Minería de datos: PRA1 - Selección y preparación de un juego de datos

Autor: Pablo Suárez

Noviembre 2023

Contents

Enunciado

Todo estudio analítico debe nacer de una necesidad por parte del **negocio** o de una voluntad de dotarle de un conocimiento contenido en los datos y que solo podremos obtener a través de una colección de buenas prácticas basadas en la Minería de Datos.

El mundo de la analítica de datos se sustenta en 3 ejes:

A. Uno de ellos es el profundo **conocimiento** que deberíamos tener **del negocio** al que tratamos de dar respuestas mediante los estudios analíticos.

B. El otro gran eje es sin duda las **capacidades analíticas** que seamos capaces de desplegar y en este sentido, las dos prácticas de esta asignatura pretenden que el estudiante realice un recorrido sólido por este segundo eje.

C. El tercer eje son los **Datos**. Las necesidades del Negocio deben concretarse con preguntas analíticas que a su vez sean viables responder a partir de los datos de que disponemos. La tarea de analizar los datos es sin duda importante, pero la tarea de identificarlos y obtenerlos va a ser para un analista un reto permanente.

Como **primera parte** del estudio analítico que nos disponemos a realizar, se pide al estudiante que complete los siguientes pasos:

1. Plantear un problema de analítica de datos detallando los objetivos analíticos y explica una metodología para resolverlos de acuerdo con lo que se ha practicado en las PEC y lo que se ha aprendido en el material didáctico.
2. Seleccionar un juego de datos y justificar su elección. El juego de datos **deberá tener capacidades** para que se le puedan aplicar **algoritmos supervisados**, **algoritmos no supervisados** y **reglas de asociación** y deberá estar alineado con el problema analítico planteado en el paso anterior.

Requisito mínimo: El juego de datos deberá tener como mínimo 500 observaciones con un mínimo de 5 variables numéricas, 2 categóricas y 1 binaria. Además **debe ser distinto**, es importante que no sea un dataset usado en las PEC anteriores.

Adjuntamos aquí una lista de portales de datos abiertos para seleccionar el juego de datos. Se pueden usar otras fuentes para obtener vuestro juego de datos, pero recordad de citarlas:

- **Datos abiertos**
  - Google Dataset Search
  - Datos abiertos España
  - Datos abiertos Madrid
  - Datos abiertos Barcelona
  - Datos abiertos Londres
  - Datos abiertos New York
- **Conjuntos de datos para aprendizaje automático e investigación**
  - UCI Machine Learning
  - Datasets for machine-learning research (Wikipedia)
  - Kaggle datasets

3. Realizar un análisis exploratorio del juego de datos seleccionado.
4. Realizar tareas de limpieza y acondicionado para poder ser usado en procesos de modelado.
5. Realizar métodos de discretización
6. Aplicar un estudio PCA sobre el juego de datos. A pesar de no estar explicado en el material didáctico, se valorará si en lugar de PCA investigáis por vuestra cuenta y aplicáis SVD (Single Value Decomposition).

- **Algunos recursos**
  - PCA para reducción de dimensiones
  - SVD Singular Value Decomposition

Recordad que para todas las PRA es **necesario documentar** en cada apartado del ejercicio práctico que se ha hecho, por qué se ha hecho y cómo se ha hecho. Asimismo, todas las decisiones y conclusiones deberán ser presentados de forma razonada y clara, **contextualizando los resultados**, es decir, especificando todos y cada uno de los pasos que se hayan llevado a cabo para su resolución. Por último, incluid una **conclusión final** resumiendo los resultados obtenidos en la práctica e indicad eventuales **citaciones bibliográficas**, fuentes internas/externas y materiales de investigación.

- **Documento entregable**

Solo hay que entregar el documento html y hay que hacerlo con el siguiente nombre 75.584-PRA1-NombreEstudiante.html

---

Recursos de programación

- Incluimos en este apartado una lista de recursos de programación para minería de datos donde podréis encontrar ejemplos, ideas e inspiración:
  - Material adicional del libro: Minería de datos Modelos y Algoritmos
  - Espacio de recursos UOC para ciencia de datos
  - Buscador de código R
  - Colección de cheatsheets en R
