IIC2413 — Bases de Datos

Control 3: manejo de datos para un problema real

Durante esta actividad trabajaremos con los datos del Sistema Nacional de Información Municipal del gobierno de Chile. Vamos a trabajar con dos fuentes de datos:

- Una planilla llamada presupuesto_2019.csv, que contiene información sobre el presupuesto municipal, para cada una de las comunas de Chile.
- Una base de datos en SQLite llamada municipios.db, que contiene información sobre la cantidad de personas, los metros de plazas y los metros de parques de cada municipalidad. Esos datos se almacenan en un esquema con tres tablas: (1) una tabla `Comuna(id, nombre)`, que asigna a cada comuna en Chile un identificador; (2) una tabla `Areas_Verdes(id_comuna, metros_plaza, metros_parque)`, que asigna a cada comuna dos números, un número es la cantidad de metros cuadrados de plazas en esa comuna, y el otro es la cantidad de metros cuadrados de parques (para ver la diferencia entre parques y plazas, pensar en que el Cerro San Cristóbal es un parque, y que la Plaza de Armas es una plaza); y (3) una tabla `Personal(id_comuna, personas)`, que asigna a cada comuna la cantidad de personas que trabajan ahí.
- Recuerda que en SQLite puedes poner `.table` para ver las tablas que hay, y puedes poner `.schema <tabla>` para ver el esquema de una tabla llamada `<tabla>`.

La idea de este control es hacer un análisis de datos en torno a la siguiente problemática:

> ¿Como se relacionan los metros de áreas verdes y el tamaño del personal de las municipalidades con su presupuesto?

Para la parte evaluada de esta tarea trabajaremos solo con las áreas verdes. Dejamos el trabajo con el personal como un ejercicio propuesto.

1. Manejo de datos

En este control, asumimos que ya sabes trabajar tanto con SQLite como con la librería `pandas` de Python. Para trabajar tus datos necesitas tener todo en pandas, pero puede ser que quieras importar datos a SQLite antes de eso. Puede ser también que prefieras hacer el join en python, y para eso no necesitas importar. Para importar a SQLite, tienes dos opciones.

Importar desde la consola de SQLite3 Esta opción requiere que instales SQLite3 en tu computador. Aquí tienes que abrir la base de datos en el cliente de SQLite3 (el comando debería ser `sqlite3 municipios.db`) y luego:

- Crea en SQL tablas de para las comunas y sus presupuestos, con atributos de acuerdo a lo que ves en el archivo .csv.
- Dentro de SQLite3, ejecuta el comando `.separator ,`, para indicar que el separador es el caracter `,`.
- Ejecuta `.import archivo.extension tabla` para importar el archivo a la tabla.

Importar los datos utilizando Python Existen varias formas de importar los datos utilizando Python. Una forma es la propuesta en `https://stackoverflow.com/questions/2887878/importing-a-csv-file-into-a-sqlite3-database-table-using-python`.

2. Tareas a realizar

Ten cuidado: puede ser que los datos estén sucios, pues son llenados por humanos. Es tu responsabilidad limpiar esos datos: para las siguientes tareas no debes tomar en cuenta datos nulos o que no han sido reportados (aunque eso signifique dejar fuera algunas comunas).

2.1. Correlación entre áreas verdes y presupuesto

1. Crea un gráfico de puntos para visualizar la correlación entre el presupuesto de las comunas y la superficie en metros cuadrados de plazas con la que cuentan esas comunas. Entrega el coeficiente de correlación entre esos dos campos.
2. Crea un gráfico de puntos para visualizar la correlación entre el presupuesto de las comunas y la superficie en metros cuadrados **total**, es decir, sumando plazas con áreas verdes, con la que cuentan esas comunas. Entrega el coeficiente de correlación entre esos dos campos.
3. En base a esos dos experimentos, argumenta qué campo es mejor para ser utilizado en una regresión: la superficie de plazas o la superficie de plazas sumadas con parques.

2.2. Regresión

Sea S el indicador de superficie que elegiste en la parte anterior (ya sea plazas o suma de plazas con parques).

4. Entrega un modelo de regresión que te permita predecir el presupuesto de una comuna, en función de S. Utiliza *cross-validation*, dividiendo en 5 partes iguales.

2.3. Regresión - ejercicio adicional (no evaluado)

Busca como implementar una regresión cuando la respuesta depende de más de un parámetro. Ajusta ahora un modelo en el que el presupuesto no depende solo de los metros cuadrados de áreas verdes, si no que también del personal de esa municipalidad. ¿Crees que tu modelo quedó mejor? ¿Cómo podrías probarlo?

3. Entrega y detalles administrativos

Este control es individual. La entrega de este control debe ser un archivo comprimido donde se encuentre un **Jupyter Notebook**, junto a cualquier archivo .csv o .db que estés llamando desde tu código. El plazo para el control es el **Lunes 6 de Julio, a las 20:00 hrs.**.
La nota se calcula como un promedio ponderado entre **dos veces** la nota de la parte 2.1 y **una vez** la nota de la parte 2.2.