UC | Chile

Aprendizaje estadístico y computacional

Tarea: Comparación de Modelos

¡Bienvenido(a)!

Te invitamos a realizar el segundo trabajo.

- **Objetivo**: Comprender la diferencia entre modelos paramétricos y no paramétricos a través de casos prácticos desarrollados en Python.
- **Tipo de actividad**: Grupal
- **Tipo de evaluación**: Sumativa
- **Ponderación**: 15%
- **Puntaje**: 60 puntos
- **Calificación**: Escala de 1 a 7, con una exigencia de 50%. La nota mínima para aprobar es 4.0.

Evaluación

Descarga el instrumento de evaluación y revísalo antes de realizar la actividad.

Instrucciones

1. Antes de comenzar, debes haber revisado las siguientes clases y la lectura: videos, tutoriales y lecturas de la semana 2 a la semana 4.
2. Leer con atención los siguientes dos casos y responde según lo indicado.
3. Esta Tarea debe ser desarrollada completamente en lenguaje de programación Python, y estructurarse en formato de Notebook (seguir buenas prácticas de

escritura y programación, e incluir comentarios o celdas de markdown suficientes para explicar claramente todos los códigos computacionales).

4. Una vez finalizada la actividad, guarda un archivo con el nombre "Tarea1_ Apellidos_Integrantes", luego suba ambos archivos a la plataforma siguiendo las siguientes instrucciones:
    - Haz clic en el botón para agregar entrega. Se abrirá una nueva ventana que permite arrastrar el archivo y subirlo.
    - Comprueba que el archivo arrastrado es el correcto y presiona el botón para guardar cambios. El documento quedará guardado en la plataforma.

Enunciado

Introducción

Esta tarea se enfoca en abordar y profundizar los aspectos de la validación cruzada tanto en el contexto de regresión y como clasificación clasificación. Además, de comparar diversos métodos de aprendizaje. En el problema de regresión se compararán la regresión lineal, el k-NN, y los árboles de clasificación. Mientras que, para el problema de clasificación, se comparará la regresión logística, el K-NN, y los árboles de clasificación.

Descripción de los problemas

I. **Caso 1 (30 puntos):** Utilice el conjunto de datos "ozone" disponibles en la librería "faraway" (buscar en https://pypi.org/project/faraway/). Considere Y=O3 como variable respuesta, todas las demás serán variables explicativas.
    1. Realizar una descomposición aleatoria de la base de datos con la proporción 70%-30% para train y test, respectivamente.
    2. Seleccione alguna de las medidas de desempeño que pueda ser utilizada a este conjunto de datos. Indique el criterio utilizado.
    3. Utilizando la muestra de entrenamiento, junto con la validación cruzada k-fold (para algún k seleccionado por usted) y la medida de desempeño escogida, compare el

modelo de regresión lineal, el k-NN y un árbol de regresión. ¿Cuál de los métodos tiene el mejor resultado, según la validación cruzada k-fold?

4. Para el modelo seleccionado en el paso anterior, ajuste los parámetros con toda la muestra de entrenamiento y utilice la muestra test para medir la calidad del ajuste. Comente los resultados.

II. **Caso 2 (30 puntos):** Utilice el conjunto de datos “prostate” disponibles en la libreria ``faraway" (buscar en https://pypi.org/project/faraway/). Además, considere Y= svi como la variable respuesta, donde 1 es presencia y 0 ausencia. Todas las demás serán variables explicativas.

1. Realizar una descomposición aleatoria de la base de datos con la proporción 70%-30% para train y test, respectivamente.
2. De acuerdo con el contexto del conjunto de datos, seleccione alguna de las medidas de desempeño vistas en clase (Tasa de Error, Exactitud, Precisión, Sensibilidad, Especificidad o F_β-Score). Explique la elección de la medida de desempeño escogida.
3. Utilizando la muestra de entrenamiento, junto con la validación cruzada k-fold (para algún k seleccionado por usted) y la medida de desempeño escogida, compare los modelos de regresión logística, K-NN y árbol de clasificación. ¿Cuál de los métodos tiene el mejor resultado, según la validación cruzada k-fold?
4. Para el modelo seleccionado en el paso anterior, ajuste los parámetros con toda la muestra de entrenamiento y utilice la muestra test para medir la calidad del ajuste. Comente los resultados.

Aspectos formales

Considera los aspectos formales que se describen a continuación:

- Letra Arial 12 normal, interlineado simple.
- Extensión: Entre 1500 - 2000 palabras.

- Utilizar formato APA en citas al interior del texto y en la bibliografía.

¡Mucho éxito!

Importante: la fecha de entrega está indicada en el calendario del curso. Cuidar la redacción y la ortografía. Si tienes alguna duda sobre los contenidos o sobre cómo realizar esta actividad, puedes utilizar la herramienta "Mensajes" y enviar tu pregunta. Recibirás la respuesta de su tutor con las orientaciones correspondientes