Actividad evaluable – Actividad 1

ASIGNATURA: Minería de Datos

Título: *Máster Universitario en Big Data y Ciencia de Datos*
Materia: *Ciencia de Datos*
Créditos: *6 ECTS*
Código: *05MBID*

Índice

1. Actividad

DESCRIPCIÓN	
Introducción	El objetivo de esta actividad es introducir los conceptos básicos de la minería de datos. Para ello el alumnado trabajará en el desarrollo de un proceso de extracción de conocimiento a partir de bases de datos (KDD del inglés).
Objetivo	Esta actividad consiste en implementar únicamente las etapas de preprocesamiento y transformación de datos de un proceso de KDD. Dicho proceso deberá resolver un problema de clasificación binaria (de dos valores de clase, como puede ser el aprobar o no un examen, otorgar o no un crédito, entre otros ejemplos).
Trabajo previo	Lectura del material docente disponible en Recursos y materiales > Material del profesor y de todos los hilos de conversación del Foro que se generen.
Metodología	En las sesiones se expondrán los conocimientos, materiales e indicaciones necesarias para que el alumnado pueda realizar esta actividad guiada sin dificultades. Se establecerán las pautas concretas y la dinámica que se deberá seguir para realizar la actividad propuesta.
Tarea para el portafolio	Implementar en un Jupyter Notebook, únicamente las etapas de preprocesamiento y transformación del proceso de KDD. Para lo cual, el alumno deberá buscar en Internet un dataset con el que trabajar. Los siguientes sitios web son sólo algunos de los tantos repositorios de datos en Internet de los cuales se puede seleccionar el dataset a utilizar: - https://archive.ics.uci.edu/datasets - https://www.kaggle.com/datasets Además, si disponen de un dataset propio que puedan y quieran utilizar, podrás hacerlo, pero recuerden que deben subir con el Notebook el fichero de datos y por tanto, deberán encargarse de anonimizar los datos si fuese necesario. También, pueden utilizar el paquete datasets de la librería de Scikitlearn para utilizar alguno de los datasets disponibles. Por lo visto, permite la misma librería importar incluso datasets que provienen del repositorio https://openml.org/. En cuanto al proceso de KDD, se deberá asumir que la etapa de selección de datos (anterior a la etapa de preprocesamiento) ya fue realizada con anterioridad y cuyo resultado es el fichero de datos que cada uno elija para resolver la actividad. Este fichero de datos será la entrada de la etapa de preprocesamiento. Recordar que el dataset a utilizar debe permitir resolver un problema de clasificación binaria, es decir, que debe contar con una variable de tipo target o clase con dos valores posibles. Esto debe ser tenido en cuenta al momento de buscar y seleccionar el dataset. A su vez, intenten seleccionar un dataset que tenga varias variables categóricas y numéricas, como así también valores, faltantes y fuera de rango. Tenga en cuenta que en el Jupyter notebook lo primero que deberá hacerse es leer el fichero de datos de entrada y lo último, generar como resultado un nuevo fichero de datos de salida, con los datos ya preprocesados y transformados correctamente de modo que estén preparados para ser utilizados en la construcción del modelo de minería de datos que tendrán que realizar en la Actividad Guiada 2. Para aprobar con la nota mínima se deberá: - Indicar el tamaño del conjunto de datos - Verificar la existencia de duplicados - Especificar las variables de cada tipo - Determinar si hay valores fuera de rango en al menos un atributo numérico - Determinar si hay valores faltantes en al menos un atributo categórico o numérico - Calcular estadísticas de una variable numérica y otra categórica - Realizar el diagrama de barras de al menos una variable categórica - Realizar el histograma de al menos una variable numérica - Identificar la variable target y verificar el balance entre clases

Para alcanzar la nota máxima, se podrá entre otras cosas:

- Tratar los valores faltantes
- Tratar los valores fuera de rango
- Realizar diagramas de dispersión
- Generar variables nuevas
- Normalizar y/o discretizar variables
- Analizar la correlación entre variables
- Mapear valores de atributos categóricos

Esta actividad no solo debe ser vista como una actividad puntual (que tiene continuidad con la Actividad Guiada 2), sino también como una posible base para proyectos futuros en otras asignaturas y/o para el Trabajo de Fin de Máster (TFM).

El alumnado deberá entregar un fichero con extensión .ipynb con el código desarrollado. Antes de entregar, verificar que el cuaderno se pueda ejecutar de inicio a fin sin errores. No olvide entregar junto con el fichero .ipynb, el fichero de datos (en la extensión que el fichero haya sido leído en el Notebook, csv por ejemplo).

2. Rúbrica de evaluación

	Suspenso (< 5)	Aprobado (> = 5)	Sobresaliente (> = 9)
Estilo (30%)	Codificación incoherente y/o con errores.	Codificación coherente y sin errores.	Codificación coherente y sin errores, yendo más allá de lo básico esperado.
Contenido (40%)	No hace lo mínimo solicitado.	Hace lo mínimo solicitado.	Hace lo mínimo solicitado y además realiza los extras.
Originalidad y pasos adicionales (30%)	No utiliza las librerías y métodos adecuados	Utiliza las librerías y métodos adecuados.	Utiliza las librerías y métodos adecuados, pero también incorpora los no convencionales.

3. Bibliografía

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM, November 1996/ Vol 39, Nº 11, 27–34.

3.1. Bibliografía de referencia

- Model, F. E., Williams, G. J., & Huang, Z. (1996). Modelling the KDD Process.
- Brachman, R. J., & Anand, T. (1994, July). The Process of Knowledge Discovery in Databases: A First Sketch. In KDD workshop (Vol. 3, pp. 1-12).

3.2. Bibliografía complementaria

- Siegel, E. (2013). Analítica predictiva. Predecir el futuro utilizando Big Data. Anaya Multimedia-Anaya Interactiva.
- Mayer-Schönberger, V., Cukier, K. (2013). Big data. La revolución de los datos masivos. Turner.