TD VI: Inteligencia Artificial

Trabajo práctico 1 (2024 2do semestre)

Objetivo

El objetivo de este trabajo práctico es que realicen un análisis completo utilizando árboles de decisión en R. El trabajo se presentará en formato R Markdown, integrando código, resultados y explicaciones en un único documento.

Hemos elegido R como lenguaje de programación para este TP por una razón fundamental: queremos que adquieran intuiciones sobre el impacto que tiene en la performance de los modelos el hecho de que el mismo árbol maneje los valores faltantes (NAs). Esta característica está implementada en R, específicamente en la librería **rpart**, pero no está disponible en las implementaciones populares de Python.

El trabajo práctico consiste en generar un archivo R Markdown (.Rmd) que incluya todo el código, análisis y explicaciones hechas por ustedes. A continuación, se detalla la estructura que debe tener este archivo.

Estructura del R Markdown a crear

1. Introducción al problema (5 puntos)

En la primera sección del R Markdown deberán presentar el conjunto de datos con el que trabajarán. La elección del conjunto de datos es libre, pero debe cumplir con las siguientes características:

- Debe abordar un problema de **clasificación binaria**. Noten que ustedes pueden construir una variable binaria a partir de una variable numérica (e.g., creando una variable que indique si el valor de la variable numérica es mayor o igual a algún otro valor fijo). También, a partir de una variable categórica con más de dos valores se puede construir una variable binaria (e.g., preguntando si el valor es igual a una categoría dada o no). Estas modificaciones pueden ser hechas por ustedes en la siguiente sección del R markdown.
- Debe incluir tanto atributos predictores numéricos como categóricos.
- Dentro del conjunto de datos, los atributos categóricos no deben estar representados como números. Si algunos atributos categóricos estuvieran codificados como números enteros, en la siguiente sección, deberán transformarlos para que aparezcan como palabras.
- El conjunto de datos no debe tener menos de 10.000 observaciones y debe tener un número razonable de predictores (ni muy pocos, ni un número excesivo, entre 12 y 200 es un número razonable). Si optan por un conjunto con más de 50.000 observaciones, realicen un muestreo aleatorio para reducirlo a uno de 50.000 observaciones, asegurándose de utilizar una semilla para replicar sus operaciones.[1] Tengan presente que se pedirá que entreguen este archivo reducido en lugar del original.
- La variable objetivo debe requerir un modelo complejo para ser predicha, no siendo suficiente una regla simple.

En esta sección deberán:

- Presentar una descripción concisa del conjunto de datos, incluyendo su origen, variables principales y el problema a resolver.
- Justificar la elección del conjunto de datos para el uso de árboles de decisión.

Noten que en esta sección no deben presentar código R.

[1] https://r-coder.com/set-seed-r/

En caso de no saber por dónde empezar a buscar el dataset adicional, a continuación se brindan enlaces a algunas páginas que proveen acceso a conjuntos de datos públicos.

- https://data.buenosaires.gob.ar/dataset/
- https://datos.gob.ar/
- https://www.gapminder.org/data/
- https://github.com/owid
- https://dataverse.harvard.edu/
- https://www.kaggle.com/datasets
- https://archive.ics.uci.edu/

2. Preparación de los datos (10 puntos)

En esta sección deberán:

- Cargar el conjunto de datos y realizar el preprocesamiento necesario.
- Incluir un análisis exploratorio de datos que contenga:
  - Estadísticas descriptivas de las variables principales.
  - Al menos dos visualizaciones relevantes.
- Comentar sobre las características observadas en los datos.

3. Construcción de un árbol de decisión básico (10 puntos)

En esta sección deberán:

- Dividir el conjunto de datos al azar en tres particiones: entrenamiento (70%), validación (15%) y testeo (15%). Utilicen una semilla para asegurar la replicabilidad.
- Construir un árbol de decisión básico usando **rpart** con los parámetros por defecto.[2] Indiquen cuáles son los valores de los hiperparámetros que se utilizan por defecto y qué implica que cada hiperparámetro tenga dicho valor.[3]
- Visualizar el árbol resultante. En caso de no poder generar una visualización razonable, comenten por qué sucede esto.
- Explicar la estructura general del árbol obtenido.Enfocándose en los primeros cortes del mismo.

4. Evaluación del árbol de decisión básico (15 puntos)

En esta sección deberán:

- Realizar predicciones en el conjunto de testeo, calculando tanto las probabilidades predichas como las clase predicha (la que se obtiene según la clase mayoritaria en cada hoja - i.e., lo que hace por defecto rpartt).
- Calcular y presentar las siguientes métricas de performance:[4]
  - Matriz de confusión
  - Accuracy
  - Precision y recall
  - F1-score
  - AUC-ROC
- Interpretar los resultados obtenidos

.

[2] Aquí pueden ver un mini tutorial de cómo utilizar esta librería: https://rubenfcasal.github.io/aprendizaje_estadistico/cart-con-el-paquete-rpart.html.
[3] Vean: https://www.rdocumentation.org/packages/rpart/versions/4.1.23/topics/rpart.control
[4] Vean: https://cran.r-project.org/web/packages/MLmetrics/MLmetrics.pdf

5. Optimización del modelo (20 puntos)

En esta sección deberán:

- Experimentar con diferentes valores de los hiperparámetros maxdepth, minsplit y minbucket, y evaluar el rendimiento de cada árbol probado en el conjunto de **validación**. Sean exhaustivos en la búsqueda. Se pide que midan la performance únicamente mediante AUC-ROC.

  IMPORTANTE: al hacer esta prueba, deberán fijar los valores de los hiperparámetros **cp** y **xval** en 0. Hacer esto les permitirá construir árboles de profundidad máxima dados los valores de maxdepth, minsplit, y minbucket utilizados.

- Realizar una o más visualizaciones que permitan ver cómo se relacionan los valores de estos hiperparámetros con la performance obtenida en validación.
- Elegir el árbol con mejor rendimiento y medir el valor de AUC-ROC en el conjunto de **testeo**.
- Comparar el rendimiento en el conjunto de testeo del árbol básico (del punto 3) con el árbol optimizado.

Se recomienda que implementen el código de esta sección de forma tal que pueda ser reusado en la sección 7 de este trabajo práctico.

6. Interpretación de resultados (10 puntos)

En esta sección deberán:

- Presentar y explicar el árbol optimizado final, indicando diferencias respecto al presentado en la sección 3.
- Identificar y discutir las variables más importantes según el árbol de decisión optimizado.

7. Análisis del impacto de los valores faltantes (25 puntos)

En esta sección deberán:

- Generar tres nuevos sets de conjuntos de datos, donde:
  - En el primer set, el 20% de los valores de cada variable predictora se reemplazará al azar por NA (tanto en entrenamiento, validación y testeo).
  - En el segundo set, el 50% de los valores de cada variable predictora se reemplazará al azar por NA (tanto en entrenamiento, validación y testeo).
  - En el tercer set, el 75% de los valores de cada variable predictora se reemplazará al azar por NA (tanto en entrenamiento, validación y testeo).

  IMPORTANTE: En caso que en el set original (aquél sin agregar NAs) se tenga que una variable predictora ya tiene al menos la proporción de valores missings solicitada, no se debe modificar dicha misma. A su vez, las **observaciones** que pertenecen a cada versión del conjunto de entrenamiento, validación y testeo, deben ser las mismas.

- Para cada set de conjuntos de datos creado (20% NA, 50% NA, 75% NA):
  - Deberán construir un nuevo árbol de decisión optimizado, buscando los valores de los hiperparámetros (maxdepth, minsplit y minbucket - con cp y xval igual a 0) que maximizan el ROC-AUC de **validación**.
  - Deberán comparar el rendimiento en **testeo** del mejor árbol obtenido con la performance obtenida por el árbol optimizado del punto 5.
- Analizar y discutir cómo cambia el rendimiento del modelo a medida que aumenta el porcentaje de valores faltantes en las variables predictoras. Para este análisis, pueden (y se sugiere) hacer uso de uno o más gráficos.

8. Conclusiones y discusión (5 puntos)

En esta sección deberán:

- Resumir los hallazgos principales del análisis.
- Reflexionar sobre la efectividad del árbol de decisión para el problema planteado.
- Sugerir posibles mejoras o direcciones futuras para el análisis.

Criterios de evaluación

Cada una de las secciones del R Markdown se evaluarán de acuerdo a los siguientes criterios:

- Claridad y profundidad de las explicaciones e interpretaciones
- Correcta implementación del análisis
- Calidad del código R y uso efectivo de R Markdown

Modalidad de entrega

- Tal como ya dijimos, el trabajo se debe realizar en **grupos de tres personas**.
- La fecha límite de entrega es el **domingo 1 de septiembre a las 23:59 h**.
- Para realizar sus consultas sobre este TP, también cuentan con el foro llamado **TP1 | Consultas** en la página de la materia en el Campus Virtual. Todas las dudas que surjan en relación al TP1 envíenlas exclusivamente a este foro; no usen ningún otro. Esto debe ser así porque este es el foro que está configurado de forma que los mensajes enviados lleguen únicamente al cuerpo docente y a sus compañeros de grupo. En otras palabras, si un integrante de un grupo envía una pregunta por acá, tanto esa pregunta como la respuesta, luego dada por el cuerpo docente, podrán ser vistas sólo por los integrantes del grupo en cuestión.
- **Únicamente 1** integrante del grupo debe realizar la **entrega**.
- Debe entregarse un archivo zip que dentro tenga dos archivos: un archivo R markdown con las secciones solicitadas completadas y el conjunto de datos utilizado. Si el conjunto de datos no fue muestreado, se pedirá que entreguen el conjunto de datos original. Si hubiera sido muestreado, se pide que entreguen la muestra utilizada en el análisis. La versión ZIP de esta carpeta debe subirse a la tarea llamada **TP1 | Entrega** en la página de la materia en el Campus Virtual. (Cuidado con los envíos a último minuto que la tarea se cierra a las 23:59:00).