6-4-2024

Tarea: Métodos de Consenso y Potenciación

Mineria de Datos

EDWIN DANIEL PROAÑO ZAPATA
UNIVERSIDAD CENTRAL DEL ECUADOR

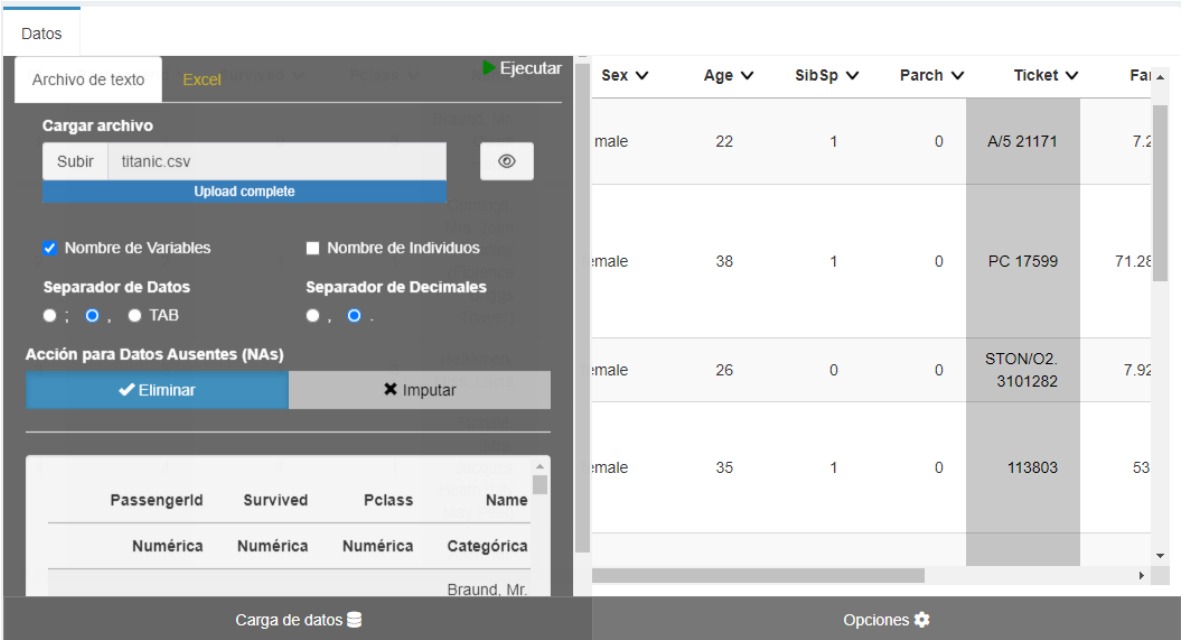
Ejercicio 1: [50 puntos] Esta pregunta utiliza los datos sobre la conocida historia y tragedia del Titanic, usando los datos (`titanic.csv`) de los pasajeros se trata de predecir la supervivencia o no de un pasajero.

La tabla contiene 12 variables y 1309 observaciones, las variables son:

- `PassegerId`: El código de identificación del pasajero (valor único).
- `Survived`: Variable a predecir, 1 (el pasajero sobrevivió) 0 (el pasajero no sobrevivió).
- `Pclass`: En que clase viajaba el pasajero (1 = primera, 2 = segunda , 3 = tercera).
- `Name`: Nombre del pasajero (valor único).
- `Sex`: Sexo del pasajero.
- `Age`: Edad del pasajero.
- `SibSp`: Cantidad de hermanos o cónyuges a bordo del Titanic.
- `Parch`: Cantidad de padres o hijos a bordo del Titanic.
- `Ticket`: Número de tiquete (valor único).
- `Fare`: Tarifa del pasajero.
- `Cabin`: Número de cabina (valor único).
- `Embarked`: Puerto donde embarco el pasajero (C = Cherbourg, Q = Queenstown, S = Southampton).

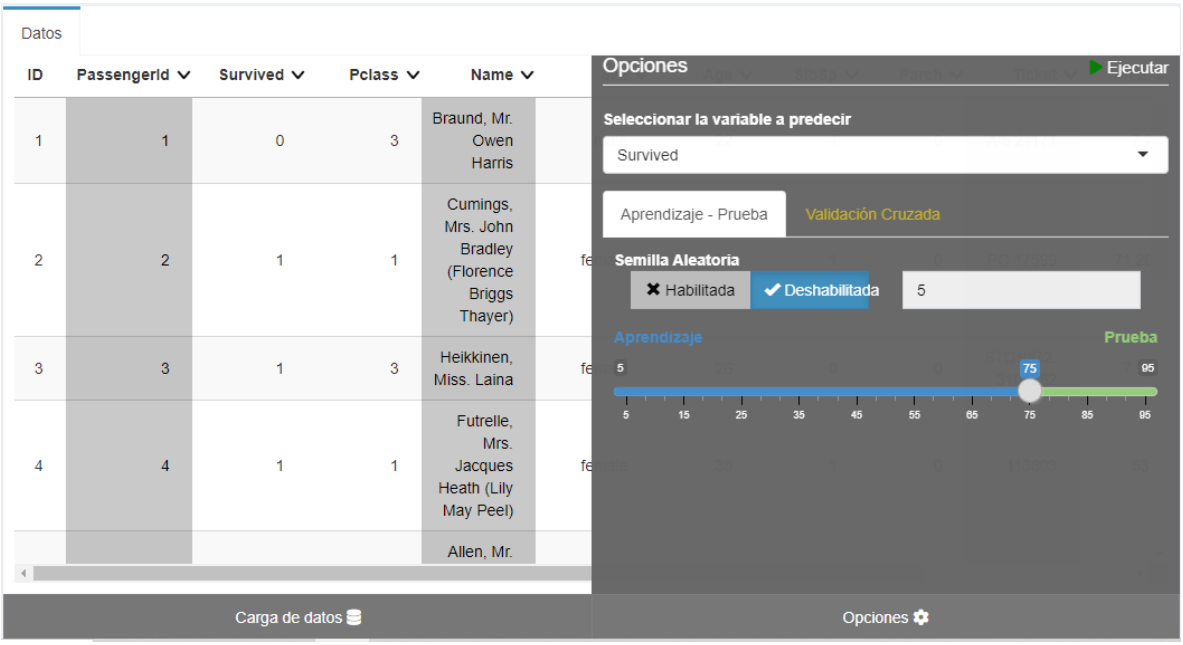
1. Cargue la tabla de datos `titanic.csv`, asegúrese re-codificar las variables cualitativas y de ignorar variables que no se deben usar.

ID	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fa
1	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.28
3	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.92
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53
				Allen, Mr.						



Una vez cargada la tabla procedemos a eliminar las variables únicas ya que estas no son consideradas para implementar un poder predictiva ya que no tiene una carga o peso sobre el cual podamos variar nuestros resultados

2. Con 100 árboles use el método de Bosques Aleatorios para generar modelos predictivos para la tabla `titanic.csv`. Para esto utilice el 75 % de los datos para la tabla aprendizaje y un 25 % para la tabla testing, luego calcule para los datos de testing la matriz de confusión, la precisión global y la precisión para cada una de las dos categorías.



Datos

ID	Partición	PassengerId	Survived	Pclass	
1	Aprendizaje	1	0	3	Braun
2	Aprendizaje	2	1	1	Cum Mrs. Br (Flo E Th
3	Prueba	3	1	3	Heikk Miss.
4	Aprendizaje	4	1	1	Fu Jac Heath May
					Alle

Opciones

Ejecutar

Seleccionar la variable a predecir

Survived

Aprendizaje - Prueba | Validación Cruzada

Semilla Aleatoria

✖ Habilitada | ✔ Deshabilitada | 5

Aprendizaje — Prueba

5 75 95

5 15 25 35 45 55 65 75 85 95

Carga de datos | Opciones

Generación del Modelo | Evolución del Error | Importancia de Variables | Predicción del Modelo | Matriz de Confusión | Índices Generales | Reglas | Probabilidad de Corte | Probabilidad de Corte con Paso

```
Call:
 randomForest(formula = as.formula(var), data = train, mtry = mtry,      ntree = ntree, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 100
```

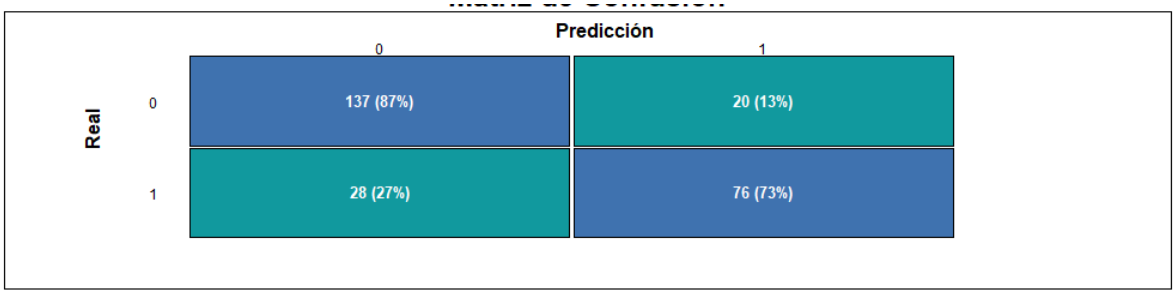
Opciones

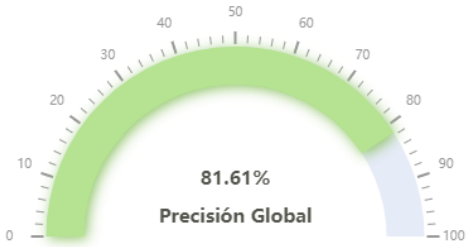
Ejecutar

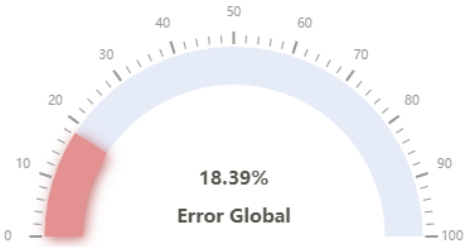
Número de Árboles: 100

Número de variables: 2

Opciones







Precision Global=	VN+VP / VN+FP+FN+VP	0,816091954	81,6091954
Precision Positiva=	VP / FN+VP	0,730769231	73,0769231
Precision Negativa=	VN / VN+FP	0,872611465	87,2611465

3. Explique una regla del árbol número 5 y una del árbol número 30.

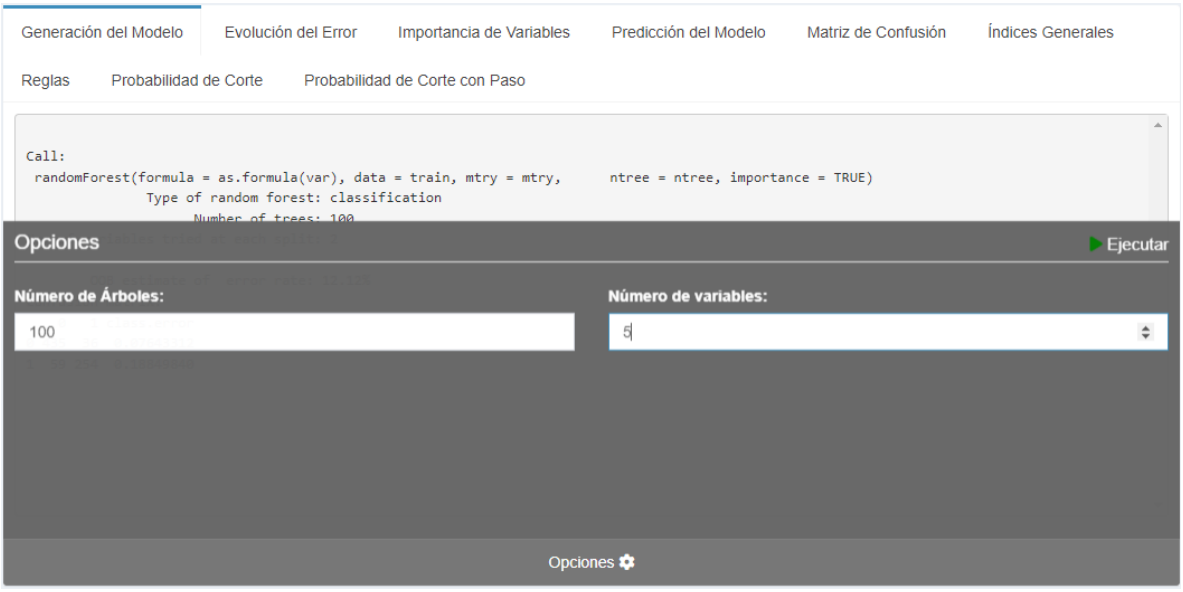
```
------------------------------------
Tree 5 Rule 4 Node 78 Decision 1

1: Parch <= 0.5
2: Sex IN ("female")
3: Age <= 23.5
4: Fare > 6.9875
5: Pclass IN ("3")
6: SibSp <= 0.5
7: Embarked IN ("", "C", "S")
8: Fare <= 7.7625
------------------------------------
```

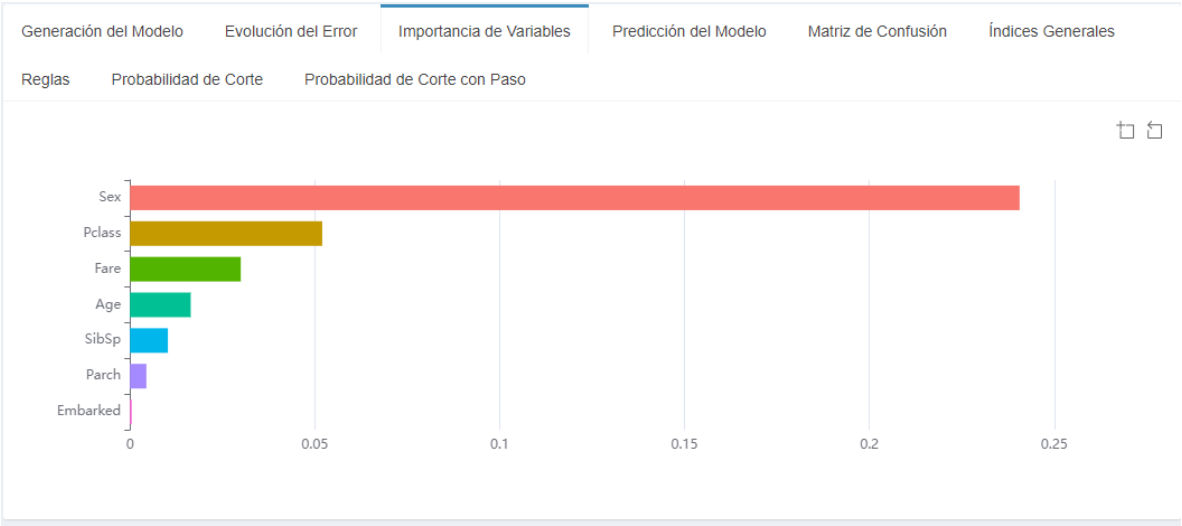
es una regla aplicada en el nodo número 78 de la estructura del árbol0, en el nodo 5, que identifica a un grupo específico de pasajeras (mujeres de 23.5 años o menos que viajaron solas o con un solo familiar cercano, en tercera clase y que embarcaron desde Cherbourg o Southampton) con una tarifa específica que cae dentro de un rango determinado. La "Decisión 1" a la que conduce esta regla predice la supervivencia en un viaje en barco

```
Tree 30 Rule 4 Node 128 Decision 1

1: Fare <= 15.64585
2: Sex IN ("female")
3: Fare <= 14.25415
4: Fare > 6.875
5: Pclass IN ("1", "3")
6: Age > 3.5
7: Fare <= 9.4125
8: SibSp <= 0.5
9: Fare <= 7.7
```

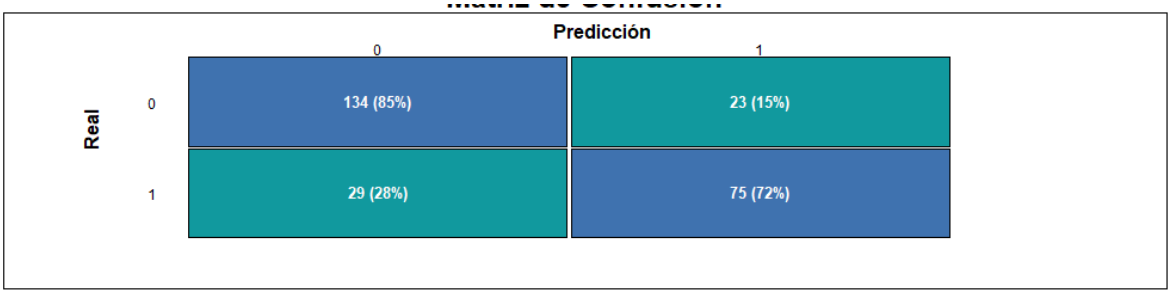
es una regla aplicada en el nodo número 128 de la estructura del árbol, que nos quiere decir si el valor de la tarifa del pasaje es menor o igual a 15.64585, y es mujer, y si mantiene una tarifa o si es menor o igual a 14.25415 y si está viajando en primera o tercera clase y si al edad del pasajero es mayor a 3.5 años de edad, verificamos nuevamente la tarifa si es menor o igual a 904125, entonces se verifica si el pasajero tiene un numero de hermano o conyugues menor a 0.5 lo que podemos redondear a 0, y verificamos nuevamente la tarifa con un valor de menor o igual a 7.7, si este pasajero cumple con todas estas condiciones podemos predecir que este pasajero si sobrevivirá en el barco.

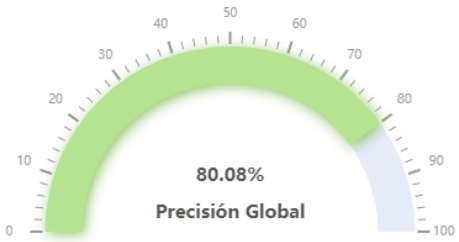
Repita los dos ejercicios anteriores pero esta vez usando solamente las 5 variables que según el gráfico de `Importancia de las Variables (MeanDecreaseGini)` tienen mejor poder predictivo. ¿Mejoró el resultado?

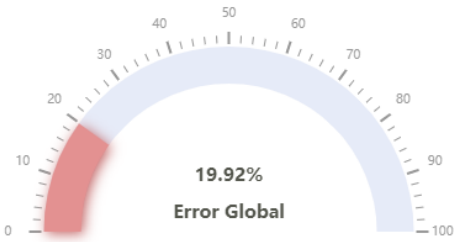




Cuando cambiamos o modificamos las variables a 5 podemos ver que les da aun mayor importancia a las variables ***Pclass*** y ***Fare***, en cambia cuando trabajamos anteriormente con dos variables no le daba mayor importancia a estas variables y se enfocaba más en la variable a predecir que es ***Survived***.







Precision Global=	VN+VP / VN+FP+FN+VP	0,800766284	80,0766284
Precision Positiva=	VP / FN+VP	0,721153846	72,1153846
Precision Negativa=	VN / VN+FP	0,853503185	85,3503185

Cuando revisamos la matriz de confusión y los valores que nos arroja ahora que modificamos las variables podemos ver que el modelo no varía en gran notoriedad lo que podemos considerar que le modelo siguen siendo bueno, y se nota una mejora ya que ahora se toma mayor importancia a otras variables que nos ayudan a predecir nuestra tabla

3. Explique una regla del árbol número 5 y una del árbol número 30.

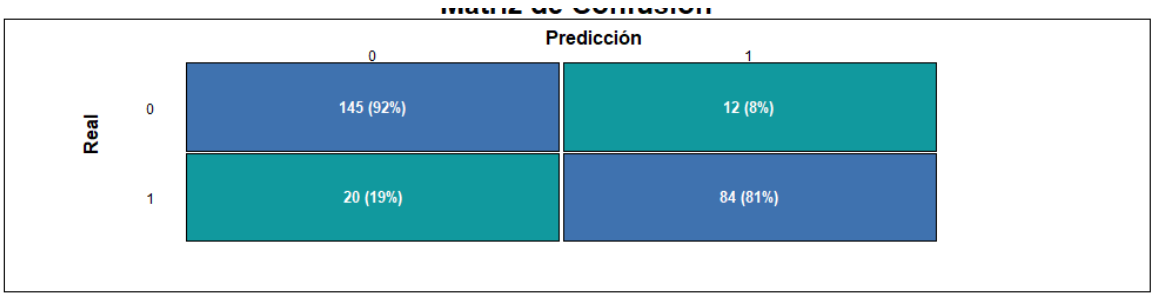
```
Tree 5 Rule 2 Node 48 Decision 1

1: Sex IN ("female")
2: Pclass IN ("1", "2")
3: Fare <= 149.0354
4: Age <= 27.5
5: Age > 25.5
6: Fare <= 17.42915
```

es una regla aplicada en el nodo número 48 de la estructura del árbol, que nos quiere decir si el pasajero a bordo corresponde al sexo femenino, verificaremos que este viajando en primera o segunda clase, entonces se verifica si tiene un boleto con tarifa menor o igual a 149.0354, continuaremos verificando si la edad del pasajero es mayor o igual a 27.5, después de este filtro se verifica nuevamente la edad si corresponde a menor de 25 años, con este filtro se verifica nuevamente la tarifa del pasajero siendo esta menor o igual a 17.42915, si el pasajero lograr cumplir con todas las verificaciones o culminar todas las ramas este pasajero sobrevivirá en el barco.

```
Tree 30 Rule 1 Node 16 Decision 0

1: Sex IN ("female")
2: Pclass IN ("1", "2")
3: Age <= 2.5
4: Pclass IN ("1")
```

es una regla aplicada en el nodo número 16 de la estructura del árbol, que nos quiere decir si del pasajero se van a seleccionar solo a las mujeres, de este filtro se procede a verificar si viajan en primera o segunda clases, a continuación, verificaremos la edad que tiene que ser menor o igual a 2.5 años, y volveremos a verificar si viajan en primera clases, si el pasajero cumple con las condiciones de las ramas entonces se puede deducir que el pasajero no sobrevivirá en el barco.

5. Con 100 árboles, con el algoritmo `real`, `Profundidad Máxima = 20` y `Mínimo para dividir un nodo = 5` use el Método de Potenciación (ADA Boosting) para generar modelos predictivos para la tabla `titanic.csv`. Para esto utilice el 75 % de los datos para la tabla aprendizaje y un 25 % para la tabla testing, luego calcule para los datos de testing la matriz de confusión, la precisión global y la precisión para cada una de las dos categorías.



	Modelo Predictivo		
	0	1	
0	145	12	
1	20	84	

Precision Global=	VN+VP / VN+FP+FN+VP	0,877394636	87,7394636
Precision Positiva=	VP / FN+VP	0,807692308	80,7692308
Precision Negativa=	VN / VN+FP	0,923566879	92,3566879

Ahora usando algoritmo real de árboles y método de (ADA Bossting) podemos ver que el modelo tuvo una mejora bastante notoria tanto en su predicción global como positiva y negativa dándonos un modelo más precisión y no muy sobreentrenado para ningún lado, sino que puede ser utilizado para relazar predicción de si y no con un acierto bastante confiable

6. Explique una regla del árbol número 5 y una del árbol número 30.

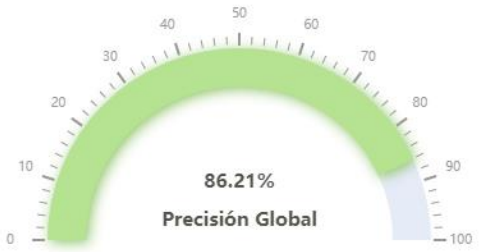
```
Rule number: 45 [Survived=1 cover=10 (1%) prob=1.00]
  Sex=male
  SibSp< 1.5
  Fare>=7.91
  Fare>=8.206
  Fare>=387.7
```

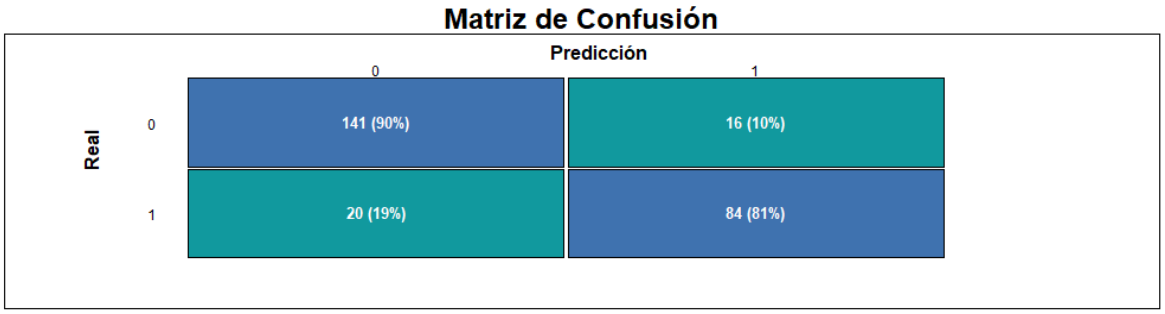
Es una regla aplicada al número 45 de la estructura del árbol, que nos quiere decir que el pasajero que este dentro de esta regla, va a sobrevivir y que este caso cubre el uno por ciento de toda nuestra bases y que la probabilidad de que un pasajero que cumpla con esta regla sobreviva es del 100%, dentro de las condiciones vemos que primero se verifica que el pasajero sea masculino o hombre, y luego se procede a verificar si tiene hermanos o conyugues menor o igual a 1.5 o redondeado a 2, se verifica ahora la tarifa que sea mayor o igual a 7.91, y una vez más se filtra por tarifa a hora mayor o igual a 8.26, y volvemos a filtrar la tarifa que sea mayor o igual a 387.7, si un pasajero cumple con estas condiciones se puede deducir que el pasajero sobrevivirá.

```
Rule number: 19 [Survived=1 cover=8 (1%) prob=1.00]
  Fare< 86.29
  Fare>=20.79
  Fare>=25.64
  Age>=72.5
```

Es una regla aplicada al número 19 de la estructura del árbol, que nos quiere decir que el pasajero que este dentro de esta regla, va a sobrevivir y que este caso cubre el uno por ciento de toda nuestra bases y que la probabilidad de que un pasajero que cumpla con esta regla sobreviva es del 100%, dentro de las condiciones vemos que primero verificamos la tarifa que sea menor a 86.29, una vez filtrado volvemos a verificar tarifa ahora que sea mayor o igual a 20.79, nuevamente después del filtro se verifica de nuevo al tarifa ahora que sea mayor o igual a 25.64, una vez filtrado verificamos la edad que sea mayor a 72.5 años, los pasajeros que cumplan con estas condiciones podrá sobrevivir en un 100% en el barco.

7. Repita los dos ejercicios anteriores pero esta vez usando solamente las 5 variables que según el gráfico de `Importancia de las Variables` tienen mejor poder predictivo. ¿Mejoró el resultado?





	Modelo Predictivo		
	0	1	
0	141	16	
1	20	84	

Precision Global=	VN+VP / VN+FP+FN+VP	0,862068966	86,2068966
Precision Positiva=	VP / FN+VP	0,807692308	80,7692308
Precision Negativa=	VN / VN+FP	0,898089172	89,8089172

Aplicando el cambio en el cual tomamos las 5 variables más importantes que el grafico nos muestra, que el modelo no tiene un gran cambio ni variación, sino que más bien se conserva dentro de los parámetros del anterior ejercicio, aun nos presenta un modelo bastante bueno y entrenado para predecir.

Explique una regla del árbol número 5 y una del árbol número 30.

```
Rule number: 39 [Survived=1 cover=61 (8%) prob=0.84]
  Fare>=8.04
  Age< 47.5
  Sex=female
  Pclass=1,2
  Age>=3
```

Es una regla aplicada al número 39 de la estructura del árbol, que nos quiere decir que el pasajero que este dentro de esta regla, va a sobrevivir y que este caso cubre el 8% de toda nuestra bases y que la probabilidad de que un pasajero que cumpla con esta regla sobreviva es del 84%, dentro de las condiciones vemos que primero verificamos la tarifa que sea mayor o igual a 8.04, luego procedemos a verificar la edad que sea menor a 47.5, una vez filtrado procedimos a seleccionar solo a las mujeres, que estén viajando en primera o segunda clases y que tengas una edad mayor o igual a 3 años, las personas que cumplan estas condiciones tendrán una posibilidad de sobrevivir de un 84%

```
Rule number: 7 [Survived=1 cover=180 (23%) prob=0.66]
  Fare>=15.4
  Age< 40.5
```

Es una regla aplicada al número 7 de la estructura del árbol, que nos quiere decir que el pasajero que este dentro de esta regla, va a sobrevivir y que este caso cubre el 23 % de toda nuestra bases y que la probabilidad de que un pasajero que cumpla con esta regla sobreviva es del 66%, dentro de las condiciones vemos que primero verificamos la tarifa que sea mayor o igual a 15.4 y luego las personas que sean mayor a 40.5 años, estas personas tendrá una posibilidad de sobrevivir de un 66%

8. Con `Número Máximo de Iteraciones = 100` (`nrounds` que es el número de árboles), con el Amplificador (booster - algoritmo) `gbtree` y `Profundidad Máxima = 20` use el Método de Potenciación (Extreme Boosting) para generar modelos predictivos para la tabla `titanic.csv`. Para esto utilice el 75 % de los datos para la tabla aprendizaje y un 25 % para la tabla testing, luego calcule para los datos de testing la matriz de confusión, la precisión global y la precisión para cada una de las dos categorías.

Opciones

Ejecutar

Seleccionar el Amplificador:

gbtree

Profundidad Máxima:

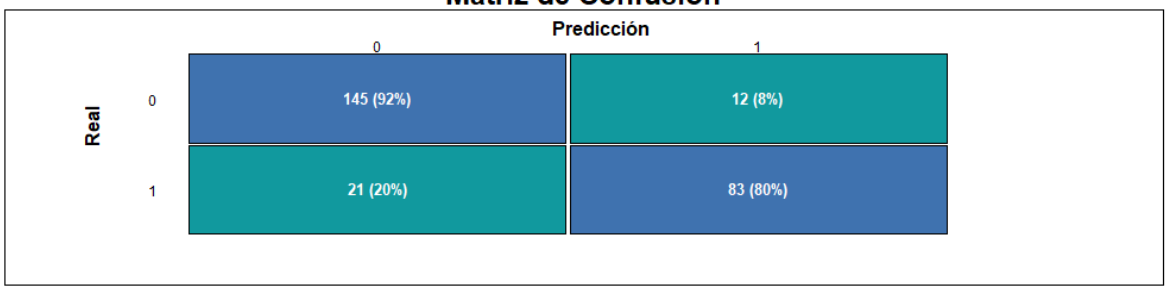
20

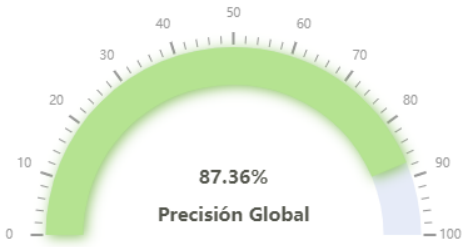
Número Máximo de Iteraciones:

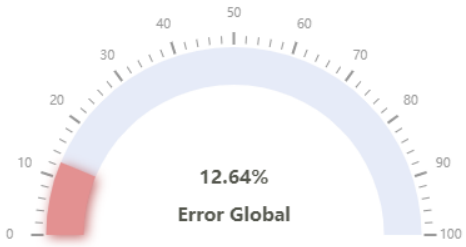
100

Opciones

Matriz de Confusión



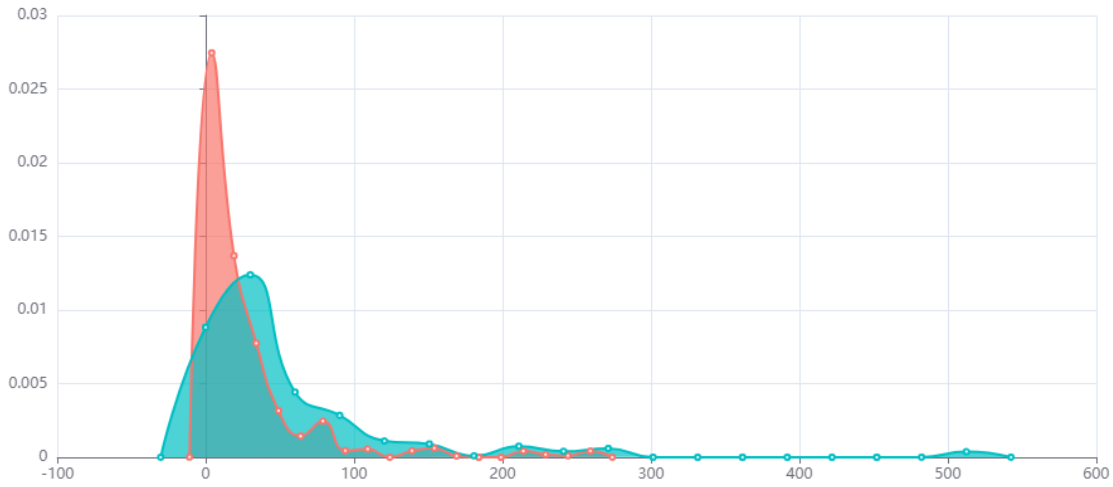


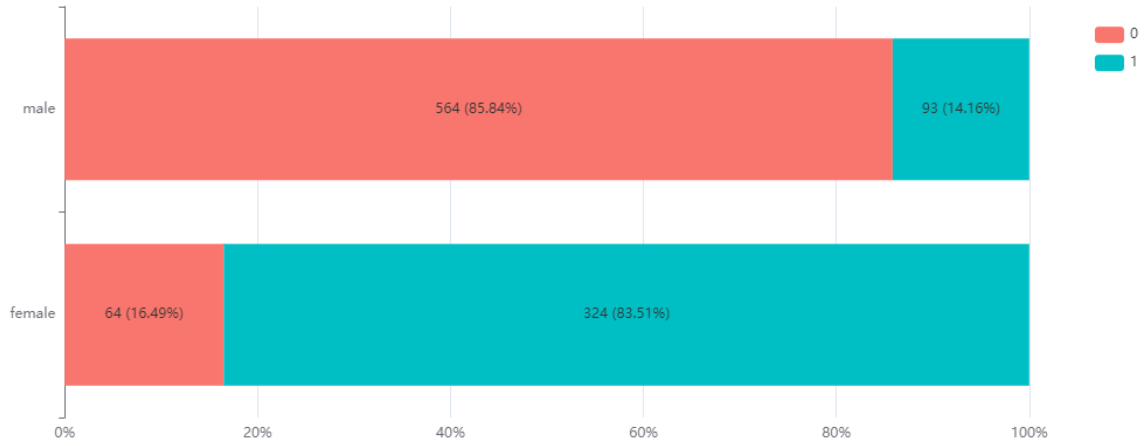


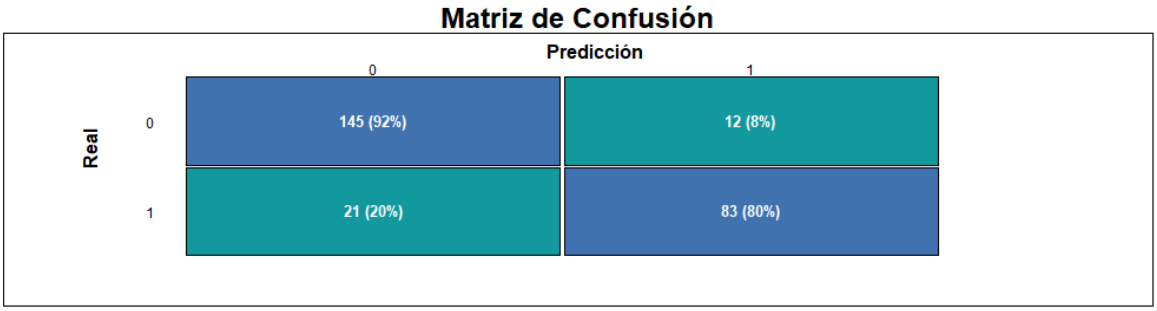
	0	1	
0	145	12	
1	21	83	

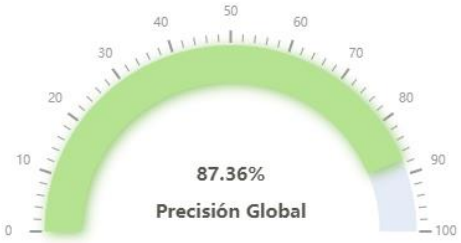
Precision Global=	VN+VP / VN+FP+FN+VP	0,873563218	87,3563218
Precision Positiva=	VP / FN+VP	0,798076923	79,8076923
Precision Negativa=	VN / VN+FP	0,923566879	92,3566879

9. Repita el ejercicio anterior pero esta vez usando solamente las 5 variables que según el gráfico de `Importancia de las Variables` tienen mejor poder predictivo. ¿Mejoró el resultado?




male
female
564 (85.84%)
93 (14.16%)
64 (16.49%)
324 (83.51%)
0%
20%
40%
60%
80%
100%
0
1


Matriz de Confusión
Predicción
0
1
Real
0
1
145 (92%)
12 (8%)
21 (20%)
83 (80%)


87.36%
Precisión Global

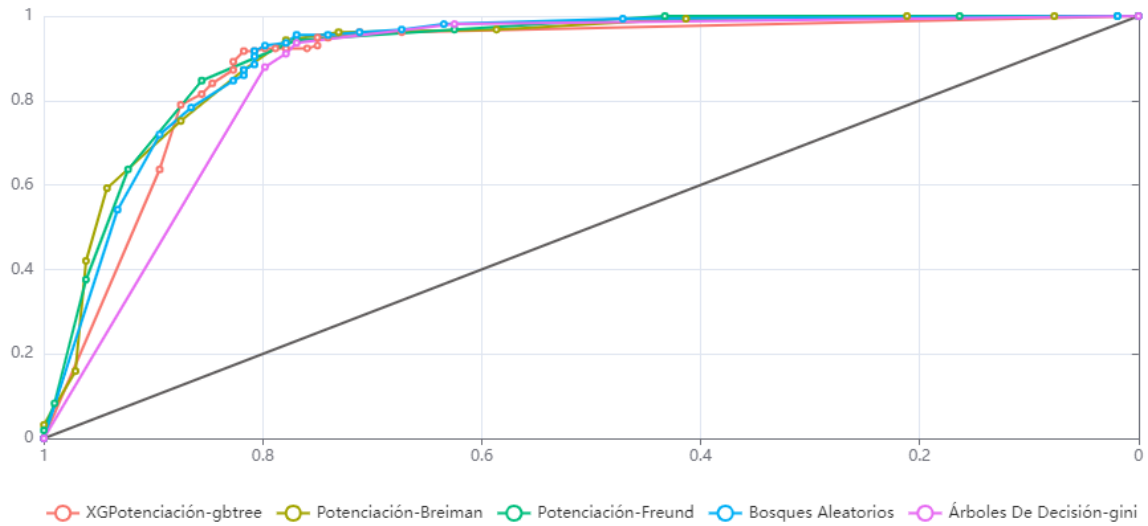

12.64%
Error Global

	0	1	
0	145	12	
1	21	83	

Precision Global=	VN+VP / VN+FP+FN+VP	0,873563218	87,3563218
Precision Positiva=	VP / FN+VP	0,798076923	79,8076923
Precision Negativa=	VN / VN+FP	0,923566879	92,3566879

10. Genere la curva ROC para todos los modelos generados en los puntos anteriores, ¿Cuál es el mejor modelo según esta curva? Compare con la curva ROC con las generadas en la tarea anterior. ¿Cuál modelo es mejor?

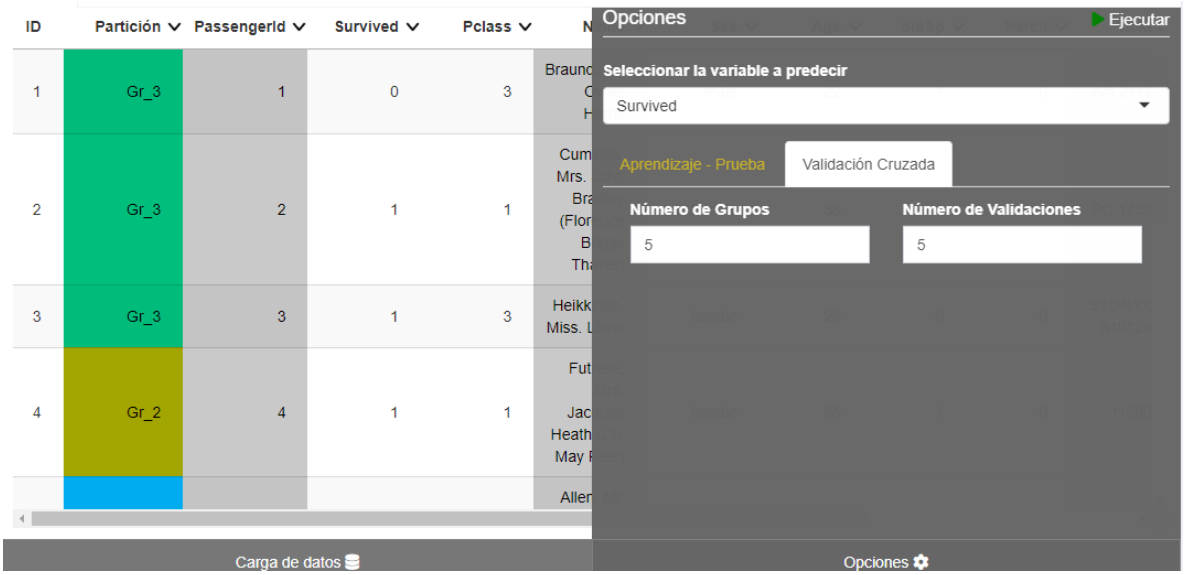
	Precisión Global	Error Global	0	1	Área de ROC
XGPotenciación-gbtree	87.35632	12.64368	92.35669	79.80769	8.63241
Potenciación-Breiman	87.35632	12.64368	92.35669	79.80769	8.36906
Potenciación-Freund	85.82375	14.17625	89.17197	80.76923	8.27107
Bosques Aleatorios	87.35632	12.64368	92.35669	79.80769	8.72122
Árboles De Decisión-gini	85.82375	14.17625	91.0828	77.88462	12.78785



Los resultados de la tabla muestran que los cuatro algoritmos tienen un buen rendimiento, con una precisión global superior al 85%. Sin embargo, el algoritmo ***XGPotenciación-gbtree*** es el que mejor rendimiento tiene, con una precisión global del 87.35632% y un error global del 12.64368%. Este algoritmo también tiene el mejor valor de área bajo la curva ROC (92.35669).

Los otros tres algoritmos tienen un rendimiento similar, con una precisión global de alrededor del 85.82375% y un error global de alrededor del 14.17625%. Sin embargo, el algoritmo ***Árboles De Decisión-gini*** tiene un mejor valor de área bajo la curva ROC (91.0828) que el algoritmo ***Potenciación-Freund*** (89.17197), en base a los resultados de la tabla, se puede concluir que el algoritmo ***XGPotenciación-gbtree*** es el que mejor rendimiento tiene de los cuatro algoritmos evaluados. Sin embargo, los otros tres algoritmos también tienen un buen rendimiento y pueden ser una buena opción dependiendo de las necesidades específicas de la aplicación.

Validacion Cruzada



Seleccionar Modelo:

XGPotenciación

Profundidad Máxima: 20

Seleccionar el Amplificador: gbtree

Número Máximo de Iteraciones: 100

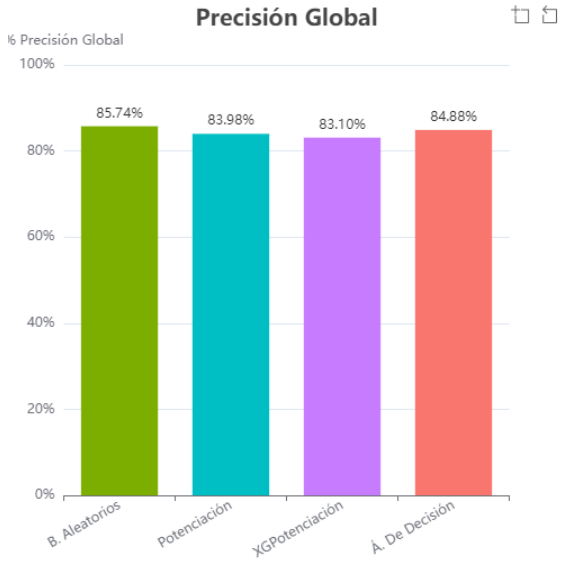
Probabilidad de Corte: 0,5

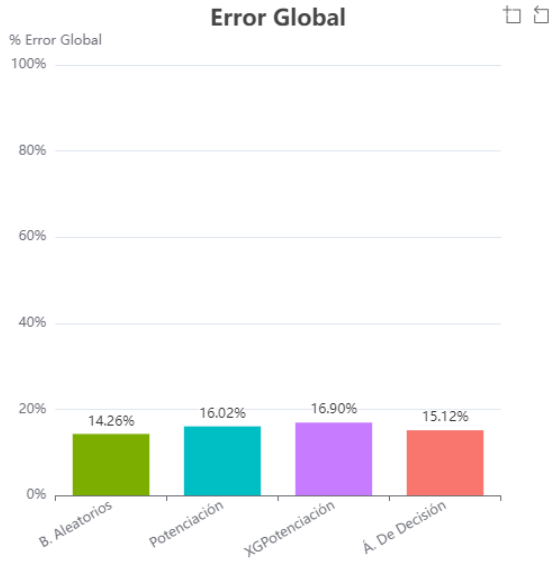
Seleccionar Categoría: 0

Generar todos los Modelos

```
    0 566  62
    1  94 323

$MCs.dtl[[4]]
     Pred
Clase   0   1
    0 569  59
    1 100 317

$MCs.dtl[[5]]
     Pred
Clase   0   1
    0 570  58
    1 104 313
```

Gráfico de Barras





	Precisión Global	Error Global	Precisión 0	Precisión 1
Potenciación	83.98086	16.01914	88.18471	77.64988
Árboles De Decisión	84.88038	15.11962	90.41401	76.54676
Bosques Aleatorios	85.74163	14.25837	91.11465	77.64988
XGPotenciación	83.10048	16.89952	87.35669	76.69065

El modelo **Bosques Aleatorios** tiene la mayor precisión global (85.74163%), seguido de **Árboles de Decisión** (84.88038%) y **Potenciación** (83.98086%). El modelo **XGPotenciación** tiene la menor precisión global (83.10048%). El modelo **XGPotenciación** tiene el mayor error global (16.89952%), seguido de **Potenciación** (16.01914%) y **Árboles de Decisión** (15.11962%). El modelo **Bosques**

Aleatorios tiene el menor error global (14.25837%). El modelo **Potenciación** tiene la mayor precisión para predecir la clase 0 (88.18471%), seguido de **Árboles de Decisión** (90.41401%) y **Bosques Aleatorios** (91.11465%). El modelo **XGPotenciación** tiene la menor precisión para predecir la clase 0 (87.35669%).

El modelo **Bosques Aleatorios** tiene la mayor precisión para predecir la clase 1 (77.64988%), seguido de **Potenciación** (77.64988%) y **Árboles de Decisión** (76.54676%). El modelo **XGPotenciación** tiene la menor precisión para predecir la clase 1 (76.69065%).

El modelo **Bosques Aleatorios** tiene el mejor rendimiento general, con la mayor precisión global y el menor error global. El modelo **Potenciación** tiene el mejor rendimiento para predecir la clase 0, mientras que el modelo **Bosques Aleatorios** tiene el mejor rendimiento para predecir la clase 1. El modelo **XGPotenciación** tiene el peor rendimiento general.

- **Ejercicio 2:** [50 puntos] Esta pregunta utiliza los datos sobre muerte del corazón en Sudáfrica (`SAheart.csv`). La variable que queremos predecir es `chd` que es un indicador de muerte coronaria basado en algunas variables predictivas (factores de riesgo) como son el fumado, la obesidad, las bebidas alcohólicas, entre otras. Las variables son:

  - `sbp`: systolic blood pressure (numérica).
  - `tobacco`: cumulative tobacco (kg) (numérica).
  - `ldl`: low densiity lipoprotein cholesterol (numérica).
  - `Adiposity`: Adiposity level (numérica).
  - `famhist`: family history of heart disease (Present, Absent) (categórica).
  - `typea`: type-A behavior (numérica).
  - `Obesity`: Obesity of the person (numérica).
  - `alcohol`: current alcohol consumption (numérica).
  - `age`: age at onset (numérica).
  - `chd`: coronary heart disease (categórica).

1. Con 75 árboles use el método de Bosques Aleatorios para generar modelos predictivos para la tabla `SAheart.csv`. Para esto utilice el 80 % de los datos para la tabla aprendizaje y un 20 % para la tabla testing, luego calcule para los datos de testing la matriz de confusión, la precisión global y la precisión para cada una de las dos categorías.

Datos

ID	sbp	tobacco	ldl	adiposity	famhist
1	160	12	5.73	23.11	Present
2	144	0.01	4.41	28.61	Absent
3	118	0.08	3.48	32.28	Present
4	170	7.5	6.41	38.03	Present
5	134	13.6	3.5	27.78	Present
6	132	6.2	6.47	36.21	Present
7	142	4.05	3.38	16.2	Absent
8	114	4.08	4.59	14.6	Present
9	114	0	3.83	19.4	Present
10	132	0	5.8	30.96	Present
	Numérica	Numérica	Numérica	Numérica	Categórica

Carga de datos

Opciones

Ejecutar

Seleccionar la variable a predecir

chd

Aprendizaje - Prueba | Validación Cruzada

Semilla Aleatoria

✖ Habilitada | ✔ Deshabilitada | 5

Aprendizaje | Prueba

5 80 95

5 15 25 35 45 55 65 75 85 95

Opciones

ID	Partición	sbp	tobacco	ldl	adiposity
1	Prueba	160	12	5.73	23.11
2	Aprendizaje	144	0.01	4.41	28.61
3	Aprendizaje	118	0.08	3.48	32.28
4	Prueba	170	7.5	6.41	38.03
5	Aprendizaje	134	13.6	3.5	27.78
6	Aprendizaje	132	6.2	6.47	36.21
7	Aprendizaje	142	4.05	3.38	16.2
8	Prueba	114	4.08	4.59	14.6
9	Prueba	114	0	3.83	19.4
10	Aprendizaje	132	0	5.8	30.96
		Numérica	Numérica	Numérica	Numérica

Carga de datos

Opciones

Ejecutar

Seleccionar la variable a predecir

chd

Aprendizaje - Prueba | Validación Cruzada

Semilla Aleatoria

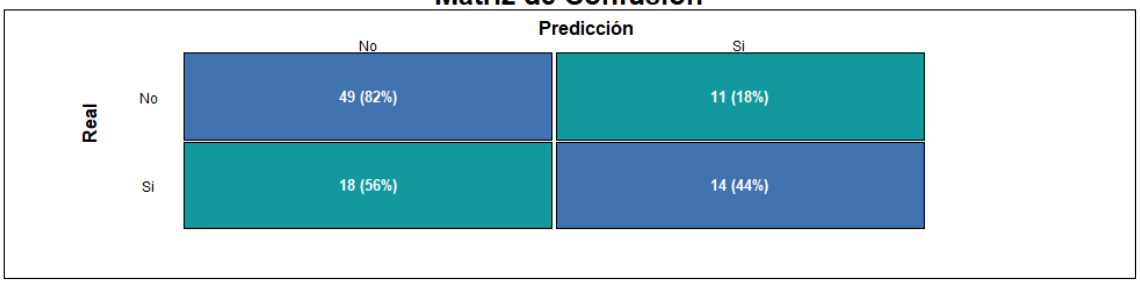
✖ Habilitada | ✔ Deshabilitada | 5

Aprendizaje | Prueba

5 80 95

5 15 25 35 45 55 65 75 85 95

Matriz de Confusión



	0	1	
0	49	11	
1	18	14	

Precision Global=	VN+VP / VN+FP+FN+VP	0,684782609	68,4782609
Precision Positiva=	VP / FN+VP	0,4375	43,75
Precision Negativa=	VN / VN+FP	0,816666667	81,6666667

Una vez realizada los bosques aleatorios podemos ver que se nos presenta un modelo muy sobreentrenado hacia el no con una precisión global bastante baja lo que nos indica que no es un buen modelo, La imagen muestra que la precisión del modelo es moderada. La calculadora es más precisa para valores reales que están entre 40 y 60, pero tiene un error mayor para valores reales que están entre 20 y 80.

2. Explique una regla del árbol número 2 y una del árbol número 10.

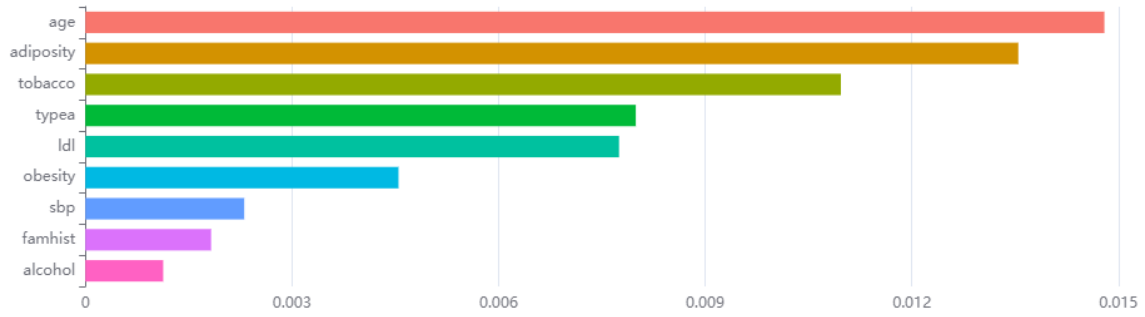
```
Tree 2 Rule 1 Node 12 Decision Si

1: ldl <= 8.25
2: age <= 40.5
3: adiposity <= 25.57
4: obesity <= 18.48
```

Es una regla aplicada al número 1 del nodo número 12 de la estructura del árbol, que nos quiere decir que el pasajero que este dentro de esta regla, pude contener un indicador de muerte coronaria dentro de las condiciones vemos que primero verificamos que el colesterol unido a lipoproteínas de baja densidad sea menor o igual a 8.25 y la edad sea menor o igual a 40.5 años, con un Nivel de adiposidad menor o igual a 25.57 y Obesidad de la persona menor o igual a 18.48, la persona que cumpla con estas condiciones tendrá una posibilidad de adquirir o sufrir de muerte coronaria es positiva.

```
Tree 10 Rule 3 Node 73 Decision No

1: tobacco <= 0.585
2: age <= 62.5
3: sbp <= 187
4: adiposity <= 38
5: alcohol <= 7.25
6: age <= 31
7: alcohol > 0.13
8: alcohol > 0.385
```

Es una regla aplicada al número 3 del nodo número 73 de la estructura del árbol, que nos quiere decir que el pasajero que este dentro de esta regla no pude contener un indicador de muerte coronaria dentro de las condiciones vemos que primero verificamos que tabaco acumulado (kg) sea menor o igual a 0.585, a su vez que tenga una edad menor o igual a 62.5, también que su presión arterial sistólica sea menor o igual a 187, igual verificamos que Nivel de adiposidad sea menor de 38, igual que su nivel de alcohol sea menor o igual a 7.25, una vez con este filtro se vuele a filtrar por edad a las personas menores o igual a 31 años, su consumo actual de alcohol mayor a 0.13 y nuevamente su consumo actual de alcohol mayor a 0.385, estas personas no tienen la posibilidad de tener muerte coronaria.

3. Repita los dos ejercicios anteriores pero esta vez usando solamente las 4 variables que según el gráfico de `Importancia de las Variables (MeanDecreaseGini)` tienen mejor poder predictivo. ¿Mejoró el resultado?



Generación del Modelo | Evolución del Error | Importancia de Variables | Predicción del Modelo | Matriz de Confusión | Índices Generales

Reglas | Probabilidad de Corte | Probabilidad de Corte con Paso

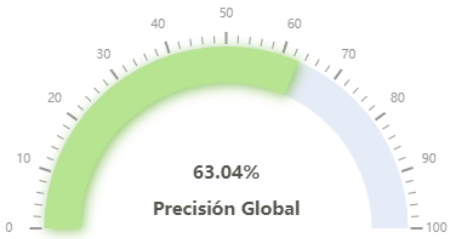
```
Call:
 randomForest(formula = as.formula(var), data = train, mtry = mtry,      ntree = ntree, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 75
```

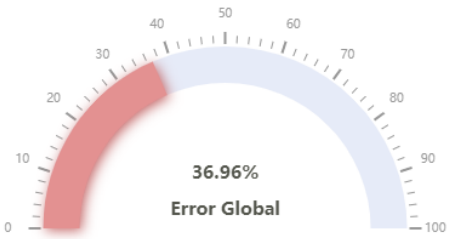
Opciones ▶ Ejecutar

Número de Árboles: 75

Número de variables: 4

Opciones ⚙

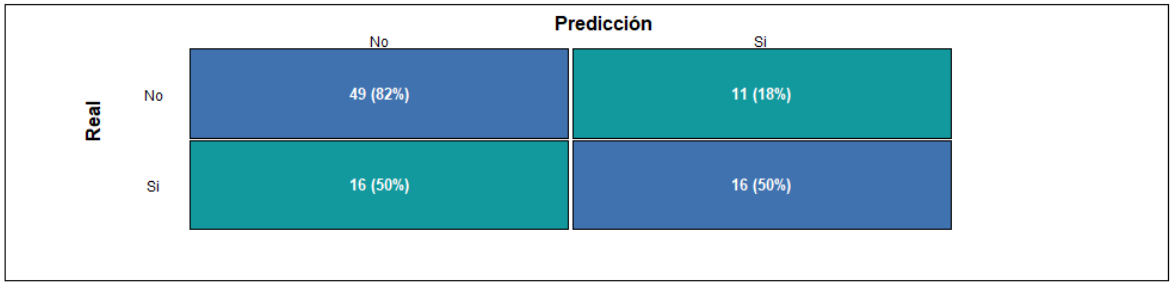




Precisión.No	Precisión.Si
71.67	46.88

Error.No	Error.Si
28.33	53.12

Matriz de Confusión



	FN	FP	
VN	49	11	
VP	16	16	
Precision Global=	VN+VP / VN+FP+FN+VP	0,706521739	70,6521739
Precision Positiva=	VP / FN+VP	0,5	50
Precision Negativa=	VN / VN+FP	0,816666667	81,6666667

En el caso de la tabla que se muestra en la imagen, la precisión global para ambas variables es de alrededor del 71%. La precisión positiva es de alrededor del 50%, mientras que la precisión negativa es de alrededor del 82%. Esto significa que la medición es más precisa para identificar valores negativos que para identificar valores positivos. La tabla proporciona información útil sobre la precisión de las mediciones realizadas para las dos variables. La precisión global es relativamente alta, pero hay algunas diferencias en la precisión positiva y negativa para cada variable. Esta información puede ser útil para evaluar la confiabilidad de las mediciones y para tomar decisiones sobre cómo utilizar los resultados de las mediciones.

Explique una regla del árbol número 2 y una del árbol número 10.

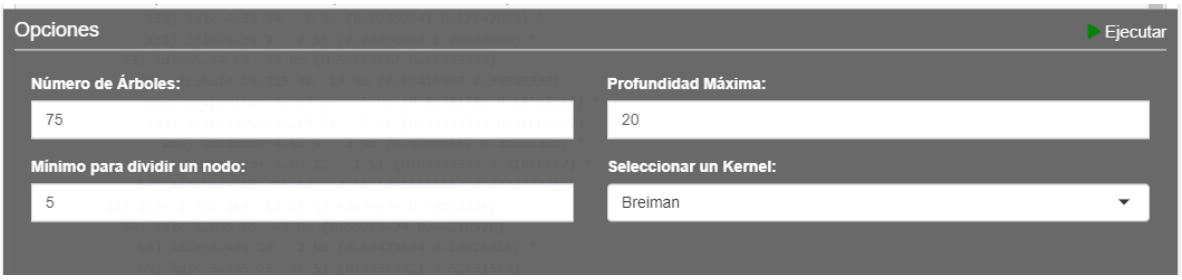
```
Tree 2 Rule 3 Node 36 Decision Si

1: adiposity <= 34.755
2: age <= 30.5
3: obesity > 18.48
4: tobacco > 1.055
5: alcohol <= 43.15
6: sbp <= 112
```

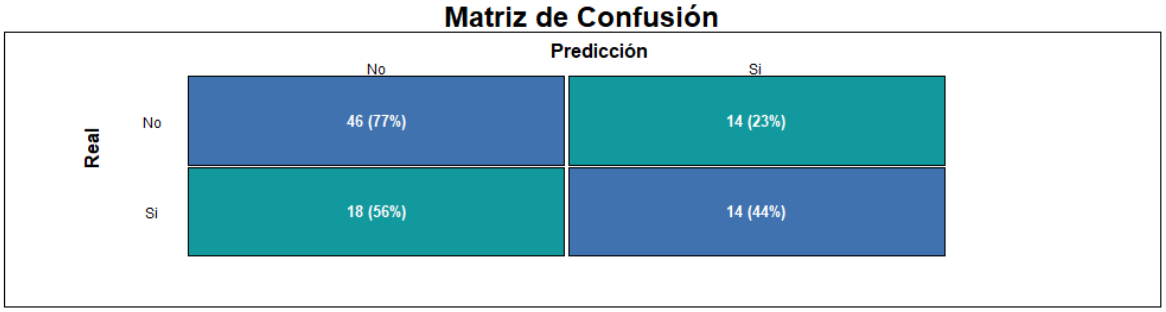
Es una regla aplicada al número 3 del nodo número 36 de la estructura del árbol, que nos quiere decir que el pasajero que este dentro de esta regla pude contener un indicador de muerte coronaria dentro de las condiciones vemos que primero verificamos que Nivel de adiposidad sea menor o igual a 34.755 igual que su edad se menor o igual a 30.5 años, Obesidad de la persona sea mayor a 18.48, el tabaco acumulado mayor a 1.055, su consumo actual de alcohol menor o igual a 43.55 y su presión arterial sistólica menor o igual a 122, la persona que presente estas condiciones si puede tener una muerte coronaria.

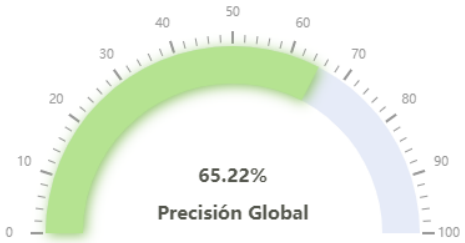
```
Tree 10 Rule 1 Node 16 Decision Si

1: age <= 37.5
2: tobacco <= 1.43
3: alcohol <= 0.385
4: obesity <= 18.92
```

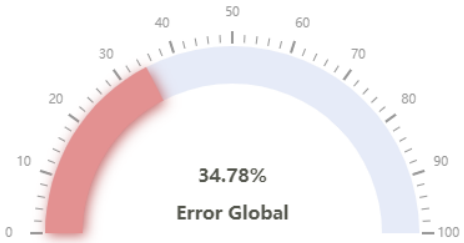
Es una regla aplicada al número 1 del nodo número 16 de la estructura del árbol, que nos quiere decir que el pasajero que este dentro de esta regla pude contener un indicador de muerte coronaria dentro de las condiciones vemos que primero verificamos que la edad sea menor o igual a 37.5, el tabaco acumulado sea menor o igual a 1.43, el consumo actual de alcohol menor o igual a 0.385 y la Obesidad de la persona menor o igual a 18.92, esta persona tendrá indicadores de una muerte coronaria.

4. Con 75 árboles, con el algoritmo `real`, `Profundidad Máxima = 20` y `Mínimo para dividir un nodo = 5` use el Método de Potenciación (ADA Boosting) para generar modelos predictivos para la tabla `SAheart.csv`. Para esto utilice el 80 % de los datos para la tabla aprendizaje y un 20 % para la tabla testing, luego calcule para los datos de testing la matriz de confusión, la precisión global y la precisión para cada una de las dos categorías.









Precisión.No	Precisión.Si
76.67	43.75

Error.No	Error.Si
23.33	56.25

	FN	FP	
VN	46	14	
VP	18	14	
Precision Global=	VN+VP / VN+FP+FN+VP	0,652173913	65,2173913
Precision Positiva=	VP / FN+VP	0,4375	43,75
Precision Negativa=	VN / VN+FP	0,766666667	76,6666667

En el caso de la tabla que se muestra en la imagen, la precisión global para ambas variables es de alrededor del 65%. La precisión positiva es de alrededor del 43%, mientras que la precisión negativa es de alrededor del 877%. Esto significa que la medición es más precisa para identificar valores negativos que para identificar valores positivos. La tabla proporciona información útil sobre la precisión de las mediciones realizadas para las dos variables. La precisión global es

relativamente alta, pero hay algunas diferencias en la precisión positiva y negativa para cada variable. Esta información puede ser útil para evaluar la confiabilidad de las mediciones y para tomar decisiones sobre cómo utilizar los resultados de las mediciones. En general, el modelo de clasificación tiene un buen rendimiento, con una precisión global del 65,2173913%. Sin embargo, el rendimiento del modelo es mejor en la clasificación de casos negativos que en la clasificación de casos positivos.

5. Explique una regla del árbol número 2 y una del árbol número 10.

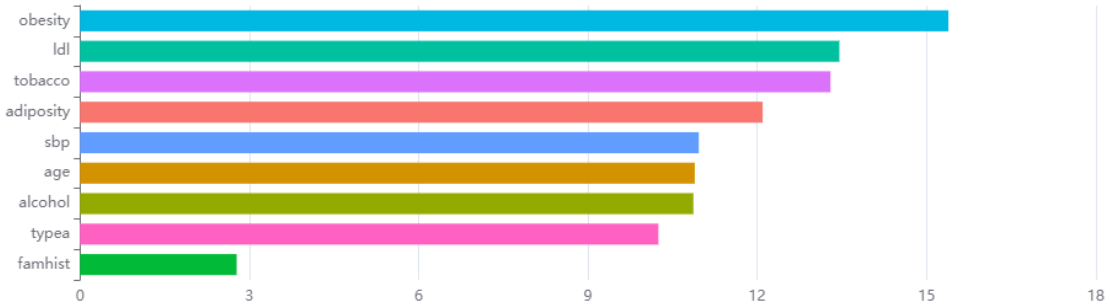
```
Rule number: 7 [chd=Si cover=15 (4%) prob=1.00]
  typea>=66.5
  sbp< 127
```

Es una regla aplicada al número de la estructura del árbol con el 4% de toda la base, que nos quiere decir que el pasajero que este dentro de esta regla pude contener un indicador de muerte coronaria dentro de las condiciones vemos que primero verificamos que comportamiento tipo A sea mayor o igual a 66.5 y su presión arterial sistólica menor a 127, esta persona tendrá la probabilidad de que si tenga una muerte coronaria con un 100%

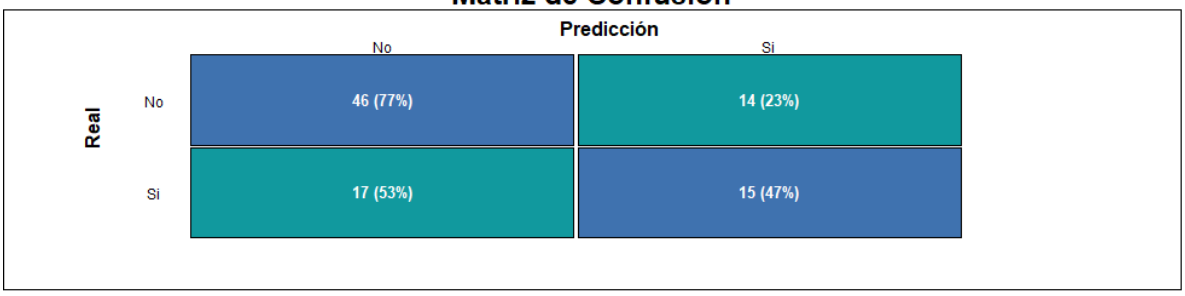
```
Rule number: 7 [chd=Si cover=25 (7%) prob=0.84]
  age>=27.5
  sbp>=175
```

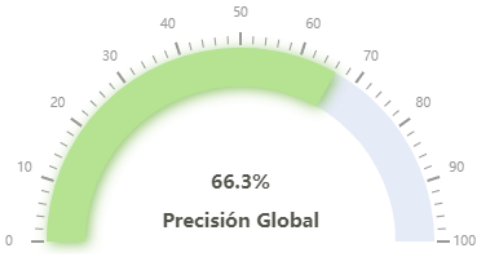
Es una regla aplicada al número 7 de la estructura del árbol con el 7% de toda la base, que nos quiere decir que el pasajero que este dentro de esta regla pude contener un indicador de muerte coronaria dentro de las condiciones vemos que primero vemos la edad sea mayor o igual a 27.5, de igual manera su presión arterial sistólica mayor o igual a 175 esta persona tendrá la probabilidad de que si tenga una muerte coronaria con un 84%.

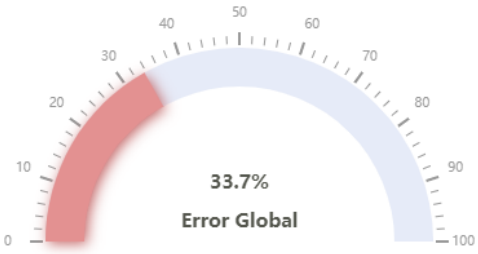
6. Repita los dos ejercicios anteriores pero esta vez usando solamente las 4 variables que según el gráfico de `Importancia de las Variables` tienen mejor poder predictivo. ¿Mejoró el resultado?



Matriz de Confusión







Precisión.No	Precisión.Si
76.67	46.88

Error.No	Error.Si
23.33	53.12

	FN	FP
VN	46	14
VP	17	15

Precision Global=	VN+VP / VN+FP+FN+VP	0,663043478	66,3043478
Precision Positiva=	VP / FN+VP	0,46875	46,875
Precision Negativa=	VN / VN+FP	0,766666667	76,6666667

En el caso de la tabla que se muestra en la imagen, la precisión global para ambas variables es de alrededor del 66%. La precisión positiva es de alrededor del 46%, mientras que la precisión negativa es de alrededor del 77%. Esto significa que la medición es más precisa para identificar valores negativos que para identificar valores positivos. La tabla proporciona información útil sobre la precisión de las mediciones realizadas para las dos variables. La precisión global es relativamente alta, pero hay algunas diferencias en la precisión positiva y negativa para cada variable. Pero seguimos sin ver una mejora en el modelo notable.

Explique una regla del árbol número 2 y una del árbol número 10.

```
Rule number: 12 [chd=Si cover=13 (4%) prob=0.15]
  tobacco>=7.55
  obesity< 26.3
  ldl>=4.58
```

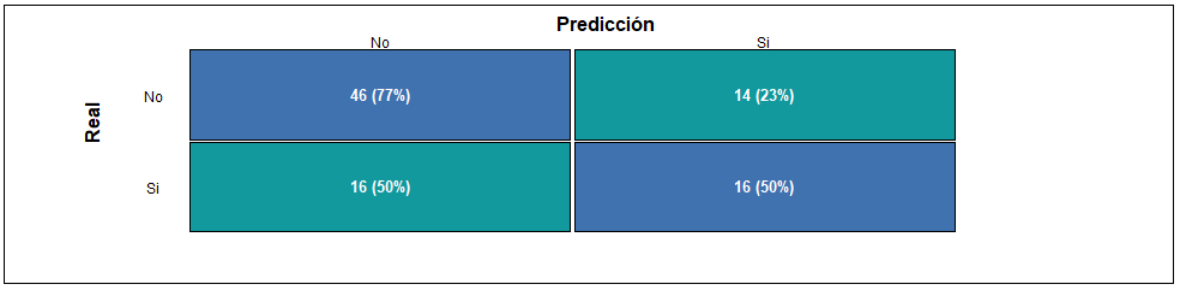
Es una regla aplicada al número 12 de la estructura del árbol con el 4% de toda la base, que nos quiere decir que el pasajero que este dentro de esta regla pude contener un indicador de muerte coronaria dentro de las condiciones vemos que primero vemos tabaco acumulado mayor o igual a 7.55, su Obesidad de la persona menor a 26.3 y su colesterol unido a lipoproteínas de baja densidad mayor o igual a 4.58, esta persona tendrá la probabilidad de que si tenga una muerte coronaria con un 15%.

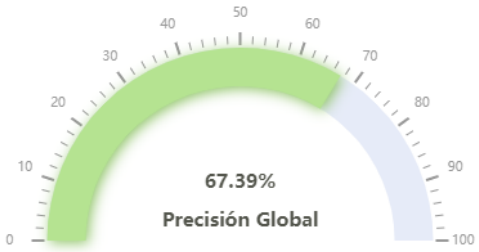
```
Rule number: 5 [chd=Si cover=8 (2%) prob=1.00]
  tobacco< 0.99
  typea>=70.5
```

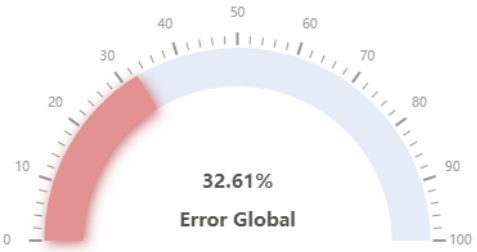
Es una regla aplicada al número 5 de la estructura del árbol con el 2% de toda la base, que nos quiere decir que el pasajero que este dentro de esta regla pude contener un indicador de muerte coronaria dentro de las condiciones vemos que primero vemos tabaco acumulado menor a 0.99 su comportamiento tipo A mayor o igual a 70.5, esta persona tendrá la probabilidad de que si tenga una muerte coronaria con un 100%.

7. Con `Número Máximo de Iteraciones = 75` (`nrounds` que es el número de árboles), con el Amplificador (booster - algoritmo) `gbtree` y `Profundidad Máxima = 20` use el Método de Potenciación (Extreme Boosting) para generar modelos predictivos para la tabla `SAheart.csv`. Para esto utilice el 80 % de los datos para la tabla aprendizaje y un 20 % para la tabla testing, luego calcule para los datos de testing la matriz de confusión, la precisión global y la precisión para cada una de las dos categorías.

Matriz de Confusión







Precisión.No	Precisión.Si
76.67	50.00

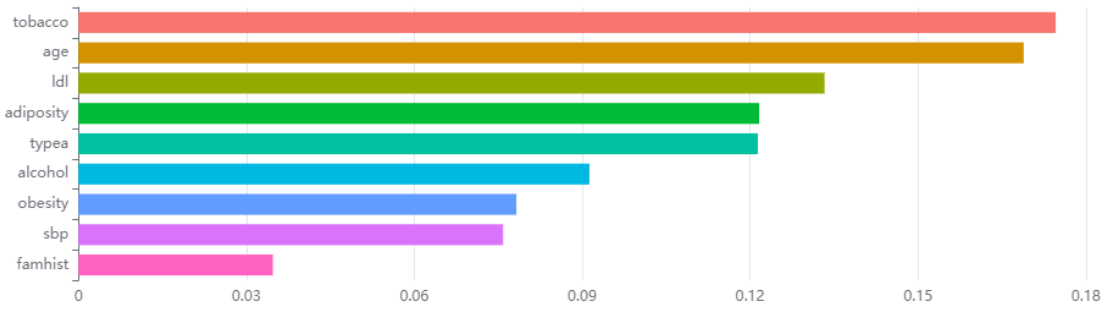
Error.No	Error.Si
23.33	50.00

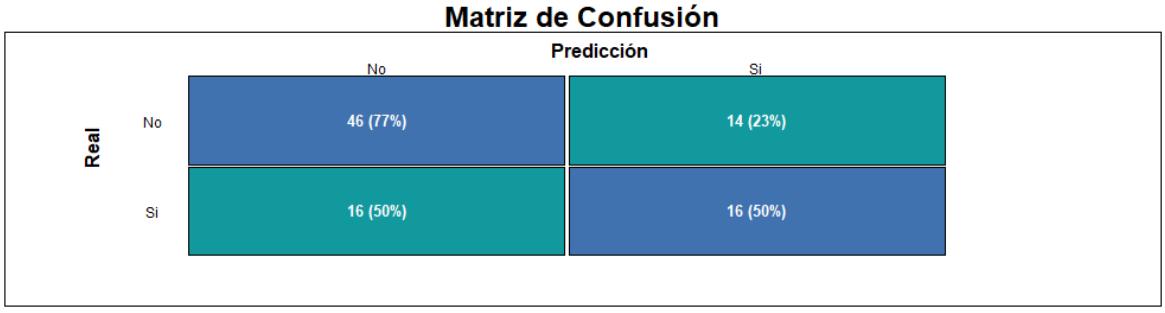
	FN	FP
VN	46	14
VP	16	16

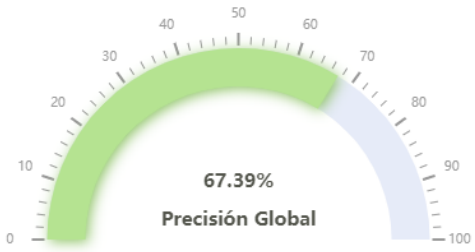
Precision Global=	VN+VP / VN+FP+FN+VP	0,673913043	67,3913043
Precision Positiva=	VP / FN+VP	0,5	50
Precision Negativa=	VN / VN+FP	0,766666667	76,6666667

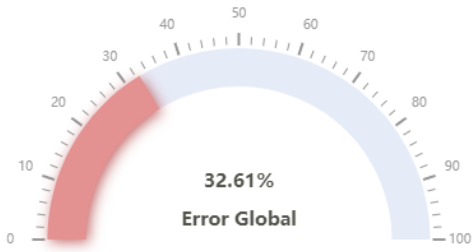
En el caso de la tabla que se muestra en la imagen, la precisión global para ambas variables es de alrededor del 67%. La precisión positiva es de alrededor del 50%, mientras que la precisión negativa es de alrededor del 76%. Esto significa que la medición es más precisa para identificar valores negativos que para identificar valores positivos. La imagen muestra que el modelo tiene una precisión global moderada (0.6739). La precisión positiva es baja (0.50), lo que indica que el modelo tiene una alta tasa de falsos positivos. La precisión negativa es alta (0.7931), lo que indica que el modelo tiene una baja tasa de falsos negativos.

8. Repita el ejercicio anterior pero esta vez usando solamente las 4 variables que según el gráfico de `Importancia de las Variables` tienen mejor poder predictivo. ¿Mejoró el resultado?









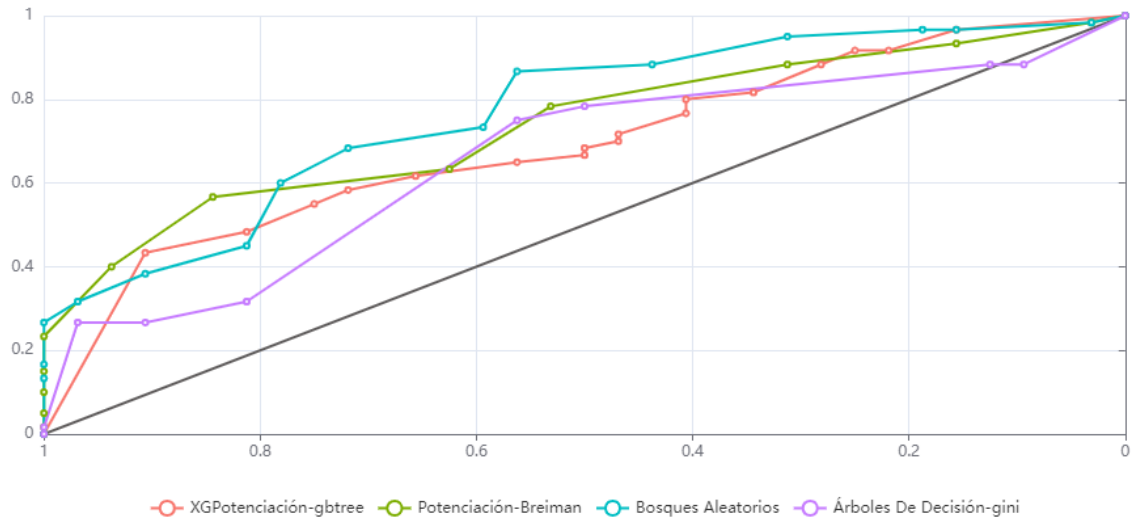
Precisión.No	Precisión.Si
76.67	50.00

Error.No	Error.Si
23.33	50.00

En el caso de la tabla que se muestra en la imagen, la precisión global para ambas variables es de alrededor del 67%. La precisión positiva es de alrededor del 50%, mientras que la precisión negativa es de alrededor del 76%. Esto significa que la medición es más precisa para identificar valores negativos que para identificar valores positivos.

9. Genere la curva ROC para todos los modelos generados en los puntos anteriores, ¿Cuál es el mejor modelo según esta curva? Compare con la curva ROC con las generadas en la tarea anterior. ¿Cuál modelo es mejor?

	Precisión Global	Error Global	No	Si	Área de ROC
XGPotenciación-gbtree	63.04348	36.95652	71.66667	46.875	70.72917
Potenciación-Breiman	65.21739	34.78261	66.66667	62.5	75.15625
Bosques Aleatorios	71.73913	28.26087	80	56.25	77.16146
Árboles De Decisión-gini	68.47826	31.52174	78.33333	50	67.57812

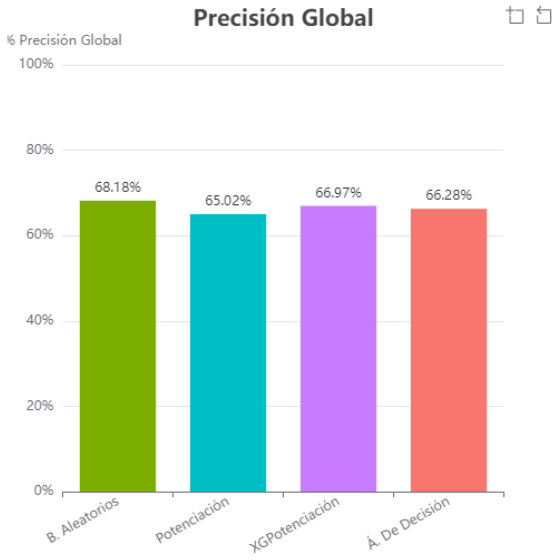


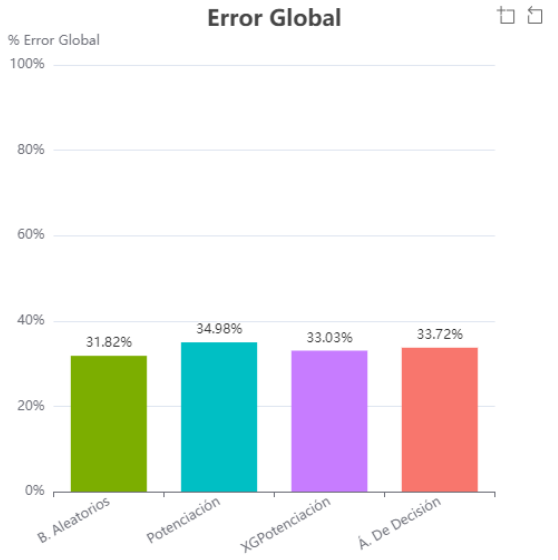
El mejor algoritmo para una tarea específica dependerá de los requisitos específicos de la tarea. Si es importante minimizar el error global, entonces **XGBoost** con potenciación o bosques aleatorios podrían ser buenas opciones. Si es importante minimizar los falsos positivos, entonces **XGBoost** con **Breiman** o árboles de decisión con **Gini** podrían ser mejores opciones. Al observar los resultados, podemos notar que los "***Bosques Aleatorios***" tienen la mayor precisión global y el menor error global entre los modelos presentados. Además, tienen la mayor área bajo la curva ROC, lo que indica un mejor rendimiento en la capacidad de discriminación entre las clases

Validacion Cruzada

ID	Partición	sbp	tobacco	ldl	adiposity
1	Gr_2	160	12	5.73	23.11
2	Gr_2	144	0.01	4.41	28.61
3	Gr_2	118	0.08	3.48	32.28
4	Gr_3	170	7.5	6.41	38.03
5	Gr_2	134	13.6	3.5	27.78
6	Gr_2	132	6.2	6.47	36.21
7	Gr_1	142	4.05	3.38	16.2
8	Gr_5	114	4.08	4.59	14.6
9	Gr_3	114	0	3.83	19.4
10	Gr_5	132	0	5.8	30.96
	Seleccionar muestra				

```
$MCs.dtl
$MCs.dtl[[1]]
      Pred
Clase  No  Si
   No 235  67
   Si  81  79

$MCs.dtl[[2]]
      Pred
Clase  No  Si
   No 226  76
   Si  86  74

$MCs.dtl[[3]]
      Pred
Clase  No  Si
   No 233  69
   Si  80  80

$MCs.dtl[[4]]
      Pred
```





	Precisión Global	Error Global	Precisión No	Precisión Si
Potenciación	65.02165	34.97835	75.43046	45.375
Árboles De Decisión	66.27706	33.72294	76.09272	47.75
Bosques Aleatorios	68.18182	31.81818	82.38411	41.375
XGPotenciación	66.9697	33.0303	78.21192	45.75

El modelo de ***Bosques Aleatorios*** tiene la precisión global más alta (68.18%) y la precisión para la clase negativa más alta (82.38%). Sin embargo, su precisión para la clase positiva es relativamente baja (41.38%). El modelo de ***Árboles de Decisión*** tiene la segunda precisión global más alta (66.28%) y una precisión para la clase positiva un poco más alta que la de ***Bosques Aleatorios*** (47.75%). El modelo ***XG Potenciación*** tiene una precisión global (66.97%) y una precisión para la clase positiva (45.75%) que se sitúan entre las de Bosques Aleatorios y Árboles de Decisión.