



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

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:** lunes 11 de abril a las 16:30 hrs. en repositorio privado.
- **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 continuar y expandir lo realizado en el ejercicio de la sesión 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 de 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 pingüino.

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.
2. **Imputación y eliminación**: cargue nuevamente los datos, elija una columna y entrene 3 modelos distintos que permitan impute 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 sklearn 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. Utilizando esta visualización discuta sobre la dificultad del problema de clasificación y como se relaciona esto con los resultados del primer ítem.