Árbol de decisión	100.00
Clasificador Bayes Ingenuo	78.70
Regresión logística	77.78

Figura 2: Matriz de confusión.

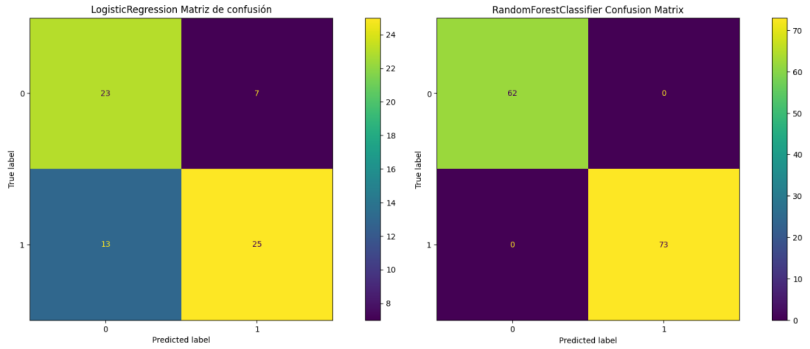


En el Cuadro 2 podemos observar que el modelo de bosque aleatorio y árbol de decisión son el mejor modelo con 100 % de precisión, intentamos replicarlo y se obtuvo una precisión muy alta en ambos modelos.

Decidimos ajustar el modelo en nuestros datos donde la mayoría de los datos serán parte del entrenamiento y el restante sera de prueba, esto para poder partir con datos ya establecidos y asi también aplicaremos el recall para detectar los casos relevantes, esto lo veremos a continuación.

Modelo	Puntuación
Bosque aleatorio	72.92
Árbol de decisión	55.88
Clasificador Bayes Ingenuo	53.12
Regresión logística	75.00

Figura 3: Matriz de confusión.



Con el nuevo ajuste de modelo observamos el Cuadro 3 y nos menciona que el mejor modelo es el de regresíon logística y le sigue el bosque aleatorio, en la Figura 3 nos muestra la matriz de confusión donde hay un total de 48 valores predictivos de forma correcta por el modelo y solo 20 valores donde el modelo se ha equivocado.

5. Conclusiones

En el primer intento obtuvimos un 100 % de precisión de parte de 2 modelos que fueron: bosque aleatorio y árbol de decisión, en los datos esto lo podemos tomar como bueno pero a veces la información es inútil. Por eso mejor decidimos ajustar el modelo para obtener la sensibilidad de los datos y poder ver los casos mas relevantes.

Al ajustar el modelo, nos dio como resultado que las técnicas de regresión logística y el bosque aleatorio eran los modelos con mejor precisión con un 75 % y 72 % respectivamente.

En resumen, con los resultados obtenidos en este proyecto solo nos deja claro que podemos utilizar mas datos o variables predictivas para llegar una predicción mas efectiva realizando distintas pruebas en un futuro, también con estos avances tecnológicos en un plano laboral esto ayuda mucho a las empresas, doctores, etc. a detectar casos de enfermedades cardíacas de una manera mas sencilla y se pueden seguir empleando mejores técnicas o darle mejoras a la que ya se tiene.