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

IIC2115 – Programación como Herramienta para la Ingeniería (I/2024)

Ejercicio Formativo 1 Capítulo 3

Aspectos generales

- **Objetivos:** familiarizarse con los elementos fundamentales de modelos predictivos.
- **Lugar de entrega:** lunes 15 de abril a las 17:30 hrs. en repositorio privado.
- **Formato de entrega:** archivo Python Notebook (**C3E1.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. Para esto, se ha hecho pública una base de datos que almacena características de pingüinos de diferentes razas.

La base de datos

La base da datos se encuentra disponible en el sitio del curso, en el archivo `data_E1.csv`. Esta contiene información de pingüinos por medio de las siguientes columnas:

1. **species**: etiqueta de los datos, correspondiente a la 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

Abra el archivo y complete la información numérica faltante utilizando el valor promedio de cada columna. A continuación, descarte los registros para los cuales hay variables categóricas con valores faltantes.

Misión 2: preparación de los datos

Prepare los datos para su uso posterior en los modelos predictivos. Para esto, siga los pasos descritos en el material del curso, con el fin de evitar *data leakage*. Específicamente, debe realizar los siguientes pasos en orden:

1. Codificación numérica de variables categóricas.
2. Separación en conjuntos de entrenamiento y test.
3. Escalamiento de variables numéricas.

Misión 3: análisis exploratorio visual

Realice un análisis visual utilizando técnicas de reducción de dimensionalidad, con el fin de caracterizar los datos. Para esto, inicialmente reduzca la dimensionalidad del **espacio de características** a 2 o 3 dimensiones y grafíquelos con un *scatter plot*, utilizando la etiqueta de los datos como color de los marcadores. Analice distintos métodos de reducción de dimensionalidad, comentando sobre los grupos generados y la existencia de *outliers*.

Misión 4: análisis de clusters

Realice un proceso de clústering sobre los datos, primero sobre los datos en su espacio de características completo y luego sobre el espacio reducido generado en la misión anterior. Para ambos casos, obtenga el número óptimo de clusters con el método del codo.
Analice visualmente los resultados del clústering óptimo, incluyendo los centroides y los puntos pertenecientes a cada clúster, para ambos espacios de características. Finalmente, comente sobre cómo se relaciona la cantidad de clusters óptimo y los distintos valores que puede tomar la etiqueta, analizando la existencia de subgrupos, ya sea dentro de cada especie, o entre especies.

Misión 5: predicción de la especie

Ya con los datos completos y analizados, 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: en base al análisis realizado en las misiones anteriores, 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 test.