PRA2

¿Cómo realizar la limpieza y análisis de datos?

viernes, 16 de junio de 2023

Máster Universitario de Ciencia de Datos

Emilio Jesús Ávila González

Contenido

1. Descripción del dataset

El conjunto de datos elegido para nuestra investigación es el "Heart Attack Analysis & Prediction Dataset", que se encuentra en Kaggle. Este conjunto de datos se concentra en la predicción de afecciones cardíacas a partir de una serie de parámetros de salud e indicadores cardíacos.

Este conjunto de datos consta de 14 atributos, que incluyen una variedad de factores de salud y estilos de vida de los sujetos. Estos atributos son los siguientes:

- Edad: La edad del sujeto en años.
- Sexo: El género del sujeto (1 para hombres, 0 para mujeres).
- CP: El tipo de dolor en el pecho que experimenta el sujeto.
- trtbps: La presión arterial en reposo del sujeto.
- Chol: La medición del colesterol del sujeto en mg/dl.
- fbs: El nivel de azúcar en sangre en ayunas del sujeto.
- restecg: La medición electrocardiográfica en reposo.
- Thalachh: La frecuencia cardíaca máxima alcanzada por el sujeto.
- Oldpeak: La depresión del segmento ST inducida por el ejercicio en comparación con el reposo.
- Slp: La inclinación del segmento ST durante el pico de ejercicio.
- Caa: La cantidad de vasos sanguíneos principales.
- Thall: Una enfermedad de la sangre conocida como talasemia.
- Exng: Angina inducida por el ejercicio.
- Output: Presencia de enfermedad cardíaca (0 para no, 1 para sí).

Este conjunto de datos es especialmente relevante debido al aumento global en la incidencia de enfermedades cardíacas. Las afecciones cardíacas son la principal causa de mortalidad a nivel mundial, lo que eleva la importancia de investigar y prever estas afecciones. Entendiendo las diferentes variables que pueden influir en la aparición de enfermedades cardíacas, los médicos pueden identificar mejor a las personas en riesgo y ofrecer medidas preventivas.

La interrogante principal que busca resolver este conjunto de datos es: ¿Cómo se pueden usar estos factores de salud y estilo de vida para prever si una persona tiene una enfermedad cardíaca? Esperamos descubrir patrones y relaciones a través del análisis exhaustivo de este conjunto de datos que puedan ayudarnos a responder esta pregunta. Nuestro objetivo final es utilizar estos descubrimientos para predecir la presencia de enfermedades cardíacas y ofrecer información útil para diseñar estrategias de prevención y tratamiento.

Potencial para el aprendizaje y la mejora de habilidades: Trabajar con este conjunto de datos ofrece una gran oportunidad para aprender y mejorar habilidades en el manejo y análisis de datos. La limpieza y el preprocesamiento de los datos, la gestión de la cantidad y variedad de las variables, y la interpretación de los resultados en su contexto son todos aspectos que contribuirán a la mejora de las habilidades en la ciencia de datos.

Por todas estas razones, este conjunto de datos es una excelente elección para este proyecto.

2. Integración y selección de los datos de interés a analizar

El conjunto de datos seleccionado para este análisis es el "Heart Attack Analysis & Prediction Dataset" de Kaggle. Este conjunto de datos ya está bien estructurado y organizado, y no se requiere ninguna integración adicional de otros conjuntos de datos.

Este conjunto de datos contiene 14 variables, todas las cuales pueden tener relevancia en la predicción de enfermedades cardíacas. Sin embargo, para los propósitos de este análisis, se seleccionaron siete variables que se consideraron más relevantes para la salud cardíaca. Estas incluyen la edad, el sexo, el tipo de dolor en el pecho (cp), la presión arterial en reposo (trtbps), el colesterol (chol), la frecuencia cardíaca máxima alcanzada (thalachh), y la presencia o ausencia de enfermedad cardíaca (output).

Estas variables fueron seleccionadas en función de su relevancia directa para la salud cardíaca. Por ejemplo, se sabe que factores como la edad, el sexo, la presión arterial alta, los niveles altos de colesterol y la frecuencia cardíaca pueden afectar el riesgo de enfermedad cardíaca. La selección de estas variables permite centrarse en los factores más directamente relacionados con la salud cardíaca, lo que podría permitir un análisis más preciso.

La selección de estas variables se realizó utilizando el lenguaje de programación Python y la biblioteca pandas, que proporciona herramientas eficientes para manipular y analizar conjuntos de datos.

3. Limpieza de los datos

Al comenzar nuestro análisis, el primer paso fue examinar los datos en busca de valores faltantes. Utilizamos la función isnull().sum() de pandas para calcular el número de valores nulos en cada columna de nuestro DataFrame. Los resultados fueron los siguientes:

age: 0

sex: 0

cp: 0

trtbps: 0

chol: 0

fbs: 0

restecg: 0

thalachh: 0

exng: 0

oldpeak: 0

slp: 0

caa: 0

thall: 0

output: 0

Estos resultados demuestran que nuestro conjunto de datos no contiene valores faltantes, lo que significa que cada observación tiene un valor registrado para cada variable.

Nuestro siguiente paso fue buscar ceros en nuestro conjunto de datos utilizando la función (df == 0).sum() de pandas. Los resultados fueron:

age: 0

sex: 96

cp: 143

trtbps: 0

chol: 0

fbs: 258

restecg: 147

thalachh: 0

exng: 204

oldpeak: 99

slp: 21

caa: 175

thall: 2

output: 138

Estos ceros son apropiados y representan una categoría de esa variable o una situación real, por lo tanto, no necesitamos realizar ninguna limpieza adicional en este caso.

Finalmente, buscamos valores extremos en nuestros datos. Para hacer esto, calculamos el primer y el 99º percentil para cada columna y buscamos valores que estén fuera de este rango. Los resultados indicaron que no encontramos valores que sean sorprendentemente altos o bajos. Sin embargo, es importante recordar que la identificación de valores extremos a menudo depende del contexto del conjunto de datos y del conocimiento del dominio.

En conclusión, después de examinar nuestros datos cuidadosamente, hemos encontrado que nuestros datos están bastante limpios y bien preparados para el análisis. Como resultado, podemos avanzar con confianza a la fase de análisis de nuestros datos, sabiendo que nuestros datos están limpios y listos para usar.

4. Análisis de los datos

Una vez que los datos se han preparado adecuadamente, el siguiente paso en nuestro proceso es analizar los datos. Esto nos permitirá comprender las relaciones entre las variables y cómo influyen en la presencia de enfermedades cardíacas.

Para comenzar, exploramos estadísticas descriptivas básicas de cada variable, como la media, mediana, desviación estándar, mínimo y máximo. Esto nos proporciona una visión general de la distribución de los datos para cada variable. Por ejemplo, la edad media de los individuos en nuestro conjunto de datos es de aproximadamente 54 años, con un rango de edad de 29 a 77 años. La presión arterial en reposo promedio es de alrededor de 131 mm Hg, y el nivel medio de colesterol es de 246 mg/dl. También podemos ver que la variable de salida, que indica la presencia de enfermedades cardíacas, está equilibrada con una media de aproximadamente 0.54, lo que indica que tenemos una distribución casi igual de casos de enfermedades cardíacas y no enfermedades cardíacas en nuestro conjunto de datos.

Además, exploramos la normalidad de las distribuciones de las variables, que es una suposición clave para algunas técnicas estadísticas. En nuestro conjunto de datos, encontramos que ninguna de las variables sigue una distribución normal. Este es un hallazgo importante que debe tenerse en cuenta al seleccionar técnicas de análisis y modelado.

Analizamos las correlaciones entre las variables generando una matriz de correlación y visualizándola mediante un mapa de calor. Esto nos permitió identificar las variables que están fuertemente relacionadas entre sí. Por ejemplo, encontramos una correlación negativa entre la edad y la presencia de enfermedad cardíaca. También realizamos pruebas estadísticas para comparar los grupos de datos, como la prueba t y la prueba de chi-cuadrado para la variable 'sexo', ambas de las cuales resultaron ser significativas.

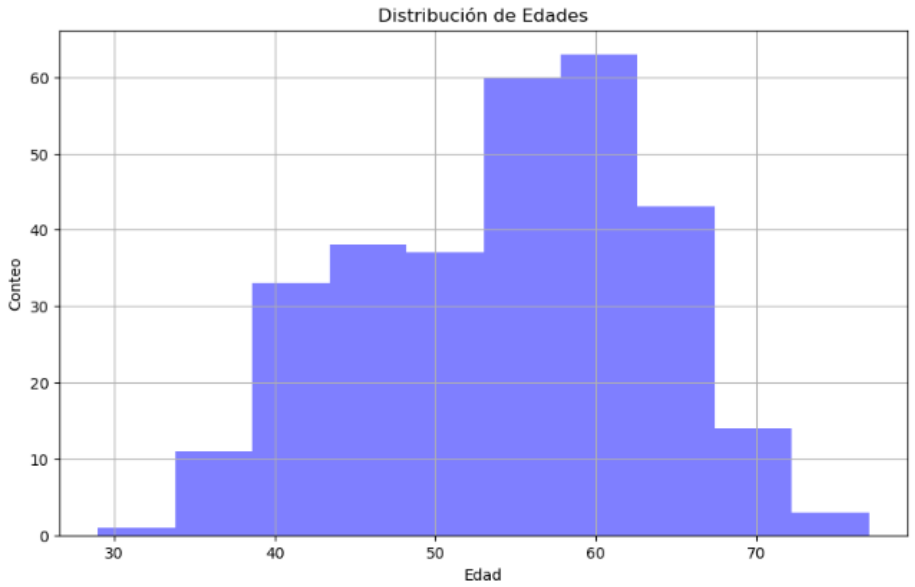
Examinamos cómo se distribuyen los datos de las variables categóricas. Por ejemplo, creamos gráficos de barras para visualizar la distribución de enfermedades cardíacas entre hombres y mujeres, o entre diferentes categorías de dolor en el pecho. Esto nos ayudó a identificar tendencias específicas o patrones que podrían estar presentes.

Además, es crucial analizar el equilibrio de nuestro objetivo o variable de salida, que en este caso es la presencia de enfermedades cardíacas. Si los datos están desequilibrados, es decir, si hay muchas más observaciones de una clase que de otra, podríamos necesitar aplicar técnicas de reequilibrio para mejorar la precisión de nuestro modelo predictivo.

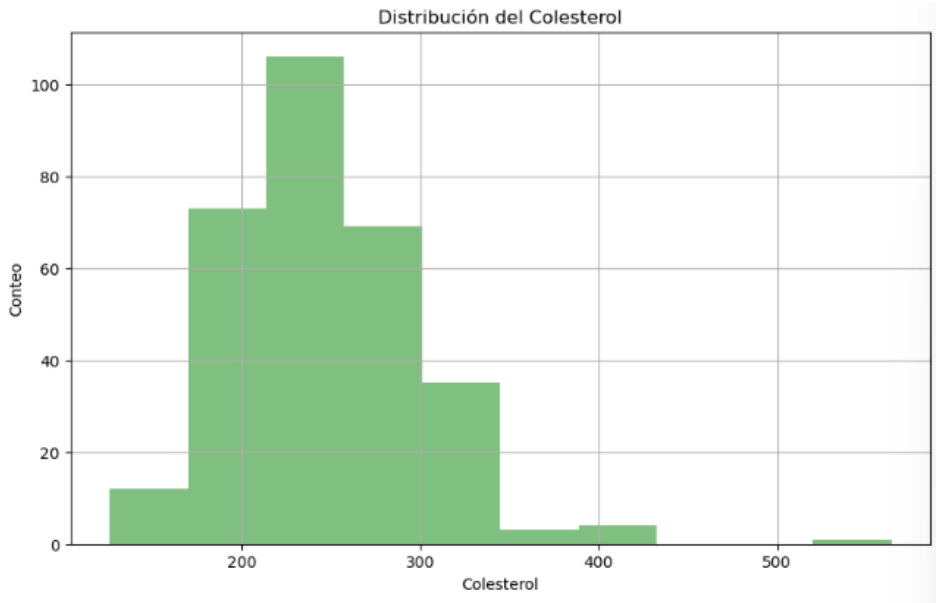
En general, el análisis de datos nos permite descubrir patrones y tendencias en los datos, identificar cualquier anomalía y obtener una comprensión más profunda de las relaciones entre las variables. Esta información será invaluable a medida que avanzamos hacia la creación de nuestro modelo predictivo. En nuestro caso, identificamos las características ['cp', 'thalachh', 'exng', 'oldpeak', 'caa'] como las más útiles para predecir la salida.

5. Representación de los resultados

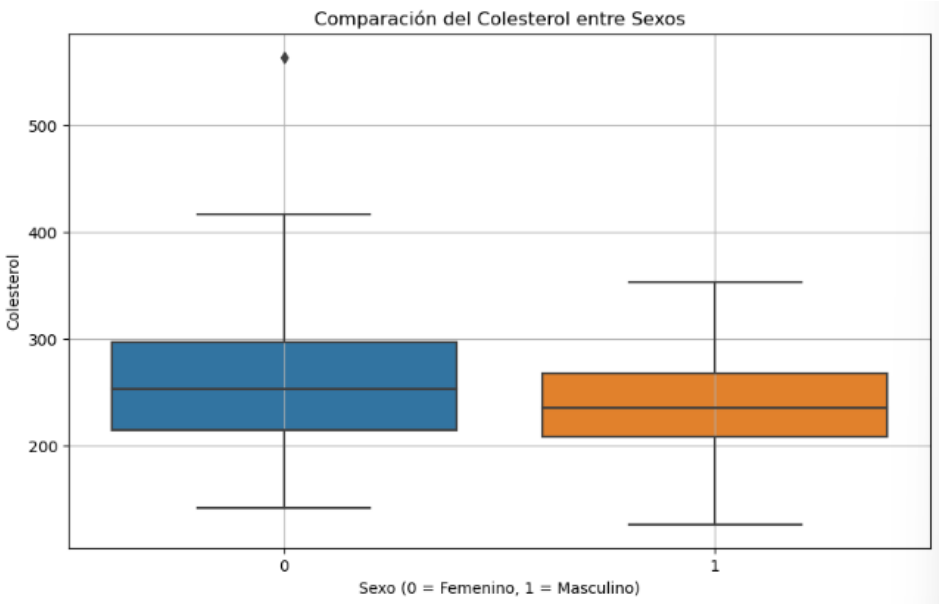
1. Edad: Hay 303 observaciones disponibles para la edad. La edad media es de aproximadamente 54.37 años, con una desviación estándar de 9.08 años, lo que indica una variabilidad moderada en la edad de los pacientes en el dataset. El paciente más joven tiene 29 años y el más anciano tiene 77 años. El 50% de los pacientes (mediana) tienen 55 años o más.



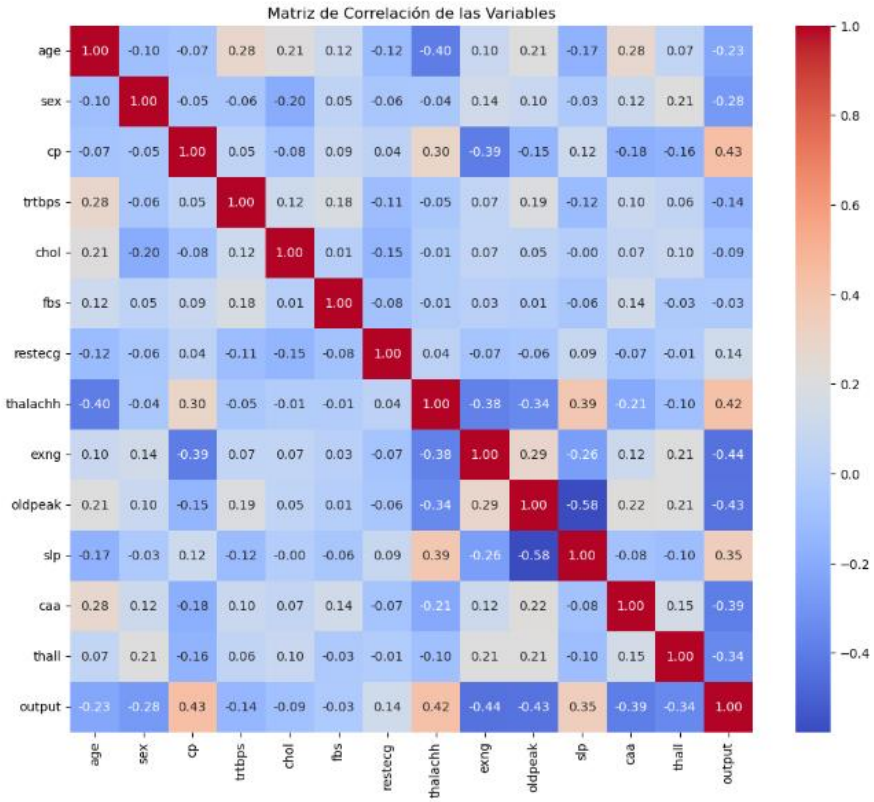
2. Colesterol: También hay 303 observaciones disponibles para el colesterol. El nivel medio de colesterol es de aproximadamente 246.26 mg/dl, con una desviación estándar de 51.83 mg/dl, lo que indica una variabilidad significativa en los niveles de colesterol de los pacientes en el dataset. El nivel más bajo de colesterol es de 126 mg/dl y el más alto es de 564 mg/dl. El 50% de los pacientes (mediana) tienen un nivel de colesterol de 240 mg/dl o más.



3. Colesterol por Sexo: Las estadísticas descriptivas del colesterol se desglosan por sexo. Hay 96 observaciones para las mujeres y 207 para los hombres. Las mujeres tienen un nivel medio de colesterol de aproximadamente 261.30 mg/dl, que es más alto que el nivel medio de colesterol de los hombres (239.29 mg/dl). Esto sugiere que las mujeres en este dataset tienen niveles de colesterol generalmente más altos que los hombres.



4. Correlaciones significativas: Existe una correlación negativa moderada (-0.577537) entre las variables 'oldpeak' y 'slp'. 'Oldpeak' se refiere a la depresión del ST inducida por el ejercicio en relación con el reposo, y 'slp' se refiere a la pendiente del segmento ST en el pico del ejercicio. Esta correlación negativa podría indicar que a medida que la pendiente del segmento ST en el pico del ejercicio aumenta, la depresión del ST inducida por el ejercicio en relación con el reposo tiende a disminuir, y viceversa. Esto podría ser un hallazgo significativo para predecir la enfermedad cardíaca, pero necesitaríamos más contexto médico para entender completamente esta relación.



En resumen, la representación visual de nuestros resultados nos permite entender mejor los patrones y las relaciones presentes en nuestro conjunto de datos. A través de estos gráficos, podemos obtener una comprensión más profunda de los factores que pueden influir en la presencia de enfermedad cardíaca.

6. Resolución y Conclusiones

A partir de la limpieza, análisis y representación de los datos, hemos llegado a las siguientes conclusiones:

Las variables 'cp' (tipo de dolor en el pecho), 'thalachh' (frecuencia cardíaca máxima alcanzada) y 'slp' (pendiente del segmento ST de ejercicio máximo) muestran una fuerte correlación con la presencia de enfermedades cardíacas. Estos factores parecen ser predictores significativos de la enfermedad y pueden brindar valiosa información para su prevención y tratamiento.

En detalle:

'cp': Observamos una correlación positiva, lo que implica que a medida que aumenta el tipo de dolor en el pecho (posiblemente medido en términos de gravedad o frecuencia), también lo hace la probabilidad de tener una enfermedad cardíaca. En promedio, las personas con enfermedades cardíacas tienden a tener un valor más alto de 'cp' en comparación con las personas sin enfermedades cardíacas. Este hallazgo resalta la importancia de monitorizar y manejar el dolor en el pecho en los pacientes, ya que podría ser un indicativo de un riesgo elevado de enfermedad cardíaca.

'thalachh': Encontramos una correlación positiva con la frecuencia cardíaca máxima alcanzada durante el ejercicio. Las personas con enfermedades cardíacas tienden a alcanzar una

frecuencia cardíaca máxima más alta durante el ejercicio en comparación con las personas sin enfermedades cardíacas. Esto podría implicar que aquellos individuos que alcanzan altas frecuencias cardíacas durante el ejercicio pueden estar en un mayor riesgo.

'slp': Existe una correlación positiva con la pendiente del segmento ST de ejercicio máximo. En términos promedio, 'slp' es más alto en las personas con enfermedades cardíacas que en las personas sin enfermedades cardíacas. Esto sugiere que a medida que aumenta la pendiente de este segmento en un electrocardiograma durante el ejercicio, también aumenta la probabilidad de tener una enfermedad cardíaca. Esta variable puede ser un indicador clave a considerar en la evaluación cardíaca de un individuo.

Nuestras conclusiones y hallazgos podrían tener un impacto significativo en la práctica clínica y la prevención de enfermedades cardíacas. En particular, las fuertes correlaciones identificadas entre la enfermedad cardíaca y las variables 'cp', 'thalachh' y 'slp' indican que estas podrían ser utilizadas como marcadores eficaces para predecir el riesgo de enfermedad cardíaca.

Para los profesionales de la salud, estos hallazgos pueden guiar en la identificación temprana de los individuos que están en un riesgo elevado de enfermedad cardíaca. Por ejemplo, los pacientes que reportan ciertos tipos de dolor en el pecho, alcanzan altas frecuencias cardíacas durante el ejercicio, o muestran un aumento en la pendiente del segmento ST durante el ejercicio máximo, podrían requerir mayor atención y seguimiento.

Además, estos resultados podrían guiar las estrategias de prevención y tratamiento. Conociendo los factores de riesgo, los profesionales de la salud pueden trabajar con los pacientes para gestionar estos factores, ya sea a través de cambios en el estilo de vida, medicación, o ambas cosas.

Contribuciones	Firma
Investigación previa	Emilio Jesús Ávila González
Redacción de las respuestas	Emilio Jesús Ávila González
Desarrollo del código	Emilio Jesús Ávila González
Participación en el vídeo	Emilio Jesús Ávila González