Sesión 3 - Integración de KNIME y WEKA

Elaborad un informe con el trabajo realizado para los ejercicios que se proponen a continuación. Empaquetad junto con el informe los flujos de trabajo exportados en un fichero comprimido. Incorporad capturas de pantalla de la configuración de los nodos y exporta imágenes de los flujos de trabajo en formato vectorial (SVG) para ilustrar tu trabajo.

Ejercicio 1

Descarga el fichero de datos "yellow-small.data" del conjunto de datos Balloons Data Set (http://archive.ics.uci.edu/ml/datasets/Balloons) y prepara un flujo de trabajo que realice las tareas que se proponen a continuación.

1. Lee el fichero y renombra las columnas de forma adecuada ("Color", "Size", "Act", "Age", "Inflated (True/False)".

2. Añade la siguiente columna de clasificación ("class"):

```
IF Color = yellow AND Size = Small
  => class = inflated
ELSE
  => class = not inflated
```

3. Añade una columna llamada "full sentence" que complete la anterior:
   "inflated is T"
   OR
   "not inflated is F"
   en la que el valor "Inflated" y "Not inflated" se tomen de la columna "Class" y "T/G" de la columna "Inflated (True/False").

   Nota: La columna "Full sentence" no deberá contener comillas.

Ejercicio 2

Realiza las tareas que se proponen usando un conjunto de datos que describe el número de visitantes a un sitio web en los meses de junio/julio de 2010 (fichero *"website1.txt"*).

1. Calcula parámetros estadísticos sobre el número de visitantes (media, desviación, ...). ¿Qué es el coeficiente de Kurtosis? Puedes buscar información sobre este estadístico en la Wikipedia ¿Qué dice este coeficiente sobre cualquier variable? Ilústralo con un ejemplo sobre el conjunto de datos del ejercicio.

2. Entrena una red Bayesiana Naïve sobre el número de visitantes para tratar de descubrir cuándo los datos recogidos para una fila se corresponden a un día de diario o a fin de semana. ¿Qué variables dan mejores resultados usando este clasificador? Explica las pruebas realizadas.

3. Dibuja la curva ROC (https://es.wikipedia.org/wiki/Curva_ROC) para visualizar el rendimiento del clasificador Bayesiano Naïve.

Ejercicio 3

En este ejercicio vamos a reutilizar workflows hechos previamente. La idea es elaborar una comparativa a gran escala de los modelos usados tanto con los nodos de KNIME como de Weka.

1. Prepara uno o varios workflows reutilizando los realizados en sesiones anteriores que entrene modelos Decision Tree (J48), Multilayer Perceptron y Naive Bayes utilizando tanto los nodos de KNIME y de Weka. Puedes utilizar metanodos para organizar el trabajo. Como conjuntos de entrenamiento / prueba utiliza los ficheros "wine.data", "iris.data" y "adult.data" (https://archive.ics.uci.edu/ml/datasets/Adult).

2. Realiza cinco ejecuciones con al menos tres proporciones distintas (p.e. 30 %, 50 % y 80 %) para los conjuntos de entrenamiento y prueba. Luego compara los resultados para todos los modelos entrenados. Ten en cuenta que, de cara a obtener resultados consistentes, debes hacer varias ejecuciones sobre los mismos conjuntos de datos para trabajar con un promedio. Aparte de utilizar las tablas con los resultados, idea la manera de representar gráficamente cuáles son los modelos que proporcionan los mejores resultados en cada caso. Puedes apoyarte en la herramienta de Reporting de KNIME o usar Excel directamente para hacer tus cálculos.

3. Documenta el workflow, describe los cambios u adaptaciones que hayas tenido que hacer en los nodos y datos y, documenta los problemas surgidos. ¿Has podido utilizar todos los modelos propuestos? ¿Te han surgido problemas? ¿Qué soluciones has aportado?