Diplomado en Minería de datos
Módulo 5. Minería de datos
Dra. Amparo López Gaona y M en I. Gerardo Avilés Rosas
Mayo 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 que, de acuerdo a los resultados obtenidos, propongas medidas prácticas para disminuir la cantidad de accidentes de tránsito occurridos.

2 Conocimiento de los datos

Se proporciona una base de datos con datos acerca de choques de autos registrados desde 1987 en los cuales se tiene al menos una persona herida.

Los datos con los que vas a trabajar están en la página en el archivo `accidentes.zip` y su descripción en el archivo `ACCIDENT.zip` Trabaja con los dataset: `accident.csv, person.csv` y `vehicle.csv` adicionalmente y de forma optativa con los otros dos datasets

Lo que se te pide en este apartado es:

1. Elaborar 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.
   - Si es numérico: tipo de distribución que parecer seguir (por ejemplo, normal).
   - ¿Tiene valores atípicos?
2. Haz una interpretación de los datos de acuerdo al estudio previo.
3. En esta etapa podrías determinar:
   (a) ¿Cuáles atributos parecen estar más ligados a riesgo de colisiones? Resume en una tabla tus hallazgos.
   (b) ¿Cuáles atributos parecen estar menos ligados a riesgo de colisiones?
   (c) Resume en una tabla tus hallazgos relativos a la predicción de los valores de cada atributo.
   (d) ¿Existen atributos correlacionados?
4. ¿Podrías eliminar variables? Justifica tu respuesta incluyendo las ventajas de esta acción.

3 Preprocesamiento de datos

En este paso se preparan los datos de acuerdo a las tareas de minería que se van a realizar. Algunos aspectos a considerar son:

1. Selección de atributos.

   Selecciona los atributos que consideres apropiados para una tarea predictiva. Justifica tu respuesta.

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.

   (b) Utiliza regresión lineal para estimar los valores perdidos de cada atributo.

3. Eliminación de atípicos.

4. Discretización de atributos numéricos. Si usas alguna discretización especifica, cuál usaste.

5. Normalización. Justifica la necesidad de la normalización.

Guarda el dataset resultante en un archivo con el nombre de `choques2.csv`
En tu reporte especifica cuáles tareas de preprocesamiento realizaste. Justificando el uso.

4 Minado de datos

4.1 Tareas de clasificación

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

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 con RIPPER 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 la herramienta lo permita).

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

5. Usa reglas de asociación para construir reglas de alta confianza para predecir `choques`.
   (a) Usa el método Apriori.
   (b) Describe los patrones obtenidos y compáralos con las conclusiones previas.
6. Haz una comparación de los resultados obtenidos con los distintos métodos

4.2 Evaluación de modelos

En el paso anterior construiste varios modelos. Necesitas evaluar la calidad de los modelos y compararlos.

1. Resume en una tabla las diferentes medidas de evaluación de cada clasificador para cada dataset.
2. ¿Qué puedes concluir?

4.3 Tareas de agrupamiento (clustering)

Investiga si hay una tendencia de agrupamiento en el dataset. Empieza agrupando los datos con el algoritmo k-medias para segmentar los registros de choques en grupos similares.

1. Perfila los clusters, es decir, qué se puede aprender del tipo de registros que hay en cada cluster. Descríbelo en Español.
2. Encuentra un valor adecuado para k. Justifica tu respuesta.
3. Usa el atributo de clase para evaluar el cluster y asegúrate que la desviación estándar se calcula sobre los atributos numéricos.
4. Saca conclusiones de las medidas numéricas desplegadas para cada cluster.

5 Conclusiones

Escribe unas conclusiones finales del proceso, para presentar al cliente, en las cuales indiques los factores de riesgo para choques que encontraste en esos datos después de analizarlos. Basado en tus hallazgos, indica algunas medidas que se puedan poner en práctica para prevenir los accidentes.

6 Entregables

El 24 de mayo harán una presentación de su trabajo. Además deberán entregar lo siguiente:

1. Un documento engargolado en que se muestre el proceso desarrollado, en él deben incluir capturas de pantallas y conclusiones de cada paso.
2. La presentación en pdf, que refleje el trabajo desarrollado. (esto es independiente de que la presentación la hagan en power point).
3. Antes del 24 de mayo deben haber enviado los scripts creados para el proyecto y una versión en pdf del documento engargolado.