TA2 - Enzo Cozza - Agustín Fernández

Ejercicio 1

b)

ID: identificador único del cliente, numérico.
Edad: la edad en años redondeada al entero más cercano.
EstadoCivil: “C” para los casados, “S” para todas las otras alternativas. Binominal.
Sexo: F = femenino; M= masculino. Binominal.
ActividadWebsite: refleja el nivel de actividad en el sitio web: Escasa, Regular o Frecuente. Polinominal.
MiroElectronicos12: indica si la persona hay mirado o no productos electrónicos en el sitio de la compañía (SI / NO) en el último año. Binominal.
ComproElectronicos12: indica si la persona ha comprado o no productos electrónicos en el sitio de la compañía en el último año (SI / NO). Binominal.
ComproMedios18: indica si la persona ha comprado o no productos digitales (ej: MP3) en el sitio de la compañía en el último año y medio (SI / NO). Este atributo NO incluye libros digitales. Binominal.
ComproLibrosDigitales: Martín cree que este atributo puede ser un muy buen indicador del comportamiento de compra para el nuevo eReader, y por ello se lo ha separado de los demás atributos que refieren a compras. En este caso se indica si el cliente alguna vez compró libros digitales, no se restringe sólo al último año. Binominal.
MetodoPago: la forma más frecuente en que el cliente ha efectuado sus pagos: o Transferencia bancaria o CuentaWebsite – el cliente ha dispuesto una tarjeta de crédito o cuenta bancaria para débito automático en el sitio o TarjetaCredito – el cliente ingresa los datos de la tarjeta y autorización en cada compra o DebitoMensual – el cliente realiza compras regularmente y recibe una factura que puede abonar mensualmente. Polinominal.
AdopcionEReader: este atributo existe sólo en el dataset de entrenamiento. Tiene los datos de los clientes que han comprado eReaders de generaciones anteriores. Los que compraron dentro de una semana del lanzamiento son registrados como “Innovadores”. Los que compraron entre una y tres semanas luego del lanzamiento, se registran como “AdoptanteTemprano”. Luego de tres semanas, pero dentro de los primeros 2 meses, se consideran “MayoriaTemprana” y los demás, “MayoríaTardía”. Este atributo servirá como etiqueta al aplicar el modelo al dataset de evaluación. Polinominal.

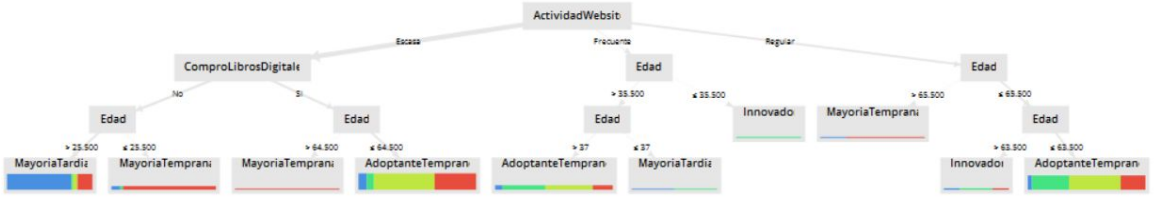
No se registraron outliers ni tampoco datos faltantes.

f)

Criterion: selecciona el criterio en el que los atributos serán seleccionados para su división. ‘Information_gain’: las entropías de los atributos son calculadas y el que tenga la entropía mínima es el elegido para la división. ‘Gain_ratio’: una variante de ‘information_gain’. ‘Gini_index’: una medida de desigualdad entre las distribuciones de las características ‘label’. Dividir en un atributo elegido resulta en una reducción en el índice promedio gini de los subconjuntos restantes. ‘Accuracy’: un atributo es seleccionado para la división, que maximiza la precisión del árbol. ‘Least_square’: un atributo es seleccionado para la división,

que minimiza la distancia cuadrada entre el promedio de los valores y cada uno de los valores.
Maximal depth: se utiliza para restringir la profundidad máxima del árbol de decisión. Si se selecciona -1, no se limita la profundidad del árbol.
Apply pruning: marcar para realizar poda luego de la generación del árbol.
Confidence: este parámetro especifica el nivel de confianza utilizado para el cálculo pesimista del error de la poda.
Apply prepruning: Especifica si se toman más criterios de detención además de la máxima profundidad del árbol. Si este parámetro se activa, se utilizan los criterios mínima ganancia, mínimo tamaño de hoja, mínimo tamaño de división y número de alternativas de prepoda.
Minimal gain: La ganancia de un nodo es calculada antes de dividirlo. El nodo es dividido si su ganancia es mayor a la mínima. Cuanto mayor es el valor de mínima ganancia, se obtiene como resultado menos divisiones y en un árbol de menor tamaño.
Minimal leaf size: Es el número de ejemplos en el subconjunto. El árbol es generado de tal forma que cada hoja tenga por lo menos el valor de este parámetro.
Minimal size for split: el tamaño de un nodo es el número de ejemplos en el subconjunto. Solo se aceptarán los nodos cuyo tamaño sea mayor o igual al número indicado en este parámetro.
Number of prepuning alternatives: cuando la pre poda se aplica a un cierto nodo, este parámetro ajusta la cantidad de alternativas que serán testeadas para la división.

g)

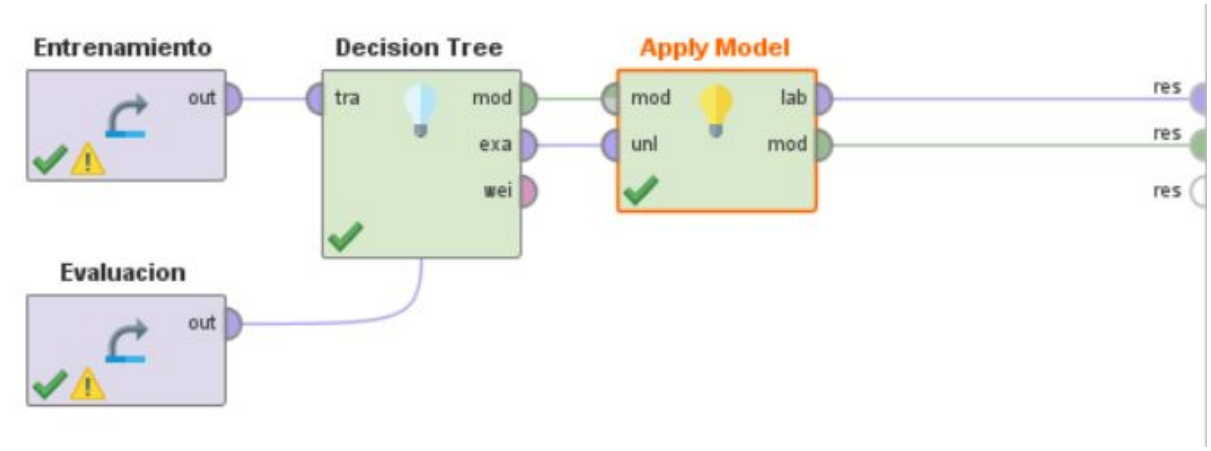


j)
El primer camino se podría decir que sí es intuitivo, ya que quienes tienen poca actividad en la plataforma tienden a tener tiempos de compra mayores al resto, y además se observa que los más jóvenes y los que compraron libros digitales tienen menores tiempos de compra que los otros dentro del mismo sector ‘Escasa’.
Dentro del sector ‘Frecuente’, se entiende que las personas más jóvenes tengan el menor tiempo de compra de los posibles. Sin embargo, no parece muy intuitiva la segunda separación por edad, ya que un rango queda simplemente para la gente que tiene 36 o 37 años. Este problema se repite para el sector ‘Regular’, ya que nuevamente queda un rango de dos años (64, 65).
Esto podría atribuirse a un sobreajuste en el entrenamiento del algoritmo, debido a que se efectúan evaluaciones con los mismos predictores (‘Edad’) y separando el espacio del conjunto de datos en rangos muy pequeños. De esta manera, el CART se vuelve muy específico, quedando cantidades pequeñas de instancias en algunas hojas (2 ítems en el primer caso y 4 en el segundo) .

Si se aumenta la profundidad, ocurre más de lo dicho anteriormente. Al aumentar solo un nivel la profundidad, más de la mitad de las hojas poseen menos de 10 ítems para la

evaluación, lo que seguramente lleve a un sobreajuste del modelo sobre los datos de entrenamiento.

Ejercicio 2



Prediction **prediction(AdopcionEReader)**	Polynominal	0	Open visualizations	Least Innovador (7)	Most AdoptanteTemprano (2...	Values AdoptanteTemprano (287), MayoriaTardia (140), MayoriaTemprana (39), Innovador (7) Details...	
Confidence_MayoriaTardia **confidence(MayoriaTardia)**	Real	0	Open visualizations	Min 0	Max 0.751	Average 0.266	Deviation 0.315
Confidence_Innovador **confidence(Innovador)**	Real	0	Open visualizations	Min 0	Max 1	Average 0.149	Deviation 0.164
Confidence_AdoptanteTemprano **confidence(AdoptanteTempran...**	Real	0	Open visualizations	Min 0	Max 0.515	Average 0.299	Deviation 0.213
Confidence_MayoriaTemprana **confidence(MayoriaTemprana)**	Real	0	Open visualizations	Min 0	Max 1	Average 0.285	Deviation 0.190

7.

De parte del árbol se puede ver que ya no se realizan varias divisiones por ‘Edad’, sino que en los casos que eso ocurría ahora hay una primera división por si la persona compró o no libros digitales. Sin embargo, ‘Edad’ sigue siendo un predictor utilizado para realizar las divisiones.

Esta vez, parece no haber sobreajuste en el entrenamiento del algoritmo, debido a que la cantidad de instancias por hoja es mayor que en el anterior.

Los valores de confianza también cambiaron. Los promedios en su mayoría siguen siendo muy similares, sin embargo en ninguno de los casos se tiene un valor de confianza 1, es decir, nunca se está garantizando que una persona va a pertenecer a una clase. Sin embargo, a pesar de mantener los promedios, se puede observar que hubo un cambio en las predicciones, ya que ‘MayoriaTemprana’ pasó a ser la más frecuente, mientras que ‘AdoptanteTemprano’, que era la más seleccionada en el primer caso, aquí terminó como la tercera más seleccionada. También se ve que las predicciones están más equilibradas.

8-9.

Prof = Profundidad

CEDN = Cantidad de Elementos para Dividir un Nodo

MCEH = Máxima Cantidad de Elementos en la Hoja

Clase Prof-CEDN-MCEH	Gain ratio	Gini index
MayoriaTemprana 4-4-2	39	169
MayoriaTardia 4-4-2	140	137
AdoptanteTemprano 4-4-2	287	123
Innovador 4-4-2	7	44

MayoriaTemprana 4-15-10	40	169
MayoriaTardia 4-15-10	137	137
AdoptanteTemprano 4-15-10	249	131
Innovador 4-15-10	47	36
MayoriaTemprana 5-15-10	40	113
MayoriaTardia 5-15-10	137	137
AdoptanteTemprano 5-15-10	231	164
Innovador 5-15-10	65	59
MayoriaTemprana 5-4-2	42	119
MayoriaTardia 5-4-2	147	137
AdoptanteTemprano 5-4-2	264	147
Innovador 5-4-2	20	70

10.
Cliente ID = 56031

Prof-CEDN-MCEH	Gain ratio	Gini index
	Clase (confianza)	
4-4-2	AdoptanteTemprano (0.439)	AdoptanteTemprano (0.458)
4-15-10	Innovador (0.520)	AdoptanteTemprano (0.458)
5-15-10	Innovador (0.520)	AdoptanteTemprano (0.538)
5-4-2	AdoptanteTemprano (0.444)	Innovador (0.529)