Profesor: Dr. Oldemar Rodríguez Rojas
CA-0404 Modelos Lineales
Métodos Basados en Árboles
Fecha de Entrega: 30 de septiembre
Hora de entrega: 1 pm

Tarea Número 4

- Las tareas son estrictamente de carácter individual, tareas idénticas se les asignará cero puntos.
- Todas las tareas tienen el mismo valor en la nota final del curso.
- Todos los ejercicios tienen el mismo valor.
- **Pregunta 1:** [10 puntos] En este ejercicio vamos a usar la tabla de datos `wine.csv`, que contiene variantes del vino "Vinho Verde". Los datos incluyen variables de pruebas fisicoquímicas y sensoriales realizadas a dicho vino.

  La tabla contiene 1599 filas y 12 columnas, las cuales se explican a continuación.

  - `fija.acidez`: Acidez fija, ácidos presentes después de la destilación.
  - `volatil.acidez`: Acidez volátil, esta determina si el vino tendrá un sabor avinagrado.
  - `citrica.acidez`: Acidez cítrica, esta es propiamente de las uvas y no por fermentación.
  - `residual.azucar`: Azúcar residual del vino que no fue fermentada.
  - `cloruros`: Cloruros, uno de los principales componentes salinos del vino.
  - `libre.sulfuro.dioxido`: Dióxido de azufre libre, antimicrobiano natural.
  - `total.sulfuro.dioxido`: Dióxido de azufre total, suma del dióxido libre y combinado.
  - `densidad`: Densidad.
  - `pH`: Potencial de hidrógeno, este reduce la sensación de acidez en el vino.
  - `sulfitos`: Sulfitos, conservan los aromas y actúa como desinfectante.
  - `alcohol`: Alcohol presente.
  - `calidad`: Calidad del vino.
  - `tipo`: Tinto o Blanco (variable a predecir).

  Para esto realice lo siguiente:

  1. Cargue la tabla de datos `wine.csv` en `R`.
  2. Usando el comando `sample` de **R** genere al azar una tabla de testing con una 20 % de los datos y con el resto de los datos genere una tabla de aprendizaje.
  3. Usando árboles de Decisión (con `rpart`) genere un modelo predictivo para la tabla de aprendizaje, grafique el árbol obtenido. Pruebe modificar los parámetros del método hasta encontrar el que minimiza el error global.

4. Genere un modelos predictivos para la tabla de aprendizaje usando Bosques Aleatorios ADABoosting y XGBoosting. Pruebe modificar los parámetros del método hasta encontrar el que minimiza el error global.
5. Construya un `DataFrame` de manera que en cada una de las filas aparezca un modelo predictivo y en las columnas aparezcan los índices *Precisión Global, Error Global Precisión Positiva (PP), Precisión Negativa (PN), Falsos Positivos (FP), los Falsos Negativos (FN), la Asertividad Positiva (AP) y la Asertividad Negativa (AN)*. ¿Cuál de los modelos es mejor para estos datos? (incluya todos los métodos que hemos estudiando en el curso).

- **Pregunta 2:** [10 puntos] Suponga que somos contratados por el banco y se nos pide volver a predecir el monto promedio de deuda en tarjeta de crédito de una cartera de clientes relativamente nuevos, basado en otra cartera de comportamiento y estructura similar de la cual sí se tiene información de deuda en tarjeta de crédito. En este ejercicio hacemos uso de la tabla de datos `DeudaCredito.csv` que contiene información de los clientes en una de las principales carteras de crédito del banco, e incluye variables que describen cada cliente tanto dentro del banco como fuera de éste.

  Esta tabla de datos contiene 400 clientes y 11 variables que los describen. Seguidamente se explican las variables que conforman la tabla.

  - `Ingreso`: Ingreso del cliente, en miles de dólares.
  - `Limite`: Límite de crédito global en tarjetas de crédito del cliente.
  - `CalifCredit`: Calificación crediticia del cliente.
  - `Tarjetas`: Cantidad de tarjetas de crédito del cliente.
  - `Edad`: Edad del cliente.
  - `Educacion`: Años de educación del cliente.
  - `Genero`: Género del cliente.
  - `Estudiante`: Indica si el cliente es estudiante o no.
  - `Casado`: Indica si el cliente es casado o no (1 = Sí, 0 = No).
  - `Etnicidad`: Indica si el cliente es caucásico, afroamericano o asiático.
  - `Balance`: Monto promedio de deuda en tarjeta de crédito del cliente, en dólares.

  Cargue la tabla de datos en `R`, asegúrese que las variables se están leyendo de forma correcta. Recodifique variables en caso de que sea necesario, tome para entrenamiento un 80 % de la tabla de datos. Realice lo siguiente:

  1. Genere Modelos de Regresión usando KNN, Regresión Múltiple, Ridge, Lasso, Árboles, Bosques Aleatorios y Potenciación incluyendo las todas las variables predictoras ¿Qué error se obtiene sobre la tabla de testing para estos modelos? ¿Cuál considera que es un mejor modelo para predecir la deuda en tarjeta de crédito? Justifique. Interprete las medidas de error obtenidas.
  2. ¿Qué observa en los gráficos de dispersión que muestra los valores reales contra la predicción de cada modelo? ¿Qué desventajas o ventajas puede observar en cada modelo? Explique.

3. Muestre e interprete la mejor regla que genera el modelo de Árboles de Regresión. Desde su punto de vista ¿Le ve sentido a esta regla? ¿Esto es bueno o malo?

- **Pregunta 3:** [20 puntos] Un cliente nos contrata esta vez para aplicar un modelo no paramétrico con el fin de estudiar una posible oportunidad de negocio, y para ver si le es rentable quiere una predicción de las ventas potenciales de asientos de niños para autos en su tienda. Para ello nos proporciona la tabla de datos `AsientosNinno.csv` que contiene detalles de ventas de asientos de niños para auto en una serie de tiendas similares a las del cliente, y además los datos incluyen variables que definen características de la tienda y su localidad. La tabla de datos está formada por 400 filas y 13 columnas. Seguidamente se explican las variables que conforman la tabla.

  - `Ventas`: Ventas de asientos de niños para autos en cada localidad, en miles de unidades.
  - `PrecioCompt`: Precio promedio por asiento de niño cobrado por la competencia en cada localidad.
  - `Ingreso`: Nivel de ingreso promedio de los habitantes de la región, en miles de dólares.
  - `CercaniaEsc`: Índice que indica que tan cercana está la tienda a zonas escolares.
  - `Publicidad`: Presupuesto que asigna cada tienda a publicidad, en miles de dólares.
  - `Poblacion`: Tamaño de la población en cada región, en miles.
  - `Precio`: Precio cobrado por la tienda por los asientos de niño para auto.
  - `CalidadEstant`: Indica la calidad de ubicación de los asientos de niño en los estantes de la tienda.
  - `Edad`: Edad promedio de los habitantes de la localidad.
  - `Educacion`: Años de aducación promedio de los habitantes de cada región.
  - `Urbano`: Indica si la tienda está localizada en una zona urbana o no (1 = Sí, 0 = No).
  - `USA`: Indica si la tienda está ubicada en Estados Unidos o no (1 = Sí, 0 = No).
  - `Desarrollo`: Índice de desarrollo de cada localidad.

  Cargue la tabla de datos en `R` y no elimine los NA. En caso de ser necesario, recodificar las variables de forma adecuada. Para medir el error tome un 20 % de la tabla de datos. Realice lo siguiente:

  1. Corra un modelo de Árboles de Regresión incluyendo las variables predictoras adecuadas. Muestre e interprete alguna de las reglas obtenidas. Por último interprete las medidas de error.
  2. Ahora genere un modelo de Bosques Aleatorios usando 200 árboles y todas las variables predictoras. Cuáles son las 3 variables más importantes (basado en disminución del RSS). Por último interprete las medidas de error.
  3. Ahora genere un modelo de Potenciación usando 200 árboles, pruebe con las diferentes opciones de distribución y escoja la mejor (`distribution = gaussian, laplace, bernoulli, adaboost, poisson,...`) y todas las variables predictoras. Cuáles son las 3 variables más importantes del mejor modelo. Por último interprete las medidas de error.

4. Prefiere usar un modelo con bajas medidas de error pero poco interpretable o uno con medidas de error un poco mayores pero que es más interpretable ¿Cuál de los modelos de los incisos anteriores le dio mejores resultados en la tabla de testing? ¿Cuál modelo prefiere de los tres? Justifique sus respuestas.

- **Pregunta 4:** [20 puntos] [**no usar** `R`] Considere los datos de entrenamiento que se muestran en la siguiente Tabla para un problema de clasificación binaria.

ID Cliente	Género	Tipo-Carro	Talla	Clase
1	M	Familiar	Pequeño	C0
2	M	Deportivo	Mediano	C0
3	M	Deportivo	Mediano	C0
4	M	Deportivo	Grande	C0
5	M	Deportivo	Extra Grande	C0
6	M	Deportivo	Extra Grande	C0
7	F	Deportivo	Pequeño	C0
8	F	Deportivo	Pequeño	C0
9	F	Deportivo	Mediano	C0
10	F	Lujo	Grande	C0
11	M	Familiar	Grande	C1
12	M	Familiar	Extra Grande	C1
13	M	Familiar	Mediano	C1
14	M	Lujo	Extra Grande	C1
15	F	Lujo	Pequeño	C1
16	F	Lujo	Pequeño	C1
17	F	Lujo	Mediano	C1
18	F	Lujo	Mediano	C1
19	F	Lujo	Mediano	C1
20	F	Lujo	Grande	C1

1. Calcule el índice de Gini para la tabla completa, observe que el 50 % de las filas son de la clase C0 y el 50 % son de la clase C1.
2. Calcule el índice de Gini Split para la variable Género.
3. Calcule el índice de Gini Split para la variable Tipo-Carro.
4. Calcule el índice de Gini Split para la variable Talla.
5. ¿Cuál variable es mejor Género, Tipo-Carro o Talla?

- **Pregunta 5:** [20 puntos] [**no usar** `rpart`] Supongamos que tenemos un árbol de decisión con tres clases A, B, C. Se tiene que decidir cómo dividir el nodo padre:

$$N = \begin{pmatrix} A & 100 \\ B & 50 \\ C & 60 \end{pmatrix}$$

para esto hay dos posibles divisiones. La primera posible división N_1 divide el nodo N en los 2 siguientes nodos:

$$N_{1,1} = \begin{pmatrix} A & 62 \\ B & 8 \\ C & 0 \end{pmatrix}, \quad N_{1,2} = \begin{pmatrix} A & 38 \\ B & 42 \\ C & 60 \end{pmatrix}.$$

La segunda opción de división N_2 para el nodo N es la siguiente en 3 nodos:

$$N_{2,1} = \begin{pmatrix} A & 65 \\ B & 20 \\ C & 0 \end{pmatrix}, \quad N_{2,2} = \begin{pmatrix} A & 21 \\ B & 19 \\ C & 20 \end{pmatrix}, \quad N_{2,3} = \begin{pmatrix} A & 14 \\ B & 11 \\ C & 40 \end{pmatrix}.$$

1. Calcule la información ganada usando el índice de Gini para las dos posibles divisiones (N_1 en 2 nodos y N_2 en 3 nodos). ¿Cuál división es la mejor?
2. Repita 1. usando el criterio de la Entropía.
3. Repita 1. usando el criterio del Error Clasificación.
4. Otro índice utilizado para decidir cuál división es la mejor es la Complejidad (que también es utilizado como criterio en la poda de un árbol). Dado un índice Q que puede Gini, la Entropía o el Error Clasificación, se define la Complejidad de un árbol T con nodos terminales $(t_j)_{1\leq j\leq m}$ (m = cantidad de nodos terminales) como sigue:

   $$C_\alpha(T) = \sum_{j=1}^{m} n_j Q(t_j) + \alpha m,$$

   con $\alpha \geq 0$ y donde n_j es la cardinalidad del nodo t_j.
   Dado un árbol grande T_L y si $T \leq T_L$ denota que T es un subárbol de T_L, entonces se define el Árbol Óptimo como sigue:

   $$\hat{T}_\alpha = \min_{T\leq T_L} C_\alpha(T).$$

   Si usamos como parámetro para la complejidad $\alpha = 25$ para cada nodo terminal y usando el índice de Gini para las dos posibles divisiones (N_1 en 2 nodos o N_2 en 3 nodos). ¿Cuál división es la mejor en el sentido de que minimiza la complejidad? ¿Para que valores de α este criterio prefiere N_1 sobre N_2?
5. Repita 4. usando el criterio de la Entropía.
6. Repita 4. usando el criterio del Error Clasificación.

- **Pregunta 6:** [20 puntos] Considere las siguientes definiciones:

  Definición 1. • *Una función* $\phi : [0,1]^r \longrightarrow [0,\infty[$ *se llama* **Función de Impureza** *si tiene las siguientes propiedades:*

  1. *Para cualquier* $i \in \{1,\ldots,r\}$, $\phi(e_i) = min\{\phi(x_1,\ldots,x_r) \mid \sum_{i=1}^r x_i = 1\}$, *donde* e_i *es el vector que tiene un* 1 *en la celda* i *y* 0 *en las restantes.*
  2. $\phi(\frac{1}{r},\ldots,\frac{1}{r}) = max\{\phi(x_1,\ldots,x_r) \mid \sum_{i=1}^r x_i = 1\}$.
  3. ϕ *es simétrica, esto es, para cualquier permutación* σ *de* r *letras:*

     $$\phi(x_1,\ldots,x_r) = \phi(x_{\sigma(1)},\ldots,x_{\sigma(r)}).$$

Definición 2. • *Sea* $p(s|v) = \frac{|E_s \cap v|}{|v|}$ *la probabilidad del grupo a priori* E_s *en el nodo* v*;* $s = 1, \ldots, r$*. La impureza de* v *es,*

$$Imp(v) = \phi(p(1|v), \ldots, p(r|v)).$$

donde ϕ *es una función de impureza.*

- *El nodo* v *es puro si* $Imp(v) = 0$*.*

Definición 3. *El descenso de la impureza (información ganada)* $\Delta Imp(v)$*, obtenido a consecuencia de la división del nodo* v *en* v_i *y* v_d *se define como:*

$$\Delta Imp(v) = Imp(v) - \left[p(v_i)Imp(v_i) + p(v_d)Imp(v_d)\right].$$

donde $p(v_i) = \frac{|v_i|}{|v|}$ *y* $p(v_d) = \frac{|v_d|}{|v|}$*.*

Pruebe lo siguiente:

1. Si la función de impureza ϕ es estrictamente cóncava[1] entonces $\Delta\text{Imp}(v) \geq 0$ y $\Delta\text{Imp}(v) = 0$ si y solo si $p(s|v) = p(s|v_i) = p(s|v_d)$ para $s = 1, \ldots, r$.
2. Sea la función $g : [0,1]^r \longrightarrow [0, \infty[$ definida por $g(x_1, \ldots, x_r) = \sum_{i \neq j}^{r} x_i x_j$, con $\sum_{i=1}^{r} x_i = 1$.

   Entonces la función g tiene las siguientes propiedades:

   a) $g(x_1, \ldots, x_r) = 1 - \sum_{i=1}^{r} x_i^2$.

   b) g es estrictamente cóncava.

   c) g es una función de impureza.

Entregables: Debe entregar un documento autreproducible HTML con todos los códigos y salidas. No olvide poner un título para cada pregunta. Las demostraciones las puede entregar en papel a mano y enviarlas escaneadas.

[1] Sea D un conjunto convexo de $\mathbb{R}^n$ y $\alpha, \beta \in [0,1]$, con $\alpha + \beta = 1$. Una función $f : D \longrightarrow \mathbb{R}$ es estrictamente cóncava si $\forall\ x, y \in D$, $\alpha f(x) + \beta f(y) \leq f(\alpha x + \beta y)$ con igualdad si y solo si $x = y$.