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

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

Ejercicio Formativo 2 Capítulo 3

Aspectos generales

- **Objetivos:** Aplicar los contenidos de modelos predictivos de aprendizaje de máquina.
- **Lugar de entrega:** miércoles 20 de septiembre a las 23:59 hrs. en repositorio privado.
- **Formato de entrega:** archivo Python Notebook (**C3E2.ipynb**) con el avance logrado para el ejercicio. El archivo debe estar ubicado en la carpeta **C3**. Utilice múltiples celdas de texto y código para facilitar el trabajo del cuerpo docente.
- **ULTRA IMPORTANTE**: todas las celdas utilizadas deben estar ejecutadas al momento de entregar el ejercicio, de modo que las salidas generadas sean visibles. En caso de no cumplir esto, su entrega no será considerada como validación del ticket de salida.

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 `data_E2.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

Misión 1: completando información

Su primera misión será identificar los registros que presentan valores faltantes y completarlos utilizando modelos predictivos especializados en la categorías a las que pertenecen los registros. Por ejemplo, para completar los valores faltantes en las filas que describen pingüinos de la especie `Adelie`, deberá entrenar modelos con registros de pingüinos de esa especia.

Misión 2: predicción de la especie

Ya con los datos completos, su objetivo es entrenar modelos que permitan predecir la especie de un pingüino dadas sus características. En particular, deberá evaluar tres posibles estrategias para construir modelos:

- Predicción tradicional: entrenamiento de modelos para predecir directamente la raza de cada pingüino.
- Predicción jerárquica: entrenamiento de dos modelos para predecir la raza del pingüino. El primero debe discriminar entre 1 raza y las otras 2, mientras que el segundo debe discriminar entre las dos que formaron el mismo grupo para el modelo anterior. Qué raza usar para cada grupo y modelo es una decisión que debe tomar ud.

- Visualización: investigue sobre el uso de las técnicas de reducción de dimensionalidad PCA y tSNE (ambas disponibles en sklearn) y luego utilícelas para visualizar en 2D los datos. Finalmente, en base a lo observado, identifique subcategorías relevantes (subgrupos claros dentro de una especie de pingüino) y entrene modelos exclusivos para cada una de ellas.

En todos los esquemas puede elegir la familia de modelos que quiera. Finalmente, evalúe el rendimiento de los modelos en un conjunto de prueba.