Investigación de Operaciones
Control 5: Árboles de Clasificación

Profesor: Nicolás Rojas M.
Ayudantes: Francisca Ramirez E. - Felipe Vega V.

Instrucciones:

- El escrito debe ser realizado en LaTeXo en Word
- Todo árbol debe ser incluído en el escrito (no adjuntar archivos extra).
- JUSTIFIQUE TODAS sus respuestas. En caso de no presentar análisis y justificación escrita en una pregunta, su puntaje será de cero puntos (aunque adjunte un árbol relacionado a la pregunta).
- Una vez realizada la evaluación, subir el escrito en formato '.pdf'.

1. **Diagnosticando diabetes** (100 puntos)

   El Ministerio de Salud necesita realizar un estudio enfocado en pacientes de Diabetes. El objetivo es predecir, considerando ciertas características de los pacientes, si es que un paciente tiene diabetes. Para este estudio, el Ministerio nos entrega un dataset llamado Pima Indians Diabetes, que considera a pacientes mujeres de al menos 21 años. Los atributos a considerar son los siguientes:

   1. Cantidad de veces embarazada
   2. Concentración de glucosa en plasma (luego de 2 horas de tomar test de tolerancia)
   3. Presión sanguínea diastólica (mm Hg)
   4. Grosor pliegue tríceps (mm)
   5. Suero de insulina a las 2 horas (mu $\frac{U}{ml}$)
   6. IMC ($(\frac{kg}{m})^2$)
   7. Función pedigree diabetes
   8. Edad (años)
   9. Variable objetivo: Paciente tiene diabetes (0 o 1)

   Utilizando el software **R**, conteste las siguientes preguntas:

   1. (40 puntos) Generar un árbol, sin utilizar poda y considerando el 70 % de los datos entregados para Training. Conteste lo siguiente:
      - Describa el árbol obtenido: cantidad de niveles obtenido, cantidad de hojas del árbol obtenido, entre otros atributos.
      - ¿Qué variables no generan particiones?, explique porqué éstas variables no participan en el árbol generado.
   2. (50 puntos) Nuevamente, sin utilizar poda y considerando el 70 % de los datos entregados para Training, ¿Qué sucede si las variables Glucosa e IMC no son consideradas en el modelo? Describa detalladamente:
      - Compare con el árbol obtenido en la pregunta (1) considerando cantidad de niveles del árbol, cantidad de hojas del árbol obtenido, entre otros atributos.
      - Considerando el Testing set, ¿Cuál es la precisión del árbol?
   3. (10 puntos) Considerando el árbol de la pregunta (2), ¿ Cual sería el diagnóstico asignado al siguiente paciente: (2,30,71,26,5.3, 27, 0.33, 40).

   NOTA: Datos disponibles en Aula. El escrito debe contener el código (texto, no imágenes) para obtener cada resultado.