Investigación de Operaciones
Informe Árboles de Clasificación

Profesores: Daniel Quinteros - Nicolás Rojas M.
Ayudantes: Diana Gil S. - Joaquín Gatica H.

Primer Semestre 2023

1. **Objetivo**

   El objetivo de este trabajo es utilizar una herramienta para construir árboles de clasificación, evaluando el efecto de utilizar diferentes conjuntos de datos, atributos y tipos de atributos. Además, permitirá conocer y aprender a utilizar parte del software R.

2. **Enunciado**

   La detección de spam es un problema importante en la industria de la tecnología, ya que el correo electrónico no deseado representa una gran cantidad de tráfico en Internet y puede ser perjudicial para los usuarios. Se han utilizado diferentes técnicas de clasificación de datos, entre ellas árboles de clasificación, para identificar el spam y filtrarlo de manera efectiva[1]. En esta tarea, se utilizarán árboles de clasificación para desarrollar un modelo de detección de spam en correo electrónico. Se proporcionará un conjunto de datos de correo electrónico etiquetados como spam o no spam, el cual debe ser utilizado para construir/entrenar un árbol que sea capaz de discriminar esta situación.

   Utilizando el software **R**, el paquete **tree**[2] y **seteando la semilla con el número de su rol, sin guión**[3], conteste las siguientes preguntas:

   1. Describa el conjunto de datos: Cantidad de datos, Tipo y Valores posibles de cada atributo[4], Indique la cantidad de registros por clase.
   2. Convierta a *factor* el atributo *target* ('*Prediction*'[5]) y el atributo '*attachment_size*', elimine los atributos que no se utilizarán (los que fueron convertidos), compruebe sus modificaciones.
   3. Genere e imprima un árbol considerando el 70 % de los datos entregados para *training*, utilice como criterio de división **gini** y **deviance**, luego conteste lo siguiente:
      - Describa el árbol obtenido: cantidad de niveles (profundidad) y de hojas del árbol.
      - Evalúe el árbol utilizando la métrica *accuracy*[6], imprima dicha métrica y comente acerca del desempeño del modelo.
   4. ¿Qué sucede si utiliza un 30 %, 50 % y 70 % de los datos entregados como *training*? Genere, imprima y evalúe los árboles obtenidos (utilice *accuracy*). Compare los resultados e indique qué árbol es el mejor.
   5. Se desea mejorar el rendimiento temporal del árbol, sin disminuir su efectividad, considerando esta vez el 80 % de los datos para *training*. Construya, imprima y evalúe los siguientes árboles, utilizando los siguientes atributos para cada caso:

[1] `https://www.semanticscholar.org/paper/Designing-Spam-Model-Classification-Analysis-using-Rajput-Arora/9984a7d06e04a347718cb8c7f645b72195bb11ce`
[2] `https://cran.r-project.org/web/packages/tree/index.html`
[3] Por ejemplo `set.seed(2023001001)`
[4] Si su descripción aplica para varios atributos puede mencionarlo, para no repetirla.
[5] Tener precaución con el atributo '*predictions*', que es distinto y se refiere a la frecuencia de dicha palabra.
[6] Casos correctamente clasificados / total de casos evaluados.

a) Solo atributo '*thanks*'
b) Atributos '*thanks*' y '*subject*'
c) '*thanks*', '*subject*' y '*attachment_size*'
d) Todos los atributos

Compárelos, indique cuál es el mejor y por qué.

3. **Requerimientos e Informaciones**

- En cada pregunta, mencione qué comandos y funciones utilizó.
- **El trabajo es individual**.
- Los informes deben ser subidos a plataforma aula **con plazo máximo** el día 5 de Junio hasta las 23:59 horas. **Informes entregados fuera de plazo serán calificados con nota** 0.
- Debe subir una carpeta, que tenga por nombre su ROL, comprimida en .zip que contenga: (1) un archivo jupyter notebook con el código y (2) un pdf con el informe.
- El software R está disponible en `https://www.r-project.org/`
- Es obligación detallar y fundamentar adecuadamente cada respuesta, así como comentar el código que escriba.
- Debe incluir imágenes de los árboles en toda pregunta que se solicite.
- **Los datos del trabajo están disponibles en aula, en la unidad correspondiente**. Considere como *hint* la siguiente descripción del *dataset*:
  - La columna que indica si un caso es o no *spam* se llama *Prediction*.
  - Existe un atributo llamado '*attachment_size*' que indica el tamaño del archivo adjunto en el correo, en caso que posea.
  - Las demás columnas representan la frecuencia de apariciones de cada palabra en correos electrónicos. Por ejemplo: Columna '*subject*' con valor 10, indica que dicha palabra aparece 10 veces en ese correo.

Cualquier duda relacionada sobre el enunciado o con los contenidos, favor publicar en aula o escribir un mail directo a sus ayudantes.