Tema: Salud

Problema: Predecir la incidencia de enfermedades cardíacas en función de factores de riesgo.

Dataset: Conjunto de datos de enfermedades cardíacas de la Clínica Cleveland (disponible en Kaggle: https://www.kaggle.com/datasets/aavigan/cleveland-clinic-heart-disease-dataset)

Fuente: UCI Machine Learning Repository

El conjunto de datos de enfermedades cardíacas de la Clínica Cleveland es un conjunto de datos popular utilizado para estudiar y predecir enfermedades cardíacas. El conjunto de datos está disponible en Kaggle (https://www.kaggle.com/ronitf/heart-disease-uci) y proviene originalmente del Repositorio de Aprendizaje Automático de UCI.

Este conjunto de datos contiene información sobre 303 individuos, con 76 atributos para cada persona. Sin embargo, solo 14 de los 76 atributos se utilizan comúnmente en la mayoría de los estudios debido a la dificultad para obtener datos completos para todos los atributos. Los 14 atributos son:

1. Edad: Edad del paciente en años

2. Sexo: Género del paciente (1 = masculino, 0 = femenino)

3. Tipo de dolor en el pecho (cp): Tipo de dolor en el pecho experimentado por el paciente, con valores que van desde 0 hasta 3.

4. Presión arterial en reposo (trestbps): La presión arterial en reposo del paciente (en mm Hg) al ingreso al hospital

5. Colesterol sérico (chol): Nivel de colesterol sérico del paciente (en mg/dl)

6. Azúcar en sangre en ayunas (fbs): Indicador de azúcar en sangre en ayunas > 120 mg/dl (1 = verdadero, 0 = falso)

7. Resultados electrocardiográficos en reposo (restecg): Resultados del electrocardiograma en reposo, con valores que van desde 0 hasta 2.

8. Frecuencia cardíaca máxima alcanzada (thalach): La frecuencia cardíaca máxima del paciente alcanzada durante una prueba de esfuerzo

9. Angina inducida por ejercicio (exang): Si el paciente experimentó angina (dolor en el pecho) durante el ejercicio (1 = sí, 0 = no)

10. Depresión del ST inducida por ejercicio en relación al reposo (oldpeak): Depresión del ST (medida en mm) inducida por ejercicio en relación al reposo

11. Pendiente del segmento ST de ejercicio máximo (slope): La pendiente del segmento ST de ejercicio máximo, con valores que van desde 0 hasta 2.

12. Número de vasos principales (ca): El número de vasos principales coloreados por fluoroscopia (0-3)

13. Talasemia (thal): Un trastorno sanguíneo llamado talasemia (3 = normal, 6 = defecto fijo, 7 = defecto reversible)

14. Objetivo: La presencia de enfermedad cardíaca en el paciente (0 = sin enfermedad, 1 = enfermedad)

El objetivo de utilizar este conjunto de datos en un contexto de aprendizaje automático es predecir la presencia de enfermedades cardíacas (el atributo "objetivo") en función de los otros 13 atributos. Los investigadores y profesionales a menudo emplean diversos algoritmos de aprendizaje automático, como la regresión logística, las máquinas de vectores de soporte y los árboles de decisión, para entrenar modelos y realizar predicciones. El conjunto de datos ha sido ampliamente utilizado para desarrollar y comparar el rendimiento de diferentes algoritmos en la predicción de enfermedades

La salud es un tema fundamental que afecta a todas las personas. La predicción de enfermedades cardíacas es especialmente relevante debido a la alta incidencia de estas enfermedades y su impacto en la salud y calidad de vida de las personas. El conjunto de datos de enfermedades cardíacas de la Clínica Cleveland ha sido una herramienta invaluable para investigadores y profesionales de la salud en el desarrollo de modelos de predicción y comprensión de los factores de riesgo asociados a estas enfermedades.

El conjunto de datos contiene información sobre 303 individuos que han sido evaluados en la Clínica Cleveland. Cada persona se caracteriza por 76 atributos, aunque solo 14 de ellos se utilizan comúnmente debido a la falta de datos completos para los demás atributos. Estos 14 atributos incluyen información relevante y clínicamente significativa que puede utilizarse para predecir la presencia de enfermedades cardíacas en un paciente.

La edad es un factor importante a tener en cuenta, ya que el riesgo de enfermedades cardíacas aumenta con la edad. El género también juega un papel en la predicción, ya que existen diferencias entre hombres y mujeres en términos de factores de riesgo y síntomas. Otro atributo relevante es el tipo de dolor en el pecho experimentado por el paciente, que puede ser indicativo de problemas cardíacos.

La presión arterial en reposo y el nivel de colesterol sérico son medidas clave para evaluar el riesgo cardiovascular. Un nivel alto de colesterol y una presión arterial elevada son factores de riesgo conocidos para enfermedades cardíacas. El azúcar en sangre en ayunas también se considera un factor de riesgo, ya que la diabetes está estrechamente relacionada con enfermedades cardíacas.

Los resultados electrocardiográficos en reposo y la frecuencia cardíaca máxima alcanzada durante una prueba de esfuerzo proporcionan información sobre el estado del corazón y su capacidad para responder al estrés. La presencia de angina durante el ejercicio y la depresión del segmento ST inducida por el ejercicio también son indicadores importantes.

El número de vasos principales coloreados por fluoroscopia es otro atributo que puede revelar información sobre la gravedad de la enfermedad cardíaca. Cuanto mayor sea el número de vasos afectados, mayor será el riesgo de enfermedades cardíacas.

Por último, el atributo "thal" se refiere a la talasemia, un trastorno sanguíneo que puede estar relacionado con enfermedades cardíacas. La presencia de talasemia y su tipo específico pueden influir en la predicción de enfermedades cardíacas.

El objetivo final de utilizar este conjunto de datos es predecir la presencia de enfermedades cardíacas en un paciente en función de los 13 atributos mencionados anteriormente. Se emplean diversos algoritmos de aprendizaje automático, como la regresión logística, las máquinas de vectores de soporte y los árboles de decisión, para entrenar modelos y realizar predicciones precisas.

El conjunto de datos de la Clínica Cleveland ha sido ampliamente utilizado en la comunidad científica y médica para desarrollar y comparar el rendimiento de diferentes algoritmos en la predicción de enfermedades cardíacas. Estas predicciones pueden ayudar a los profesionales de la salud a identificar a

las personas con mayor riesgo y proporcionarles un tratamiento preventivo y cuidados adecuados, lo que puede tener un impacto significativo en la reducción de la morbimortalidad asociada a las enfermedades cardíacas.

En resumen, la predicción de enfermedades cardíacas basada en factores de riesgo es un campo de estudio importante en la salud. El conjunto de datos de enfermedades cardíacas de la Clínica Cleveland proporciona una fuente valiosa de información para desarrollar modelos de predicción y comprender mejor los factores de riesgo asociados con estas enfermedades. El uso de algoritmos de aprendizaje automático puede ayudar a mejorar la precisión de las predicciones y, en última instancia, contribuir a la prevención y tratamiento temprano de las enfermedades cardíacas.  
  
Maria Elena BAIN  
Bien con la selección del dataset y la descripción de variables que realizan. Ahora bien, han pensado que será necesario definir otra variable para un mejor desempeño del modelo? ¿En ese caso cuál?

Hola profesora, buen día, a continuación le dejo lo que reflexionamos:

Podemos definir variables adicionales que mejoren el desempeño del modelo de predicción, dependiendo de la naturaleza y la calidad de estas nuevas variables. Aunque el conjunto de datos de la Clínica Cleveland ya contiene muchos atributos útiles, podríamos considerar agregar algunos más si se dispone de la información.

1. Índice de Masa Corporal (IMC): El IMC es una medida simple de peso en relación con la altura que se utiliza comúnmente para evaluar si una persona tiene un peso saludable. Dado que tanto el peso bajo como el sobrepeso pueden aumentar el riesgo de enfermedades cardíacas, el IMC podría ser un buen predictor de dichas enfermedades.

2. Historial Familiar de Enfermedades Cardíacas: El riesgo de enfermedades cardíacas puede ser mayor en personas con un historial familiar de estas enfermedades. Si es posible, podría ser útil incluir una variable que indique si el paciente tiene parientes cercanos con enfermedades cardíacas.

3. Estilo de vida (Hábito de fumar, consumo de alcohol, nivel de actividad física, etc.): Estos factores también son importantes determinantes de la salud del corazón. El tabaquismo y el consumo excesivo de alcohol pueden aumentar el riesgo de enfermedades cardíacas, mientras que la actividad física regular puede disminuir el riesgo.

4. Síntomas Asociados: Algunos síntomas como la falta de aliento, la fatiga y el hinchazón de las extremidades pueden indicar enfermedades cardíacas.

5. Medicación Actual: Algunos medicamentos pueden afectar la salud del corazón. Sería útil tener una variable que registre cualquier medicamento que esté tomando el paciente.

6. Condiciones de Salud Asociadas: Las enfermedades como la hipertensión, la diabetes y la enfermedad renal pueden aumentar el riesgo de enfermedades cardíacas.

Por supuesto, la adición de estas variables dependería de la disponibilidad de esta información para cada paciente en el conjunto de datos. Además, se deben tener en cuenta los problemas de privacidad y confidencialidad al manejar este tipo de información personal y de salud.

¡Háganos saber si sugiere cambios a esta reflexión que hicimos con José!

Muchas gracias.