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

IIC2115 – Programación como Herramienta para la Ingeniería (I/2023)

Ejercicio Capítulo 2b

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:** jueves 22 de septiembre a las 16:40 hrs. en sitio del curso.
- **Formato de entrega:** archivo Python Notebook (**C2b.ipynb**) con el avance logrado durante la sesión. El archivo debe estar ubicado en la carpeta **C2b**. Utilice múltiples celdas de texto y código para facilitar el trabajo del cuerpo docente.

Introducción

Con el fin de ejercitar los contenidos de modelos de machine learning en Python, en este ejercicio deberá aplicar lo aprendido en esta sesión y la anterior, tanto para completar datos faltantes como para realizar predicciones. 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.

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. **Predicción simple**: cargue los datos y elimine todas las filas con valores faltantes. Utilizando los datos resultantes, entrene 3 modelos distintos para predecir la especie de los pingüinos. Indique cuál de los tres es el mejor en base al rendimiento en un set independiente de test. Se recomienda revisar la página de scikit-learn y estudiar los algoritmos de clasificación que existen.
2. **Reemplazo de datos**: cargue nuevamente los datos, elija una columna y entrene 3 modelos distintos que permitan rellenar los valores faltantes en esta columna, en base al resto de los atributos.
3. **Reducción de dimensionalidad y visualización**: investigue sobre técnicas de reducción de dimensionalidad en scikit-learn y genere a partir de esto una visualización, que muestre cada pingüino como un punto en un espacio bidimensional y donde el color de cada punto este asociado a su especie. Se recomienda revisar las visualizaciones de matplotlib, como `scatter`. Utilizando esta visualización discuta sobre la dificultad del problema de clasificación y como se relaciona esto con los resultados del primer ítem.