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 2 Capítulo 3

Aspectos generales

- **Objetivos:** Aplicar los contenidos de modelos predictivos.
- **Lugar de entrega:** lunes 15 de abril a las 17:30 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

Considere el conjunto de datos almacenado en el archivo `data_E2.csv`, que contiene datos obtenidos a lo largo de los años sobre los niveles de Ozono (O_3) y material particulado de 2.5 micrómetros (PM2.5). Además de esta información, cada registro está categorizado en cuatro niveles, en base al riesgo ambiental que presentan las mediciones de O_3 y PM2.5 para la fecha: bajo, medio, alto y extremo. En base a toda esta información, complete las misiones indicadas a continuación.

IMPORTANTE

Recuerde codificar numéricamente los valores de las columnas categóricas (`Year`, `Month`, `Day`, `Environmental_risk`) y normalizar las numéricas (`O3` y `PM2.5`). Sea cuidadoso con el momento en que codifica y normaliza los valores, para evitar *data leakage*.

Misión 1: predicción de variables numéricas

Utilizando solo registros que no tengan valores faltantes para las columnas `Year`, `Month`, `Day`, `O3` y `PM2.5`, construya al menos dos modelos predictivos que permitan inferir el valor de la variable `PM2.5` en base a las otras variables recién indicadas. Evalúe el rendimiento de estos modelos en un set de prueba independiente, usando como métrica el *error cuadrático medio* (MSE), *error absoluto medio* (MAE) y *error porcentual absoluto medio* (MAPE). Finalmente, utilice el modelo con mejor rendimiento para completar los valores faltantes de la columna `PM2.5`, solo en aquellos registros que no tengan valores faltantes para las columnas `Year`, `Month`, `Day` y `O3`.

Misión 2: predicción de variables numéricas parte 2

Repita el procedimiento de la misión anterior, esta vez para completar los valores de la columna `O3`. Al finalizar este proceso, la base de datos solo debería tener valores faltantes para la columna `Environmental_risk`.

Misión 3: predicción de variables categóricas

Con la base de datos ya preparada, entrene al menos 2 clasificadores para predecir el valor de la variable `Environmental_risk`, a partir de todas las otras variables. Evalúe el rendimiento de estos modelos en un set de prueba independiente, usando como métrica el *balanced accuracy*. Finalmente, utilice el modelo con mejor rendimiento para completar los valores faltantes de la columna `Environmental_risk`.

Misión 4: comparación

Compare y comente los resultados obtenidos en la misión anterior con los de la Misión 4 del ejercicio C2E2. Indique cuáles parecen ser más adecuados, justificando sus argumentos.

Misión 5: análisis visual

Realice un análisis visual, utilizando técnicas de reducción de dimensionalidad y clústering, con el fin de caracterizar los datos. Comente sobre la separabilidad de los datos, existencia de subgrupos y *outliers*, entre otros.

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.