CURSOS DE VERANO 2014

TÍTULO DEL CURSO: **APROXIMACIÓN PRÁCTICA A LA CIENCIA DE DATOS Y BIG DATA: HERRAMIENTAS KNIME, R, HADOOP Y MAHOUT**

TÍTULO PONENCIA: **KNIME. RESOLUCIÓN DE CASOS PRÁCTICOS**

NOMBRE PROFESOR: **CRISTÓBAL J. CARMONA**

MATERIAL ADICIONAL

- http://tech.knime.org/getting-started
- https://www.youtube.com/user/KNIMETV
- knime_tutorial.pdf
- http://informationandvisualization.de/blog/knime-interactive-views
- http://www.dataminingreporting.com/blog/category/knime
- http://tech.knime.org/forum/knime-general

Caso 2 – Preprocesamiento

Objetivo:

- **Analizar la importancia en la preparación de los datos.**
- **Aplicar distintas técnicas de preprocesamiento.**
- **Justificar estrategias de actuación sobre el problema realizado.**

CASO 2: Preprocesamiento

DESCRIPCIÓN

- Analizar la importancia de la preparación de los datos.
- Se analizarán los datos obtenidos mediante el "General Estimate System" de EEUU.
- Componente de la Admon. Nac. de Seg. del Tráfico.
- Datos se obtienen de una muestra de 6,4 millones de accidentes al año.

CASO 2: Preprocesamiento

DESCRIPCIÓN

- Accidentes incluyen los:
  - mortales,
  - causan lesiones, y
  - causan daños materiales.

- Sin embargo, el *GES* se centra en los accidentes con mayor preocupación para la comunidad.

CASO 2: Preprocesamiento

DESCRIPCIÓN

- Principal objetivo es identificar áreas con problemas de seguridad, base para información al consumidor, normativas, análisis de costes y beneficios de la seguridad, etc.

- Se trabajará con una base de datos pública con casi 56 mil accidentes.

AÑO 2001 - http://www.transtats.bts.gov/Fields.asp?Table_ID=1158

DESCRIPCIÓN

- Conjunto de datos complejo:
  - 55964 accidentes
  - 45 variables (categóricas y numéricas)
- Obtener un modelo basado en árboles de decisión para predecir la gravedad del daño causado.
- Sin embargo, datos presentan múltiples deficiencias.

TAREAS A REALIZAR

- Análisis previo del conjunto de datos.

- Transformar las tres variables objetivo en una única variable:
  - FATALITIES
  - INJUTY_CRASH
  - PRPTYDMG_CRASH

TAREAS A REALIZAR

- Reemplazar los datos desconocidos por valores en blanco.

- Generar un nuevo fichero excel o una nueva hoja para realizar las modificaciones.

TAREAS A REALIZAR

- Cargar el conjunto de datos en la herramienta KNIME.
- Trabajaremos siempre con el algoritmo de clasificación “C4.5“.
- Trabajar con una partición entrenamiento-prueba (80%-20%).

TAREAS A REALIZAR

- Los análisis deben de basarse en precisión, comprensibilidad y visualización.
- Crear modelos a partir de:
  - Variables inputadas (Finalizan en "_I")
  - Variables no-inputadas
- En total serían 27 variables más 1 Clase

TAREAS A REALIZAR

DISCRETIZACIÓN

- Analizar mediante C4.5 la división de las variables numéricas.
- Probar a discretizar variables como la hora del accidente o velocidad mediante cuartiles o manualmente.

TAREAS A REALIZAR

DISCRETIZACIÓN

- Analizar mediante CAIM (previo a C4.5) sobre las variables numéricas.
- Analizar las variables categóricas para una posible reducción de valores.
- Comparar todos los resultados y obtener conclusiones.

TAREAS A REALIZAR

VALORES PERDIDOS

- Analizar el comportamiento de C4.5 sobre los valores perdidos.
- Imputar valores perdidos con media o moda y comprobar los resultados.
- Eliminar instancias con algún valor perdido y analizar los resultados.

TAREAS A REALIZAR

VALORES PERDIDOS

- Emplear un algoritmo de predicción para imputar valores perdidos y analizar los resultados.
- Comparar todos los resultados y obtener las conclusiones.

TAREAS A REALIZAR

SELECCIÓN DE CARACTERÍSTICAS

- Analizar el comportamiento del algoritmo C4.5 en el descarte de algunas variables.
- Analizar la utilización de una selección de características envolvente hacia atrás. Se puede forzar a eliminar o conservar variables.
- Utilizar método envolvente basado en KNN, o Bayes, por ejemplo.

TAREAS A REALIZAR

SELECCIÓN DE CARACTERÍSTICAS

- Comparar todos los resultados y obtener las conclusiones sobre la selección de características.

TAREAS A REALIZAR

SELECCIÓN DE INSTANCIAS - DUDOSO

- Se permite trabajar con el conjunto de datos obtenido en la etapa anterior (siempre que se mejore la precisión).
- Analizar la utilización del algoritmo CNN.
- Analizar la utilización del algoritmo ENN.

TAREAS A REALIZAR

SELECCIÓN DE INSTANCIAS

- Datos no se encuentran balanceados.
  - Analizar la utilización de algoritmos de reducción de datos mediante muestreo aleatorio, para equilibrar la frecuencia de la clase.
  - Realizar un incremento de los ejemplos para la clase minoritaria.

TAREAS A REALIZAR

SELECCIÓN DE INSTANCIAS

- Comparar todos los resultados y obtener las conclusiones sobre la selección de características.

CURSOS DE VERANO 2014

TÍTULO DEL CURSO: **APROXIMACIÓN PRÁCTICA A LA CIENCIA DE DATOS Y BIG DATA: HERRAMIENTAS KNIME, R, HADOOP Y MAHOUT**

TÍTULO PONENCIA: **KNIME. RESOLUCIÓN DE CASOS PRÁCTICOS**

NOMBRE PROFESOR: **CRISTÓBAL J. CARMONA**