Contenido

[Introducción 1](#_Toc36926599)

[Descripción de Datos 1](#_Toc36926600)

[Problema 2](#_Toc36926601)

[Pregunta para investigar 2](#_Toc36926602)

[Conclusiones 3](#_Toc36926603)

[Resumen 8](#_Toc36926604)

[Anexos 9](#_Toc36926605)

Proyecto Data Science Capstone With Python, completo en GitHub:

<https://github.com/rrical/Task-4-Complete-a-Data-Science-Capstone-Project>

# Introducción

* El conjunto de datos proporciona una serie de variables junto con una condición objetivo de tener o no tener una enfermedad cardíaca.
* Los datos se utilizan para crear un modelo que intente predecir si un paciente tiene esta enfermedad o no.
* El campo "**target**" se refiere a la presencia de enfermedad cardíaca en el paciente.

# Descripción de Datos

El conjunto de datos contiene: 1025 observaciones (filas) y 14 variables (columnas).

Las columnas definidas en el conjunto de datos, el tipo de datos y una breve explicación de su contenido, se describen a continuación:

* **age**: edad de la persona en años
  + tipo de datos: int64
* **sex**: sexo de la persona [1: hombre, 0: mujer]
  + tipo de datos: int64
* **cp**: el dolor experimentado en el pecho [0: angina típica, 1: angina atípica, 2: dolor no anginal, 3: asintomático]
  + tipo de datos: int64
  + angina típica: dolor en el pecho causado cuando el músculo cardíaco no recibe suficiente sangre rica en oxígeno.
  + angina atípica: las mujeres tienen una presentación más sutil llamada angina atípica.
  + asintomático: sin causar ni exhibir síntomas de enfermedad.
* **trestbps**: presión arterial en reposo de la persona en mm Hg al ingresar al hospital
  + tipo de datos: int64
  + condición normal: 120 – 140 mm
* **chol**: medición de colesterol en la persona en mg/dl
  + tipo de datos: int64
  + condición normal: 100 – 129 mg/dl
* **fbs**: azúcar en sangre en ayunas de la persona > 120 mg/dl [1: verdadero, 0: falso]
  + tipo de datos: int64
  + condición normal: menor a 100 mg/dl
* **restecg**: medición electrocardiográfica en reposo [0: normal, 1: con anormalidad de la onda ST-T, 2: mostrando hipertrofia ventricular izquierda probable o definitiva según los criterios de estrés]
  + tipo de datos: int64
* **thalach**: frecuencia cardíaca máxima alcanzada por la persona, por minuto
  + tipo de datos: int64
  + condición normal: de 60 a 100 latido por minuto
* **exang**: angina inducida por el ejercicio [1: sí; 0: no]
  + tipo de datos: int64
* **oldpeak**: depresión del ST inducida por el ejercicio en relación con el descanso ('ST' se refiere a las posiciones en la gráfica del ECG - Electrocardiograma)
  + tipo de datos: float64
* **slope**: pendiente del segmento ST en picos de ejercicio [0: subida, 1: plano, 2: bajada]
  + tipo de datos: int64
* **ca**: número de vasos mayores [0-3]
  + tipo de datos: int64
* **thal**: trastorno sanguíneo llamado talasemia [1: normal, 2: defecto fijo, 3: defecto reversible]
  + tipo de datos: int64
* **target**: presenta enfermedad cardíaca [0: no, 1: sí]
  + tipo de datos: int64

# Problema

El conjunto de datos contiene muchos indicadores médicos, el objetivo es hacer un análisis exploratorio de datos sobre el estado de la enfermedad cardíaca. Donde el conjunto de datos contiene el historial médico de pacientes utilizado en la predicción de enfermedades cardíacas.

# Pregunta para investigar

En este estudio, el objetivo es predecir si una persona tiene una enfermedad cardíaca o no, en función de los atributos de presión arterial, latidos cardíacos, angina inducida por el ejercicio, azúcar en sangre en ayunas, entre otros.

# Conclusiones

* No hay datos perdidos, ni valores nulos; en el conjunto de datos.
* Existen algunos valores atípicos (outliers) en los atributos Presión Arterial [PA] y Colesterol [Colest], mediante el uso de zScore se eliminan estos valores, cuando están fuera de la desviación estándar 3.
* El 94.5% de los datos están dentro de la desviación estándar de la media.
* Correlación positiva:
  + **Entre Edad y PA**: a medida que aumenta la edad de una persona, también aumenta el valor de la presión arterial.
  + **Entre Edad y DepST**: a medida que aumenta la edad de una persona, aumenta el valor de depresión del ST inducida por el ejercicio en relación con el descanso.
  + **Entre Colest y PA**: el colesterol y la presión arterial muestran una correlación positiva, lo que tiene sentido ya que el colesterol alto está asociado con un riesgo elevado de enfermedad cardiovascular.
* Correlación negativa
  + **Edad vs FCM**: la edad y la frecuencia cardíaca máxima, tiene sentido porque a medida que aumenta la edad de una persona, disminuye el valor de la frecuencia cardíaca.
  + **FCM vs DepST**: frecuencia cardíaca máxima y la depresión del ST inducida por el ejercicio en relación con el descanso, lo que tiene sentido a medida que aumenta la frecuencia cardíaca de una persona, el valor de depresión del ST disminuye.
* Variables continuas
  + Edad
    - El valor de asimetría y la prueba normal, muestra con claridad que la distribución no es normal. La mediana de edad de los pacientes fue de 54 años, siendo los más jóvenes y mayores de 29 y 77, respectivamente. Los pacientes enfermos tienen una mediana más alta de 58 años en comparación con los pacientes no enfermos que tienen una mediana de 52 años. Para mostrar la diferencia entre los grupos (ausencia y presencia), utilizo el teorema del límite central, para reducir la variabilidad alrededor de las medias de dos grupos.
    - Suposición: la media del grupo no enfermo es menor que la media del grupo enfermo. Hipótesis nula: no hay diferencia en la media del grupo enfermo y no enfermo. Hipótesis alternativa: existe una diferencia en la media del grupo enfermo y no enfermo.
    - Dado el valor p > 0.05. Acepto la hipótesis alternativa de que no hay diferencia en la media del grupo enfermo y no enfermo. Por lo tanto, concluyo que las personas que son un poco mayores tienen más posibilidades de tener una enfermedad cardíaca. Por lo tanto, la edad sería una característica predictiva.
  + Presión sanguínea en reposo
    - Al igual que en el caso anterior, esta distribución tampoco es normal. Para mostrar la diferencia entre los grupos (ausencia y presencia), utilizo el teorema del límite central.
    - Dado el valor p < 0.05. Rechacé la hipótesis nula de que no hay diferencia en la media del grupo enfermo y el grupo no enfermo. La presión arterial en reposo agregada para todo el conjunto de datos exhibió un valor medio de 131 y para los grupos enfermos y no enfermos (es decir, 129 y 133 respectivamente). Por lo tanto, concluyo que las personas que tienen presión arterial ligeramente alta tienen más posibilidades de tener un corazón enfermo. Por lo tanto, la presión arterial en reposo es una buena característica predictiva.
  + Colesterol
    - Como en el caso anterior, esta distribución tampoco es normal. Para mostrar la diferencia entre los grupos (ausencia y presencia), utilizo el teorema del límite central.
    - Dado el valor p < 0.05. Rechace la hipótesis nula de que no hay diferencia en la media del grupo enfermo y la del grupo no enfermo. Los niveles de colesterol en general (mediana = 244 mg/dL) fueron más bajos en comparación con los pacientes no enfermos (mediana = 249 mg/dL). Por lo tanto, colesterol no puede ser una buena característica predictiva.
  + Depresión ST inducida
    - Dado el valor p < 0.05. Rechace la hipótesis nula de que no hay diferencia en la media del grupo enfermo y la del grupo no enfermo. La depresión ST inducida por el ejercicio difiere entre los grupos de enfermos y no enfermos, y la mayoría de los pacientes con enfermedad cardíaca mostraron una media y mediana más alta para el grupo de enfermos. Por lo tanto, la depresión ST inducida por el ejercicio en relación con el descanso puede ser una buena característica predictiva.
  + Frecuencia cardíaca máxima
    - Dado el valor p < 0.05. Rechace la hipótesis nula de que no hay diferencia en la media del grupo enfermo y no enfermo. La frecuencia cardíaca máxima fue mayor para el grupo de enfermos (media = 159, mediana = 162) en comparación con pacientes no enfermos (media = 139, mediana = 141). Se anticipó que esta característica debería tener un alto poder predictivo.
* Variables categóricas
  + Género
    - Dado el bajo valor p (6.3062518551189255e-21), se rechaza la hipótesis nula y el resultado de la prueba detecta una relación significativa entre Género y Objetivo.
  + Dolor de pecho
    - Dado el bajo valor p (8.291170809908133e-53), se rechaza la hipótesis nula y el resultado de la prueba detecta una relación significativa entre CP y Objetivo. En el gráfico de presencia de enfermedad con dolor en el pecho se denota una alta incidencia en comparación con otros dolores en el pecho, por lo que el dolor asintomático puede tener un alto poder predictivo.
  + Glucemia en ayunas
    - Dado el valor de p, por lo que no podemos rechazar la hipótesis nula y el resultado de la prueba detecta una relación no significativa entre Fbs y Objetivo. La mayoría de las personas no tenían niveles de azúcar en sangre en ayunas superiores a 120 mg / dL. Esto no cambió mucho cuando los datos se dividieron en función de la presencia de enfermedad, por lo que FBS no es una característica predictiva.
  + Resultados de ECG en reposo
    - La mayoría de los pacientes exhibieron resultados electrocardiográficos normales en reposo. Sin embargo, una mayor proporción de pacientes enfermos tenía hipertrofia, lo que sugiere que esta característica puede contribuir con cierto poder predictivo.
  + Angina inducida por ejercicio
    - Dado el valor p, rechazamos la hipótesis nula y el resultado de la prueba detecta una relación significativa entre Exang y Objetivo. Significativamente más pacientes en el grupo enfermo mostraron angina inducida por el ejercicio. Esta característica debería ser muy predictiva.
  + Ejercicio pico segmento ST
    - Dado el bajo valor de p, se rechaza la hipótesis nula y el resultado de la prueba detecta una relación significativa entre Pendiente y Objetivo. Significativamente más pacientes en el grupo no enfermos mostraron PenSegST-Plana. Esta característica podría ser muy predictiva. La pendiente del segmento ST de ejercicio máximo difiere entre el grupo de no enfermos y enfermos, y la mayoría de los pacientes con enfermedad cardíaca exhibieron una pendiente ST plana (valor = 2). Esto también puede tener un buen poder predictivo.
  + Cantidad de vasos sanguíneos
    - Dado el bajo valor p, se rechaza la hipótesis nula y el resultado de la prueba detecta una relación significativa entre CA y Objetivo. Significativamente más pacientes tiene un número de vasos sanguíneos mayor que 1. Esta característica debería ser muy predictiva.
  + Talasemia
    - Dado el bajo valor de p, se rechaza la hipótesis nula y el resultado de la prueba detecta una relación significativa entre Thal y Objetivo. Significativamente más pacientes en el grupo enfermos tiene defecto reversible. Esta característica debería ser muy predictiva.
  + Género, tipo de dolor en el pecho y ECG en reposo
    - La mayoría de las personas diagnosticadas con enfermedad cardíaca eran hombres con signos típicos de dolor de angina asintomática (valor = 4) y que muestran hipertrofia ventricular izquierda probable o definitiva (restECG = 2)
  + Género, tipo de dolor en el pecho y angina inducida por ejercicio
    - La mayoría de las personas diagnosticadas con enfermedad cardíaca eran hombres con signos típicos de dolor de angina asintomática (valor = 4) y con angina inducida por el ejercicio (AngIndEje = 1).

La exploración de los datos indicó que DepST, FCM, DP (dolor asintomático), CA (> 1), talasemia (defecto reversible) son posibles características útiles para predecir la presencia de enfermedad cardíaca. La edad, AngIndEje, PenSegST, PA, Colest, Genero, AzuSan y ResulECG también tenían un poder predictivo potencialmente menor.

Fuertes atributos de poder predictivo: DepST, FCM, DP (dolor asintomático), CA (> 1), talasemia (defecto reversible).

* Los pacientes que tienen un valor de pico antiguo mayor que 1, tienen más probabilidad de tener una enfermedad cardíaca que los pacientes con un valor de pico antiguo menor que 1.
* Los pacientes con enfermedad cardíaca tienen una frecuencia cardíaca más baja, menos de 140; en comparación con el paciente que no tiene enfermedad cardíaca, más de 140.
* Los pacientes que padecen enfermedades cardíacas tienen más probabilidades de tener dolor torácico asintomático que los pacientes que no tienen enfermedades cardíacas.
* El grupo de enfermos tiene más posibilidades (aproximadamente, seis veces) de tener un defecto reversible de talasemia que el grupo no enfermo.

Atributos de poder predictivo moderado: Edad, AngIndEje, PenSegST.

* Los pacientes mayores tienen más posibilidades de tener una enfermedad cardíaca.
* El grupo de enfermos tiene más posibilidades de tener dolor en el pecho después del ejercicio que el grupo de no enfermos.
* El grupo de enfermos tiene más posibilidades de tener una pendiente de onda plana que el grupo de no enfermos.

Atributos de poder predictivo débiles: - PA, Colest, Genero, AzuSan, ResulECG.

* Estos atributos no mostraron ningún poder predictivo o no pueden distinguir entre la cohorte con enfermedad y sin enfermedad en base a estos atributos.

Aplicando modelos predictivos.

* Crear variables dummy: antes de aplicar cualquier modelo, cree variables dummy de atributos categóricos. Que permitan utilizar una sola ecuación de regresión para representar múltiples grupos. Convirtiendo la variable categórica en nuevos atributos k-1, con k = niveles del atributo categórico.
* Escalando las variables numéricas: lo que me permitió estandarizar el rango de variables independientes o características de datos; utilizando el método min-max, y así re escalar el rango de características al rango en [0,1].
* División de conjunto de datos en Train y conjunto de Test: para implementar el modelo de algoritmo, separé las variables dependientes e independientes del conjunto de datos, luego dividí el conjunto de datos en entrenamiento y prueba para evaluar modelos.

Regresión logística: mediante la aplicación de la regresión logística, trataré este problema de clasificación.

* Modelo de entrenamiento: dada la regularización C = 0.4 y l2 son los mejores hiperparámetros para mi modelo.
  + Ajuste de hiperparámetros usando GridSearchCV, que construirá y evaluará metódicamente el modelo para cada combinación de parámetros de algoritmo especificados. Dada la regularización C = 0.3 y l1 son los mejores hiperparámetros para mi modelo. Todos los modelos tendrán el valor de estos hiperparámetros.
* Entrenando el modelo con todos los atributos
  + Sensibilidad / recuerdo: qué tan buena es una prueba para detectar los positivos. Una prueba puede hacer trampa y maximizar esto al devolver siempre "positivo".
  + Especificidad: qué tan buena es una prueba para evitar falsas alarmas. Una prueba puede hacer trampa y maximizar esto al devolver siempre “negativo”.
  + Precisión: cuántos de los clasificados positivamente eran relevantes. Una prueba puede engañar y maximizar esto al solo dar positivo en un resultado en el que tiene más confianza.
  + Puntaje F1: un buen puntaje F1 significa que tiene bajos falsos positivos y bajos falsos negativos, por lo que está identificando correctamente las amenazas reales y no le molestan las falsas alarmas.
  + Modelo de entrenamiento con los atributos de poder más predictivos concluidos en la parte EDA: como se concluyó en el análisis exploratorio de datos, estas cuatro columnas ‘DepST’, ‘FCM’, ‘CA’, ‘Talasemia con defecto reversible’, ‘DP dolor asintomático’ mostraron un fuerte poder predictivo. Al entrenar el modelo en estas 5 columnas, la precisión del entrenamiento es del 84% y la precisión de la prueba es del 87%.
* Validación cruzada K-FOLD: técnica para evaluar el rendimiento de los modelos de aprendizaje automático. K-Fold CV divide el conjunto de datos dado en K de secciones / pliegues, donde cada pliegue será un conjunto de prueba en algún momento.
  + El 84% de precisión promedio de todos los pliegues demuestra que no hay sobreajuste.
  + A partir de la tabla de importancia de características, eliminaré las columnas que tienen menos importancia. El motivo principal de aplicar un modelo es obtener una buena precisión con menos características. Por lo tanto, la CA es la característica más importante, es decir, el número de vasos principales, lo que tiene sentido, ya que lo que sucede cuando el suministro de sangre de su corazón se bloquea o se interrumpe por una acumulación de sustancias grasas en las arterias coronarias ", parece lógico que los vasos más importantes son algo bueno, y por lo tanto reducirá la probabilidad de enfermedad cardíaca.
* Aplicar modelo después de eliminar el atributo menos significativo, para obtener así más probabilidades en el modelo de sobreajuste.
  + Entonces, incluso después de eliminar las columnas menos significativas, la precisión no cambia. La mejor precisión obtenida es el 88% de TEST y el 85% de precisión del TRAIN.
  + El valor de AUC es 0,88, lo que significa que nuestro modelo puede distinguir entre pacientes con enfermedad y sin enfermedad con una probabilidad de 0,88; siendo un buen valor. Nuestro clasificador debe ser sensible a los falsos negativos. En ese caso, la persona realizará más pruebas y concluirá que no tiene ningún problema cardíaco, por lo que mi modelo debe tener una alta sensibilidad. El corte óptimo sería donde tpr es alto y fpr es bajo
* Aplicar la reducción del error tipo II
* Ahora aplicar la regresión logística, pero con PCA

Comparando todo los modelos logísticos

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Modelo | Accuracy | Senstivity | Specificity | Precision | F1score |
| Logistic(todas las variables) | 87% | 0.85 | 0.80 | 0.83 | 0.84 |
| Logistic(5 variables) | 88% | 0.84 | 0.89 | 0.90 | 0.87 |
| Logistic(menos sign) | 88% | 0.96 | 0.77 | 0.83 | 0.89 |
| Logistic(PCA) | 76% | 0.84 | 0.89 | 0.90 | 0.87 |

Random Forest: colección de árboles de decisión cuyos resultados se agregan en un resultado final. Su capacidad para limitar el sobreajuste, sin aumentar sustancialmente; y el error debido al sesgo, es la razón por la cual son modelos altamente poderosos. Los bosques aleatorios consisten en múltiples árboles individuales, cada uno basado en una muestra aleatoria de los datos de capacitación. Por lo general, son más precisos que los árboles de decisión única.

* Ajuste de hiperparámetros usando GridSearch
* Importancia de la característica

Comparando todo los modelos logísticos

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Modelo | Accuracy | Senstivity | Specificity | Precision | F1score |
| Random Forest(todas las variables) | 97% | 0.98 | 0.96 | 0.96 | 0.97 |

* Aplicar modelo después de eliminar atributos menos significativos

Entonces, eliminar columnas menos significativas en un bosque aleatorio reduce la precisión, solo disminuye el tiempo de entrenamiento.

Árbol de decisión

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Modelo | Accuracy | Senstivity | Specificity | Precision | F1score |
| Logistic(todas las variables) | 87% | 0.85 | 0.80 | 0.83 | 0.84 |
| Logistic(5 variables) | 88% | 0.84 | 0.89 | 0.90 | 0.87 |
| Logistic(menos sign) | 88% | 0.96 | 0.77 | 0.83 | 0.89 |
| Logistic(PCA) | 76% | 0.84 | 0.89 | 0.90 | 0.87 |
| Random Forest(todas las variables) | 97% | 0.98 | 0.96 | 0.96 | 0.97 |

# Resumen

El objetivo era predecir si una persona tiene una enfermedad cardíaca o no en función de los atributos de presión arterial, latidos cardíacos, angina inducida por ejercicio, azúcar en la sangre y otros.

El modelo de Random Forest con todas las variables se desempeñó mejor con una precisión del 97%, sensibilidad del 98%, especificidad del 96%, precisión del 96% y puntaje de f1 97%.

Características importantes: DesST, FCM, DP asintomático, CA (> 1), talasemia (defecto reversible) estas características mostraban un alto poder predictivo después de la edad.

DesST, CA, AngIndEje\_Yes, FCM, Talasemia\_Reversible son las características importantes para la clasificación después de aplicar los modelos.

Entonces, el número de vasos sanguíneos mayores aumenta, la probabilidad de enfermedad cardíaca disminuye. Eso tiene sentido, ya que significa que puede llegar más sangre al corazón. A medida que aumenta la frecuencia cardíaca, también aumenta la probabilidad de enfermedad cardíaca. Eso tiene sentido, ya que significa que tener una frecuencia cardíaca alta conduce a tener una enfermedad cardíaca.

# Anexos



