Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

IIC2115 – Programación como Herramienta para la Ingeniería (I/2022)

Ejercicio Capítulo 2a

Aspectos generales

- **Objetivos:** Aplicar los contenidos de análisis exploratorio de datos para completar una base de datos incompleta y responder consultas sobre la misma.
- **Lugar de entrega:** lunes 11 de abril a las 16:30 hrs. en repositorio privado.
- **Formato de entrega:** archivo Python Notebook (**C2a.ipynb**) con el avance logrado durante la sesión. El archivo debe estar ubicado en la carpeta **C2a**. Utilice múltiples celdas de texto y código para facilitar el trabajo del cuerpo docente.

Introducción

Con el fin de ejercitar los contenidos de análisis exploratorio de datos en Python, en este ejercicio deberá realizar los pasos básicos para completar una base de datos con información faltante. El cómo hacerlo en cada caso será una decisión de uds., que deberá ser tomada y **JUSTIFICADA** en base a las características de los datos analizados. Además de esto, una vez teniendo la base de datos completa, deberá contestar una serie de consultas con respecto a los datos, que requerirán el uso de técnicas de agregación, agrupación y visualización.

Descripción del problema

Debido a los problemas climáticos presentes en la tierra, muchos investigadores están sumamente preocupados por las reducciones en el hábitat de pingüinos. Para poder ayudarlos, es importante poder identificar las distintas especies y así brindarles la ayuda específica.

Por suerte, se ha hecho pública una base de datos que almacena características de pingüinos de diferentes razas. Lamentablemente existen algunos registros nulos, los que deberá corregir de la mejor forma posible para luego construir un modelo predictor de la raza.

La base de datos

La base da datos se encuentra disponible en el sitio del curso, en el archivo `penguins.csv`. Esta contiene información de pingüinos por medio de las siguientes columnas:

1. **species**: especie a la que pertenece el pingüino.
2. **island**: isla de procedencia del pingüino.
3. **culmen_length_mm**: largo de la parte superior del pico del pingüino.
4. **culmen_depth_mm**: profundidad de la parte superior del pico del pingüino.
5. **flipper_length_mm**: largo de la aleta del pingüino.
6. **body_mass_g**: masa del cuerpo del pingüino.
7. **sex**: sexo del pigüino.

Misiones

1. **Carga y exploración**: cargue el archivo con los datos y describa su contenido, indicando qué columnas tienen información incompleta. Finalmente, visualice las variables, con el fin de evaluar la existencia de *outliers*.
2. **Imputación y eliminación**: para cada una de las columnas con elementos faltantes, impute los valores en base a algún criterio basado en los datos. Además de esto, analice la posible eliminación de filas y columnas completas, en base a los valores faltantes y la relación entre las columnas.
3. **Consultas**: conteste cada una de las siguientes consultas, justificando los análisis y supuestos realizados:

   a) ¿Existe alguna diferencia importante entre pingüinos, dada por la isla de procedencia?

   b) ¿Cuáles son las 2 variables que mejor predicen el sexo de un pingüino?

c) Describa una agrupación de los pingüinos en base a los valores de las variables medidas, con el fin de realizar la clasificación de las especie de estos de la manera más precisa posible (es decir, que cada grupo contenga idealmente solo pingüinos de la misma especie).