Aprendizaje Automático – 63GIIN

Actividad 1 – Portafolio

Gagliardo Miguel Angel

23 de Diciembre de 2024

1. Realizar un análisis exhaustivo del problema presentado, identificando sus características y las necesidades clave para su resolución.

Las enfermedades cardíacas son un término que abarca cualquier trastorno del corazón. Estas, se han convertido en una preocupación importante a la que hay que hacer frente, ya que, si bien los estudios muestran que el número de muertes por enfermedades cardíacas ha disminuido significativamente en las últimas décadas[1]; aún sigue siendo una de las principales causas de muerte en el mundo. Por lo tanto, prevenir las enfermedades cardíacas se ha vuelto más que necesario.

El dataset del archivo **Infarto.csv** es uno de los conjuntos de datos más conocidos[2] y utilizados en problemas de clasificación, especialmente en el campo de la medicina y el análisis de datos, y tiene como objetivo predecir la presencia o ausencia de enfermedad cardíaca en pacientes, basándose en varias características médicas.

A continuación, se describen las columnas y los tipos de datos presentes en este conjunto:

Columna	Descripción	Tipo de Dato
age	Edad del Paciente	Entero
sex	Sexo del Paciente • 0 = femenino • 1 = masculino	Entero (Binario)
cp	Tipo de dolor en el pecho • 1 = angina típica • 2 = angina atípica • 3 = dolor no anginoso • 4 = asintótico	Entero
trtbps	Tensión Arterial en Reposo (mm Hg)	Entero
chol	Nivel de Colesterol (mg/dl)	Entero
fbs	Compara el valor de glucemia en ayunas de un individuo con **120 mg/dl**. Si la glucemia en ayunas es > 120 mg/dl, entonces: • 0 = falso • 1 = verdadero	Entero (Binario)
restecg	ECG en Reposo: • 0 = normal • 1 = anomalía	Entero (Binario)
thalachh	Frecuencia cardíaca máxima alcanzada	Entero
exng	Angina inducida por ejercicio: • 0 = no • 1 = sí	Entero (Binario)
oldpeak	Desnivel ST en el ECG • 1 = ascendente • 2 = plano • 3 = descendente	Float
slp	Slope del segmento ST	Entero
caa	Número de Vasos sanguíneos	Entero

thall	Tipo de Talasemia: • 3 = normal • 6 = defecto fijo • 7 = defecto reversible	Entero
output	Presencia o ausencia de enfermedad cardíaca: • 0 = ausencia • 1 = presente	Entero (Binario)

Observaciones:

- **Variables numéricas:** Como la edad, tensión arterial, colesterol y frecuencia cardíaca.
- **Variables categóricas (rangos):** Como sexo, tipo de dolor en el pecho, nivel de azúcar en sangre, electrocardiograma, entre otras.
- **La variable de salida (output)** es binaria, indicando si el paciente tiene o no.

2. A partir del análisis, identificar y evaluar las distintas técnicas de aprendizaje automático que podrían aplicarse al problema.

Para abordar este problema, las técnicas más relevantes de aprendizaje automático que podrían aplicarse, son:

- **Modelos Lineales (Regresión Logística):** Dado que la variable objetivo es binaria, la regresión logística **es un modelo sencillo y muy utilizado para clasificación binaria**. **Puede funcionar bien** si las relaciones entre las características y la salida son principalmente lineales o si el dataset no contiene demasiadas interacciones complejas.
    - **Ventajas:** Es fácil de interpretar, rápido de entrenar y no requiere mucha preparación de los datos.
    - **Desventajas:** Su rendimiento puede ser limitado si hay relaciones no lineales o interacciones entre características.
- **Árboles de Decisión, Random Forest:** Son adecuados para datos que contienen interacciones **no lineales** entre las características. **Dado que este dataset tiene varias variables categóricas y numéricas**, esta técnica puede servir bien para manejar estas relaciones sin necesidad de transformar excesivamente los datos.
    - **Ventajas:** No requieren que las relaciones sean lineales, **pueden manejar tanto datos categóricos como numéricos**, y son robustos frente a overfitting gracias al ensamblaje de árboles.
    - **Desventajas:** Son más complejos de interpretar que la regresión logística, aunque

ofrecen la ventaja de poder calcular la importancia de las características.

- **Máquinas de Vectores de Soporte (SVM): P**uede ser efectivo para problemas de clasificación binaria, especialmente cuando hay una clara separación entre las clases. **Sin embargo, puede ser menos eficiente en datasets con muchas características categóricas, como en este caso.**
  - **Ventajas:** Son efectivos en espacios de alta dimensión y proporcionan buenos resultados con un adecuado preprocesamiento.
  - **Desventajas:** Puede ser sensible al ajuste de los hiperparámetros, esto es, depende en gran medida de la configuración de ciertos parámetros que deben ser definidos **antes de entrenar el modelo**. Si estos parámetros no se ajustan correctamente, el modelo puede no generalizar bien a los datos, lo que lleva a un mal rendimiento. Por otro lado, no maneja tan bien datos de tipo categórico sin una transformación adecuada.
- **Redes Neuronales:** Pueden ser útiles para problemas más complejos donde las relaciones no lineales son muy pronunciadas. Sin embargo, para un dataset relativamente pequeño como este, las redes neuronales pueden **no ofrecer un rendimiento significativamente mejor que Random Forest o regresión logística**, y su capacidad de interpretación es mucho más limitada.
  - **Ventajas:** Pueden modelar relaciones altamente no lineales y complejas.
  - **Desventajas:** Requieren más datos para generalizar bien, son más lentas de entrenar y menos interpretables.

- **K-Vecinos Más Cercanos (KNN):** Puede ser útil para clasificaciones basadas en la proximidad entre los datos, pero **suele ser computacionalmente costoso cuando los datos son grandes o cuando hay muchas características, como en este caso**.
    - **Ventajas:** Fácil de implementar y no requiere entrenamiento explícito.
    - **Desventajas:** Sensible a las escalas de las características y puede ser muy lento cuando el tamaño del conjunto de datos es grande.

3. Elegir la técnica de aprendizaje más adecuada para resolver el problema propuesto, justificando su elección en base a los aspectos teóricos y prácticos estudiados.

En base a los aspectos teóricos y prácticos estudiados, **mi elección es Random Forest.** Como antes mencionado, esta técnica es capaz de manejar relaciones no lineales entre las características y la variable objetivo, lo cual **es probable en un conjunto de datos de características médicas como este**. Las interacciones entre características como edad, presión arterial, colesterol, y tipo de dolor en el pecho son **probablemente complejas y no lineales**.

El **overfitting** es un problema común en el aprendizaje automático, donde un modelo "aprende demasiado bien" los detalles y el **ruido** de los datos de entrenamiento, hasta el punto de que **pierde su capacidad para generalizar a nuevos datos no vistos.** Hecha esta aclaración, este dataset en particular tiene un número moderado de registros, lo que puede hacer que sea susceptible a overfitting si se usan modelos más complejos, como **redes neuronales**. Random Forest al ser un **modelo de ensamblaje,** o sea, combina modelos varios modelos más simples (como árboles de decisión), **es más robusto y menos susceptible a este problema**.

El dataset incluye variables numéricas, booleanas y categóricas (rangos). **Random Forest es muy adecuado para manejar ambos tipos de datos sin necesidad de transformación**

compleja de las características.

Una de las ventajas más destacadas de **Random Forest es la capacidad de determinar la importancia de las características**. Esto es **valioso en un contexto médico**, donde se desea saber qué factores son los más relevantes para la predicción de la enfermedad cardíaca.

Aunque la **regresión logística es simple y rápida, podría no ser capaz de capturar todas las interacciones y complejidades del conjunto de datos**. Random Forest, por otro lado, es más flexible y tiene el potencial de ofrecer un rendimiento superior al modelar interacciones complejas entre las variables.

A pesar de que Random Forest puede ser computacionalmente más costoso que la regresión logística, su capacidad para manejar tanto datos numéricos como categóricos, junto con su resistencia al overfitting lo hace ideal para este problema.

4. Elaborar un esquema sobre cómo implementar la técnica elegida, incluyendo pasos clave, estructura y lógica del algoritmo.

- **Preprocesamiento de los datos:**
  - **Revisión de datos faltantes:** Si existen valores nulos, decidir cómo tratarlos (eliminación de filas, imputación, etc.).
  - **Codificación de variables categóricas:** Convertir las variables categóricas (**cp, fbs, thall**) en variables numéricas usando técnicas como **one-hot encoding** o **label encoding.**
  - **Escalado de variables numéricas:** Aunque esta técnica no es sensible a la escala, es una buena práctica normalizar o estandarizar las variables numéricas si el modelo se va a combinar con otros algoritmos.
- **División de los datos:** Dividir el conjunto de datos en **conjunto de entrenamiento** (80%) y **conjunto de prueba** (20%), o usar **validación cruzada** si se desea una evaluación más robusta.
- **Entrenamiento:** Crear un modelo Random Forest usando la implementación de bibliotecas como **Scikit-learn**[3].

CONCLUSIONES

Como hemos visto, este dataset contiene datos médicos de pacientes con el objetivo de **predecir la presencia de enfermedad cardíaca**. Las variables incluyen tanto características numéricas, booleanas y categóricas. El análisis de estas variables permite identificar patrones asociados a factores de riesgo para enfermedades cardíacas.

En cuanto a las técnicas de aprendizaje automático, **tanto regresión logística como Random Forest son adecuadas para este tipo de problema**. **Mientras que la regresión logística es simple y eficiente para relaciones lineales, Random Forest es más robusto y maneja interacciones complejas entre variables**.

BIBLIOGRAFIA UTILIZADA

[1] Death rate from cardiovascular diseases, 1980-2021. Our World In Data. https://shorturl.at/1bXNn

[2] Heart Disease dataset. Kaggle. https://archive.ics.uci.edu/dataset/45/heart+disease

[3] RandomForestClassifier**.** Scikit-learn**.** https://shorturl.at/rapw5