

**Universidad Nacional Autónoma de México**

**Facultad de Ciencias**

**Minería de Datos**

**Proyecto Final**

**Integrantes:**

**Alcantar Arenas Rogelio**

**Garduño Vargas Juan Carlos**

**Semestre: 2018-2**

**1.- Descripción del problema**

El problema consiste en determinar, a partir de los atributos en el dataset proporcionado, los factores que contribuyen a que una persona padezca enfermedades cardiacas.

**2.- Conocimiento de los Datos**

En esta tarea se trabajará con el dataset acerca de enfermedades cardiacas, almacenado en la página del curso: heart.csv y heartDescription

2.1. Elabora una tabla que contenga la siguiente informaciónon para cada atributo:

• Tipo de atributo (nominal, ordinal, numerico, etc)

• Porcentaje de valores perdidos.

• Valor minimo, maximo, media, desviaciónon estandar.

• ¿Existen registros que tengan un valor para ese atributo que no aparezca en otros registros?

• ¿Tiene valores atipicos?

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Atributos | Tipo | Porcentaje Valores Perdidos | Valor Mínimo | Valor máximo | Media | Desviación Estándar | Valor Atípico | Registros |
| Age | Numérico | 0% | 29 | 77 | 54.366 | 9.082 | No | Si |
| Sex | Nominal | 0% | - | - | - | - | No | No |
| Cp | Nominal | 0% | - | - | - | - | No | No |
| Trestbps | Numérico | 0% | 94 | 200 | 131.62 | 17.538 | No | Si |
| Chol | Numérico | 0% | 126 | 564 | 246.26 | 51.831 | No | Si |
| Fbs | Nominal | 0% | - | - | - | - | No | No |
| Restcg | Nominal | 0% | - | - | - | - | No | No |
| Thalach | Numérico | 0% | 71 | 202 | 149.54 | 22.905 | No | Si |
| Exang | Nominal | 0% | - | - | - | - | No | No |
| Oldspeak | Numérico | 0% | 0 | 6.2 | 1.04 | 1.161 | No | Si |
| Slope | Nominal | 0% | - | - | - | - | No | No |
| Ca | Numérico | 1.65% | 0 | 3 | 0.674 | 0.938 | Si | No |
| Thal | Nominal | 0.66% | - | - | - | - | Si | No |
| Num | Nominal | 0% | - | - | - | - | No | No |

**2.2 Haz una interpretación de los datos de acuerdo al estudio previo. Trata de determinar los atributos que aumentan el riesgo de enfermedades cardiacas.**

**Información de los datos:**

1.- age: Edad en años

2.- sex: genero (1 = hombre; 0 = mujer)

3.- cp: Tipo de dolor de pecho

* asympt: asymptomatic(asintomático)
* non\_anginal: non-anginal pain(dolor no anginal)
* atyp\_angina: atypical angina(angina atípica)
* typ\_angina: typical angina(angina típica)

4.- trestbps: Presion Arterial en reposo en mm Hg al ingreso en el hospital

5.- chol: colesterol sérico en mg / dl

6.- fbs: azúcar en sangre en ayunas > 120 mg/dl (1 = true; 0 = false)

7.- restecg: Resultados electrocardiograficos en reposo

* normal
* left\_vent\_hyper: showing probable or definite left ventricular hypertrophy by Estes' criteria(mostrando hipertrofia ventricular izquierda probable o definida según los criterios de Estes.).
* st\_t\_wave\_abnormality: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) - tener anormalidad de la onda ST-T (inversiones de la onda T y / o elevación del ST o depresión> 0.05 mV).

8.- thalach: maxima frecuencia cardiaca lograda

9.- exang: angina inducida por el ejercicio (1 = yes; 0 = no)

10.- oldpeak: Depresión ST inducida por el ejercicio en relación con el reposo.

11.- slope: la pendiente del segmento ST de ejercicio pico

* -- Value 1: upsloping
* -- Value 2: flat
* -- Value 3: downsloping

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

13.- thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

14.- num: diagnóstico de enfermedad cardíaca, estado angiográfico de la enfermedad

* <50: < 50% diameter narrowing(50% de diámetro de estrechamiento)
* >50\_1: > 50% diameter narrowing(50% de diámetro de estrechamiento).

**Analisis de los atributos que aumentan el riesgo de enfermedades cardiacas**

Age: La edad es importante ya que va muy ligado a al sexo, los hombres tienen más riesgo cardiovascular después de los 50 años de edad. En si su riesgo es alto. El factor protector de las mujeres son los estrógenos, empiezan con un riesgo cardiovascular después de la menopausia.

La presión de sangre: Es el principal factor de riesgo para el desarrollo de enfermedad cardiaca y evento vascular cerebral.

Colesterol: Las placas de ateroma son principales causas de infarto.

Angina inducida por el ejercicio: tenemos los síndromes coronarios agudos que son: angina estable, angina inestable y infarto agudo amiocardio.

la pendiente del segmento ST de ejercicio: Es un dato electrocardiográfico que habla de infarto, la elevación del segmento ST nos va hablar de infarto.

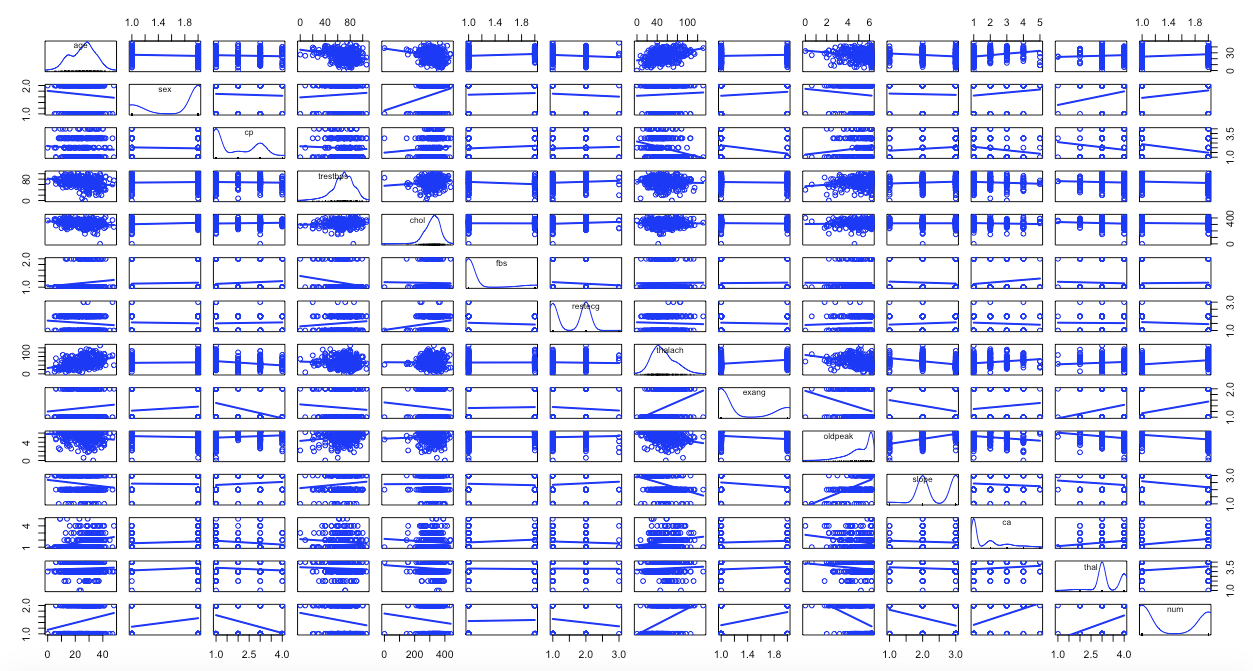
Discretización

Colesterol +200 es riesgo cardiovascular y siempre debe de estar a menos de 200

Trestbps += 140 es riesgo de infarto

Ca: Es los vasos que están obstruidos 0 significa 0 vasos obstruidos, y más igual a 1 es que están obstruidos

**2.3 Mediante una matriz de gráficas de dispersión, determina:**

****

1. ¿Cuáles atributos parecen estar más ligados a enfermedades cardiacas?

Age = edad

Trestbps = Presion de sangre

chol = Colesterol

thalach = frecuencia cardíaca máxima lograda

Oldpeak = depresión inducida por el ejercicio en relación al descanso

b) ¿Cuáles atributos parecen estar menos ligados a enfermedades cardiacas?

Sex = sexo

cp = Dolor de pecho

fbs= glucemia en ayunas

restecg = resultado de electrocardiográfico en reposo

exang = angina inducida por el ejercicio -

slope= la pendiente del segmento ST de ejercicio -

ca = número de vasos principales coloreados por fluroscopia -

thal = 3 = normal; 6 = fixed defect; 7 = reversable defect

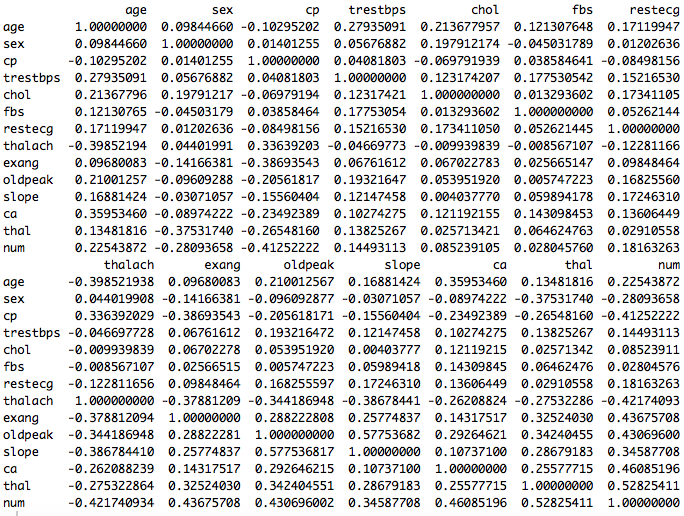
num: diagnóstico de enfermedad cardiaca

c) Resume en una tabla tus hallazgos relativos a la predicción de los valores de cada atributo.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

d) ¿Existen atributos correlacionados?

Obtenemos la matriz de correlación, por medio de la herramienta R



Si existen atributos correlacionados,

2.4 Investiga posibles asociaciones de atributos con el atributo de clase. Es decir, estudia las gráficas de dispersión elaboradas en el punto anterior y trata de identificar posibles áreas “densas” de enfermedades cardíacas.

-Si hubiera áreas “densas” en alguna(s) gráfica(s) de dispersión, cuantifica las enfermedades cardíacas en ellas con respecto al dataset completo.

PENDIENTE

**3 Preprocesamiento de datos(Preparación de datos).**

En este paso se preparan los datos de acuerdo a las tareas de minería que se van a realizar.

3.1 Selección de atributos

* Selecciona los atributos que consideres apropiados para una tarea predictiva. Justifica tu respuesta. Guarda los atributos seleccionados en un archivo llamado heart-c1.csv.

Los valores que se seleccionaron para un mejor análisis son:

* age
* sex
* trestbps
* chol
* fbs
* restecg
* thalach
* exang
* oldpeak
* slope
* ca
* thal
* num

Dado que ya se vio un análisis previo de los datos en el diagrama de dispersión y sabiendo que tan correlacionados están los datos, podemos asegurar que estos atributos nos arrojan mejores resultados de predicción para cualquier técnica de predicción.

3.2 Manejo de valores perdidos.

Considera los siguientes métodos para tratar con valores perdidos:

1. Reemplaza los valores perdidos por la media o la moda del atributo, de acuerdo al tipo de dato del atributo. Guarda el dataset resultante en un archivo con el nombre de heart-c2.csv

Listo

b) Utiliza regresión lineal para estimar los valores perdidos de cada atributo. Guarda el dataset resultante en el archivo heart-c3.csv.

LISTO

3.3 Eliminado de atípicos. Elimina los registros atípicos y guarda el resultado en el archivo heart-c4.csv

PENDIENTE

**4 Minado de datos(Construcción del modelo).**

Tareas de clasificación

Repetir los pasos descritos a continuación para cada dataset creado en el paso anterior y el dataset original (si es posible).

Utiliza un clasificador OneR

¿Qué se puede concluir? Compara estas conclusiones con las establecidas en el punto 2.

Compara la precisión del clasificador sobre el conjunto de entrenamiento con la estimaci´on de precisión obtenida mediante validación 10 ’fold-cross’. Si hay alguna diferencia, como la explicas.

Uso de un clasificador RIPPER

Describe los patrones obtenidos y compáralos con las conclusiones previas.

Usa un árbol de decisión C4.5

Utiliza diferentes valores para parámetros tales como podado y cantidad mínima de registros en las hojas.

Describe los patrones obtenidos y compáralos con las conclusiones previas.

Usa una red neuronal.

Utiliza diferentes valores para parámetros tales como momentum, tasa de aprendizaje, n´umero de épocas, cantidad de capas ocultas y/o n´umero de nodos en ellas (siempre que la herramienta lo permita).

Describe los patrones obtenidos y compáralos con las conclusiones previas.

Tareas de agrupamiento

Investiga si hay una tendencia de clustering en el dataset. Empieza agrupando los datos con el algoritmo k-medias, para algunas k, 2 ≤ k ≤ 10.

No uses el atributo de clase num.

Encuentra un valor adecuado para k. Justifica tu respuesta.

Usa el atributo de clase para evaluar el cluster y asegurate que las desviaciones estándar se calculan sobre los atributos numéricos.

Saca conclusiones de las medidas numéricas desplegadas para cada cluster

Investiga la posibilidad de usar la información del cluster para construir un clasificador para la variable num. Compara los resultados con los obtenidos con las técnicas anteriores.¿Cuál es mejor clasificador?

Conclusión(Prueba y evaluación).

En los pasos anteriores construiste varios modelos. Ahora debes compararlos para hacer una conclusión definitiva.

Elige alguna de las medidas de rendimiento para sustentar tu conclusión.

Resume en una tabla las medidas de rendimiento de cada clasificador y para cada dataset.

¿Qué puedes concluir?

Entregables(Despliegue).

………...