Almacenes y minería de datos
Proyecto Final
Dra. Amparo López Gaona y M en I. Gerardo Avilés Rosas
7 de Junio 2018

1 Descripción del problema

El objetivo de esta tarea es desarrollar un sistema de minería de datos, utilizando la metodología CRISP.

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.csve` y `heartDescription.txt`

1. Elabora una tabla que contenga la siguiente información para cada atributo:
   - Tipo de atributo (nominal, ordinal, numérico, etc)
   - Porcentaje de valores perdidos.
   - Valor mínimo, máximo, media, desviación estándar.
   - ¿Existen registros que tengan un valor para ese atributo que no aparezca en otros registros?
   - ¿Tiene valores atípicos?
2. Haz una interpretación de los datos de acuerdo al estudio previo. Trata de determinar los atributos que aumentan el riesgo de enfermedades cardiacas.
3. Mediante una matriz de gráficas de dispersión, determina:
   (a) ¿Cuáles atributos parecen estar más ligados a enfermedades cardiacas?
   (b) ¿Cuáles atributos parecen estar menos ligados a enfermedades cardiacas?
   (c) Resume en una tabla tus hallazgos relativos a la predicción de los valores de cada atributo.
   (d) ¿Existen atributos correlacionados?

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.

3 Preprocesamiento de datos

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

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`.

2. Manejo de valores perdidos.

   Considera los siguientes métodos para tratar con valores perdidos:

   (a) 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`

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

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

4 Minado de datos

4.1 Tareas de clasificación

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

1. Utiliza un clasificador `OneR`

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

   (b) Compara la precisión del clasificador sobre el conjunto de entrenamiento con la estimación de precisión obtenida mediante validación 10 'fold-cross'. Si hay alguna diferencia, cómo la explicas.

2. Uso de un clasificador `RIPPER`.

   (a) Describe los patrones obtenidos y compáralos con las conclusiones previas.

3. Usa un árbol de decisión C4.5

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

   (b) Describe los patrones obtenidos y compáralos con las conclusiones previas.

4. Usa una red neuronal.

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

   (b) Describe los patrones obtenidos y compáralos con las conclusiones previas.

4.2 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, $2 \leq k \leq 10$.

(a) No uses el atributo de clase `num`.

(b) Encuentra un valor adecuado para k. Justifica tu respuesta.

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

(d) Saca conclusiones de las mediadas numéricas desplegadas para cada cluster.

(e) 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?

5 Conclusión

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

1. Elige alguna de las medidas de rendimiento para sustentar tu conclusión.
2. Resume en una tabla las medidas de rendimiento de cada clasificador y para cada dataset.
3. ¿Qué puedes concluir?

6 Entregables

El 7 de junio harán una presentación de su trabajo. Además deberán entregar lo siguiente:

1. Un documento engargolado que incluya:
   - El proceso desarrollado, en él deben incluir capturas de pantallas y conclusiones. (Documento técnico
   - Una sección con las conclusiones finales del proceso, para presentar al cliente, en las cuales indiques los factores de riesgo para enfermedades cardíacas que encontraste, según esos datos.
2. Antes del 7 de junio deben haber enviado:
   - Si trabajaste con R, los scripts creados para el proyecto.
   - Si trabajaste en Weka:

   (a) El modelo generado en el Explorer, una vez que han ejecutado el algoritmo (se genera un archivo .model).
   (b) El archivo CSV generado, en el Experimentador, con los resultados.
   (c) El flujo de conocimiento (Knowlege Flow) que genera en un archivo .kf

- Una versión en pdf del documento engargolado.
- La presentación en pdf, que refleje el trabajo desarrollado. (esto es independiente de que la presentación la hagan en power point).