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

IIC2115 – Programación como Herramienta para la Ingeniería (II/2020)

Actividad Práctica 4 - Análisis exploratorio de datos

Objetivos

- Aplicar los contenidos de análisis exploratorio de datos para completar una base de datos incompleta y realizar predicciones sobre la misma.

Entrega

- **Lenguaje a utilizar:** Python 3.6 o superior
- **Lugar:** repositorio privado en GitHub. Recuerde incluir todo en una carpeta de nombre **P04**.
- **Entrega: jueves 5 noviembre a las 23:59 hrs.**
- **Formato de entrega:** archivo Python notebook (**P04.ipynb**) y archivo Python (**P04.py**) con la solución de los problemas. Ambos archivos deben estar ubicados en la carpeta **P04**. No se debe subir ningún otro archivo a la carpeta. En el archivo **.ipynb** utilice múltiples celdas de texto y código para facilitar la revisión de su programa.
- **NO SE ADMITEN ENTREGAS FUERA DE PLAZO**
- **Entregas con errores de sintaxis y/o que generen excepciones serán calificadas con nota 1.0.**
- Las *issues* del Syllabus en GitHub son parte de este enunciado.

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

M1 **Completando información (3.0 ptos.):** Su primera misión será identificar qué columna(s) presenta(n) pérdidas de información y completarla(s) utilizando algún criterio que respete la distribución de los datos de la(s) columna(s).

M2 **Predicción de la especie (3.0 ptos.):** Ya con los datos completos, su objetivo construir modelos que permitan predecir la especie de un pingüino dadas sus características. En particular, deberá evaluar dos 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.

En ambos esquemas, puede elegir la familia de modelos que quiera. Finalmente, evalúe el rendimiento de ambos modelos en base a su capacidad de generalización.

Corrección

La corrección de su actividad, se basará completamente en la lógica de los procedimientos utilizados. En particular, los procesos de imputación de datos, de entrenamiento de modelos y selección de estos deben estar adecuadamente justificados en base a los datos y a los criterios vistos en clase.

Política de Integridad Académica

"Como miembro de la comunidad de la Pontificia Universidad Católica de Chile me comprometo a respetar los principios y normativas que la rigen. Asimismo, prometo actuar con rectitud y honestidad en las relaciones con los demás integrantes de la comunidad y en la realización de todo trabajo, particularmente en aquellas actividades vinculadas a la docencia, el aprendizaje y la creación, difusión y transferencia del conocimiento. Además, velaré por la integridad de las personas y cuidaré los bienes de la Universidad."

En particular, se espera que mantengan altos estándares de honestidad académica. Cualquier acto deshonesto o fraude académico está prohibido; los alumnos que incurran en este tipo de acciones se exponen a un procedimiento sumario. Ejemplos de actos deshonestos son la copia, el uso de material o equipos no permitidos en las evaluaciones, el plagio, o la falsificación de identidad, entre otros. Específicamente, para los cursos del Departamento de Ciencia de la Computación, rige obligatoriamente la siguiente política de integridad académica en relación a copia y plagio: Todo trabajo presentado por un alumno (grupo) para los efectos de la evaluación de un curso debe ser hecho individualmente por el alumno (grupo), sin apoyo en material de terceros. Si un alumno (grupo) copia un trabajo, se le calificará con nota 1.0 en dicha evaluación y dependiendo de la gravedad de sus acciones podrá tener un 1.0 en todo ese ítem de evaluaciones o un 1.1 en el curso. Además, los antecedentes serán enviados a la Dirección de Docencia de la Escuela de Ingeniería para evaluar posteriores sanciones en conjunto con la Universidad, las que pueden incluir un procedimiento sumario. Por "copia" o "plagio" se entiende incluir en el trabajo presentado como propio, partes desarrolladas por otra persona. Está permitido usar material disponible públicamente, por ejemplo, libros o contenidos tomados de Internet, siempre y cuando se incluya la cita correspondiente.