# MINERIA DE DATOS USANDO SISTEMAS INTELIGENTES

# PRACTICA 3 – REGLAS DE CLASIFICACION

***Material de Lectura: Capítulo 11 del Libro Intr. a la Minería de Datos de Hernández Orallo***

## Ejercicio 1

La tabla **Curso1erAnio.xls** contiene datos referidos al desempeño de algunos alumnos en cierto curso de primer año de la Facultad de Informática de la UNLP en años anteriores.
Utilice RapidMiner para obtener, a partir de la información de la hoja **CursoNumerico** de la tabla **Curso1erAño.xls**, las reglas de clasificación que permitan predecir si un alumno aprobará o no el curso.

a) Utilice dos operadores "**Discretize by User Information"**: uno para discretizar el atributo "TRABAJA" en tres casos: No trabaja, trabaja a lo sumo 25 hs y trabaja más de 25 hs y otro para discretizar el atributo "ASISTENCIA" en dos grupos distinguiendo los que vinieron a lo sumo al 50% de las clases de los que vinieron a más del 50% de las clases.
b) Obtenga el conjunto de reglas correspondientes con los métodos ZeroR, OneR y PRISM.
c) Compare la performance de cada modelo al ser aplicado sobre el conjunto de datos de la hoja **CursoNumerico**. Complete la siguiente tabla

| Método | Cantidad de Reglas | Long. promedio del antecedente | Precisión (accuracy) |
|---|---|---|---|
| ZeroR | | | |
| OneR | | | |
| PRISM | | | |

d) Aplique el modelo obtenido en cada caso para clasificar los ejemplos de la hoja **CursoTesteo** contenida dentro del mismo archivo Excel discretizando como se indicó en el ejercicio 1.a).

## Ejercicio 2

El archivo **Globos.xlsx** contiene datos referidos a un experimento psicológico. De cada instancia se sabe el color del globo, el tamaño, si fue inflado por un adulto o un niño y si se estira o no. El objetivo es construir un modelo predictivo para determinar si un globo permanecerá inflado o no.

a) A partir de los datos disponibles determine, utilizando los métodos OneR y PRISM, el conjunto de reglas de clasificación que permita predecir si un globo dado permanecerá inflado o no (atributo "Inflado?"). Calcule el soporte, la confianza y el interés de cada una de las reglas obtenidas.

   La resolución de este ejercicio debe realizarse de forma manual. En ambos casos deberá detallar los cálculos realizados para determinar las reglas indicadas.

b) En base a las reglas conseguidas responda a las siguientes preguntas:

- ¿hay algún tipo de globo que no debió haber sido comprado?
- ¿Hay algún tipo de globo que sólo pueden inflar los adultos?

c) Utilice RapidMiner Studio para obtener las reglas del punto a) y realizar un árbol de clasificación ID3 para los mismos datos. Compare las características de los modelos obtenidos haciendo uso de los operadores "Apply Model" y "Performance" para analizar la precisión de cada modelo. Tenga en cuenta que como se dispone de pocas muestras no se dividirá el juego de datos en entrenamiento y testeo.

## Ejercicio 3

A partir de los datos del archivo **Drug5.xlsx**, que contienen las muestras de pacientes a los que se les ha suministrado una droga determinada:

a) Use el operador "**Discretize by bins**" con 2, 5 y 10 intervalos. Obtenga con el método PRISM las reglas de clasificación que sean capaces de decidir la droga a suministrarle a un paciente. Observe la cantidad de reglas obtenidas en cada caso.

b) Observe la relación entre los atributos **Na** y **K** utilizando las técnicas de visualización vistas. ¿Se puede concluir algo respecto a cómo están repartidas las clases?

c) Utilice el operador "**Generate Attribute"** para generar un atributo nuevo que sea el cociente entre **Na** y **K**. Analice el atributo generado utilizando las técnicas de visualización vistas.

d) Proponga intervalos utilizando el operador **"Discretize by user specification"** en base al análisis realizado en b) que permitan reducir la cantidad de reglas obtenidas en c).

e) ¿Qué puede decir respecto a la discretización según la cantidad de reglas obtenida en cada caso?

## Ejercicio 4

Indique el soporte, la confianza y el interés de cada una de las reglas halladas en el ejercicio 1.b) para el método PRISM. En cada caso, incluya el cálculo utilizado para obtener la medida solicitada. Explique el resultado obtenido.