UC | Chile

Aprendizaje estadístico y computacional

Tarea: Clasificación con Clases Desbalanceadas

¡Bienvenido(a)!

- **Objetivo**: Para un caso práctico desarrollado en Python, aplicar diferentes técnicas de clasificación a un conjunto de datos con clases desbalanceadas, donde también se aplicarán diferentes alternativas para balancear las clases.
- **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 6.
2. Leer con atención el siguiente caso 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 "Tarea3_ 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 clasificación cuando las clases están desbalanceadas, considerando especialmente el método de SVM para clasificar. Además, de comparar diversos métodos de aprendizaje cuando el conjunto de datos de entrenamiento se balancea y cuando no.

Descripción del problema

Utilice el conjunto de datos "BBDD_post_pabellón.xlsx" disponible en la plataforma coursera. Considere Y="Hospitalización" como variable respuesta (leer detalladamente la descripción de cada variables y de ser necesario buque información adicional).

1. Realizar una descomposición aleatoria de la base de datos, estratificada por la variable respuesta, con la proporción 70%-30% para train y test, respectivamente.
2. Seleccione una medida de desempeño, una de las técnicas de validación cruzada (simple, k-fold o leave one out) y tres técnicas de clasificación (obligatoriamente SVM y otras dos entre Naive Bayes, LDA, QDA, y Random Forest). Justifique cada elección.
3. Utilizando la muestra de entrenamiento, junto con la técnica de validación cruzada y la medida de desempeño escogidas, compare los métodos de clasificación. ¿Cuál de los

métodos tiene mejor rendimiento? En los métodos que calculan la probabilidad posterior, ¿cuál es el punto de corte óptimo?

4. Aplique dos técnicas de undersampling para balancear la muestra de entrenamiento, luego utilice la muestra balanceada para comparar los métodos de clasificación. ¿Cuál de los métodos tiene mejor rendimiento?
5. Aplique dos técnicas de oversampling para balancear la muestra de entrenamiento, luego utilice la muestra balanceada para comparar los métodos de clasificación. ¿Cuál de los métodos tiene mejor rendimiento?
6. Combine una técnica de undersampling y una de oversampling para balancear la muestra de entrenamiento, luego utilice la muestra balanceada para comparar los métodos de clasificación. ¿Cuál de los métodos tiene mejor rendimiento?
7. ¿Cuál de las estrategias anteriores (ítems 3, 4, 5, y 6) es la mejor? Para aquella estrategia ajuste los parámetros con toda la muestra de entrenamiento y utilice la muestra test para medir la calidad del ajuste. Comente sobre los resultados y responda la pregunta: ¿Es posible identificar los factores más importantes?

Aspectos formales

- Entrega un jupyter notebook (.ipynb)

¡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