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

IIC2115 – Programación como Herramienta para la Ingeniería (I/2021)

Taller 4b

Objetivos

- Aplicar los contenidos de modelos predictivos basados en *Machine Learning*.

Entrega

- **Lenguaje a utilizar:** Python 3.6 o superior
- **Lugar:** repositorio privado en GitHub. Recuerde incluir todo en una carpeta de nombre **T4b**.
- **Entrega: lunes 7 de junio a las 16:50 hrs.**
- **Formato de entrega:** archivo python notebook (**T4b.ipynb**) y archivo python (**T4b.py**) con la solución de este enunciado. Los archivos deben estar ubicados en la carpeta **T4b**. No se debe subir ningún otro archivo a la carpeta. 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.**

Descripción del problema

Considere el conjunto de datos almacenado en el archivo `data.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 (antes o después de crear los conjuntos de entrenamiento y prueba).

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 recien indicadas. Evalúe el rendimiento de estos modelos en un set de prueba independiente, usando como métrica el *error cuadrático medio*. 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 4 del taller 4a con los de la misión anterior. Indique cuáles parecen ser más adecuados, justificando sus argumentos.

Corrección

La corrección de este taller se basará en lo adecuado de los mecanismos utilizados para realizar cada misión. En otra palabras, no existe *a priori* un resultado correcto para cada misión, por lo que cada misión se corregirá en base a lo adecuado y justificado que se encuentre el análisis. Cualquier supuesto que se haga para completar las misiones debe quedar claramente indicado.
ULTRA IMPORTANTE: todas las celdas utilizadas deben estar ejecutadas al momento de entregar el taller, de modo que las salidas generadas sean visibles. En caso de no cumplir esto, **SU TALLER NO SERÁ CORREGIDO**.

Objetivo de participación

Para verificar la participación durante la clase, debe completar la Misión 1.

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.