Taller 1 Minería Datos: Pre processing

Samantha Reid

Octubre 2022

Fecha de entrega: 16 de Octubre, 23:59.

1 Indicaciones

- El taller deberá ser enviado hasta el día 16 de Octubre a las 23:59.
- Cualquier envío que se haga pasado esa hora tendrá una penalización de 1.0 puntos por cada hora de atraso, **sin posibilidad de apelación**.
- El mínimo de integrantes es uno y el máximo es dos alumnos. **No habrán excepciones de ningún tipo.**
- Se deberá hacer envío de las preguntas en un documento (ya sea Word o PDF) y el código comprimido o enviando el código de Colab.
- Los códigos deben escribirse en **Python 3.x**. Cualquier otro código o versión será evaluado con nota mínima de manera inmediata.
- Se deberá indicar dentro de los documento el nombre de los integrantes.
- No se podrá apelar posteriormente por ninguna falta de un integrante, ya sea por motivos de: 'mi compañero no aportó nada al trabajo' o 'al momento de enviar del trabajo se me olvidó colocar su nombre'.
- El envío de la resolución del taller deberá ser enviado al correo s.reidc@utem.cl con el título 'Taller 1 *xx yy*', donde *xx* corresponde al nombre del primer estudiante e *yy* al nombre del segundo estudiante si lo hubiera.

2 Introducción

El presente trabajo tiene como objetivo comprender y analizar los datos seleccionados del *datasets* (set de datos), llamados "Licencias Médicas", los cuales están publicados en la Superintendencia de Salud y pueden ser accedidos mediante este link.

El dataset corresponde a las licencias médicas emitidas durante el año 2022 en el segundo semestre. Usted tiene como objetivo realizar un pre procesamiento de los datos, junto a un análisis descriptivo básico del mismo. Recordar que en la misma página se encuentra el diccionario de los datos, el cual se recomienda leer en conjunto.

Es importante mencionar que es posible que se enfrente con datos que presenten elementos faltantes o datos extremos, los cuales pueden afectar el análisis pertinente. En esos casos es necesario realizar una limpieza previa.

Además, cada una de estas preguntas deberá ser contestada con los datos del **2022 del segundo semestre**.

Por último, recordarles que cada supuesto que ustedes realicen deberá estar documentado, tanto en el documento escrito como en el código como comentario.

1. Identificar los tipos de variables en el set de datos y proponer potenciales usos.
2. Aplicar transformación de datos según corresponda.
3. Realizar agrupamiento de categorías similares en los datos categóricos
4. Limpiar datos nulos (tanto en filas como en columnas).
5. Binarizar variables categóricas.
6. Realizar un análisis descriptivo básico, donde se muestren las medidas de tendencias generales y su variabilidad. (*)

(*) **HINT:** Pueden ayudarse con el ejemplo de este link
Cada pregunta equivale a un punto de nota base.