

# Minería de Datos usando Sistemas Inteligentes

## Guía para la realización del trabajo integrador

El trabajo integrador tiene como objetivo practicar y demostrar el uso de los modelos desarrollados durante la cursada mediante la aplicación del proceso KDD. No obstante, no es el objetivo entregar gráficos o resultados de los modelos sin analizar. Por ese motivo no hay una longitud mínima o máxima establecida. El trabajo no tendrá una valoración mayor cuanto más extenso sea.

Se debe priorizar sobre todas las cosas el análisis personal de los datos y las conclusiones que obtengan sobre los mismos. En ese espíritu, el informe debe ser lo más concreto y claro posible, y sólo incluir la información relevante para las conclusiones que presentan.

A continuación, se detallan items que pueden incluirse en el informe:

### 1) La descripción del conjunto de datos elegido:

#### - Introducción

- Breve explicación del dominio. Por ejemplo: Si el conjunto de datos trata sobre géneros musicales, explicar aquellos que no sean muy conocidos y las diferencias entre los mismos. Por ejemplo: "La Bossa Nova es un género musical de origen brasilero que mezcla Jazz y Samba".

- Cantidad de ejemplos, cantidad y tipo de atributos, explicación de los atributos. Por ejemplo: "El atributo GN indica el género musical, y está codificado como string representando un atributo polinomial. Hay 25 géneros distintos. En el siguiente gráfico podemos ver la prevalencia de los distintos géneros \_\_\_\_\_"

- Forma de recolección del conjunto de datos (encuesta online, encuesta presencial, scraping, etc.) y detalles del proceso de recolección. Por ejemplo: "Los datos fueron recolectados mediante un formulario online (imagen del formulario) en donde los usuarios voluntariamente llenaban los campos \_\_\_\_\_"

- Otra información que resulte relevante como el análisis de datos faltantes, redundancia entre atributos, unificación de valores, etc.

2) Hipótesis, preguntas de interés y objetivos del proceso de minería de datos a desarrollar. Por ejemplo: "Para una emisora de música en la radio, resulta útil saber cuáles géneros musicales pueden combinarse en un mismo programa. Por ende, vamos a realizar un clustering de los gustos musicales del conjunto de datos de los oyentes de manera de poder establecer una mejor programación. Además, se quiere \_\_\_\_\_"

3) La descripción del preprocesamiento realizado al conjunto de datos: eliminación/creación/modificación de atributos, selección de ejemplos, corrección de errores, unificación de archivos, etc. Por ejemplo: "Para poder ejecutar el algoritmo K-Medias, al atributo GN se le realizó una conversión de polinomial a binomial, resultando 25 atributos con \_\_\_\_\_"

4) Detalle de los experimentos realizados junto con sus resultados. Por ejemplo: "Realizamos un clustering con el algoritmo K-Medias sobre tal partición de datos para observar los grupos de géneros que se forman. El valor de K fue determinado mediante \_\_\_\_\_. Los centros resultantes son \_\_\_\_\_."

5) Análisis de los resultados y conclusiones. Por ejemplo: "En base al clustering realizado con  $K=3$ , encontramos que hay un grupo pequeño pero muy compacto de oyentes que prefiere escuchar mayormente rock progresivo, jazz rock, jazz y bossa nova; por otro lado, los otros dos grupos combinaban una gran cantidad de géneros, lo que sugiere que \_\_\_\_\_."