UT04 – TA02

Atributos	Categorías	Distribución
Edad: la edad en años redondeada al entero más cercano.	Variable continua rango: 16-66	
EstadoCivil	S - C	
Sexo	F - M	
ActividadWebsite: refleja el nivel de actividad en el sitio web	Escasa – Frecuente - Regular	
MiroElectronicos12: indica si la persona hay mirado o no productos electrónicos en el sitio de la compañía en el último año	Sí – No	
ComproElectronicos12: indica si la persona ha comprado o no productos electrónicos en el sitio de la compañía en el último año	Sí – No	
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. Este atributo NO incluye libros digitales	Sí – No	
ComproLibrosDigitales: Se indica si el cliente alguna vez compró libros digitales, no se restringe sólo al último año.	Sí - No	
MetodoPago	CuentaWebsite Transferencia TarjetaCredito DebitoMensual	

No se identifican outliers.

Observa los parámetros que tiene el operador, y cada una de sus alternativas. ¿Qué algoritmo de base utiliza? ¿cuáles son los criterios de división que admite? ¿cómo funcionan?

El algoritmo utilizado de base es el CART, y admite como criterios de división:

information_gain: Se calculan las entropías de todos los atributos y aquellos con menor entropía son seleccionados. Este método tiene un sesgo hacia la selección de atributos con muchos valores

gain_ratio: Variación del anterior que ajusta la información para cada atributo para permitir una uniformidad en los valores

gini_index: Medida de la inequidad entre distribuciones de las características.

accuracy: Se selecciona un atributo que maximiza la precisión de todo el árbol

least_square Se selecciona un atributo para la division que minimiza las distancias de los errores cuadrado entre los promedios de los valores de los nodos.

j) Analiza los caminos a las hojas, y observa las reglas correspondientes

a. ¿tienen un sentido intuitivo?

b. Cambia la profundidad máxima y observa los caminos de evaluación

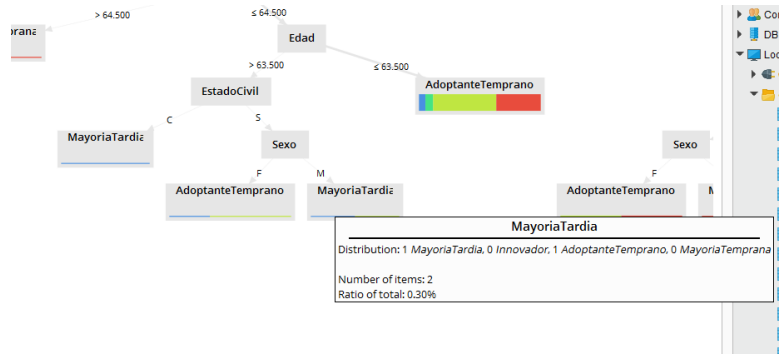
Se observa que sólo la edad y la variable compró libros digitales son relevantes para el modelo cuando se setea la profundidad en 4.

Tiene sentido que los jóvenes sean Innovadores, sería de esperar que la edad fuese un factor determinante a la hora de decidir si comprar y a qué ritmo sobre todo en innovaciones tecnológicas por un tema generacional.

Se observa a su vez que los mismos atributos por ejemplo la edad son utilizados nuevamente en diferentes subramas del árbol.

Al aumentar la profundidad se observa cómo se complejiza el árbol, incorporando nuevos, como ser método de pago, estado civil y sexo.

l) Al posicionarse sobre una hoja, observa la información desplegada (cantidad de ejemplos referenciados en esa hoja, y proporciones de la clase). Vemos que en algunas hojas existen varias posibilidades, y la hoja refiere a la clase mayoritaria. Al aplicar el modelo, veremos que esto se traduce en factores de confianza.



Se observa la distribución, número de ítems y ratio para cada nodo.

Ejercicio 2 – Modelado

Row No.	ID	prediction(A...	confidence(...	confidence(I...	confidence(...	confidence(...	Sexo	Edad	EstadoCivil	ActividadWe...	MiroElectro...
2	25913	AdoptanteTe...	0.037	0.329	0.460	0.174	F	51	C	Regular	Si
3	19396	MayoriaTardia	0.846	0.038	0.038	0.077	M	41	C	Escasa	Si
4	93666	MayoriaTemp...	0.250	0	0	0.750	M	66	S	Regular	Si
5	72282	MayoriaTardia	0.846	0.038	0.038	0.077	F	31	S	Escasa	Si
6	64466	MayoriaTemp...	0.250	0	0	0.750	M	68	C	Regular	Si
7	76655	MayoriaTardia	0.842	0.018	0.061	0.079	F	51	S	Escasa	Si
8	48465	MayoriaTardia	0.500	0.500	0	0	F	36	S	Frecuente	Si
9	19889	AdoptanteTe...	0.037	0.329	0.460	0.174	M	29	C	Regular	Si
10	63570	MayoriaTemp...	0	0	0	1	M	61	C	Frecuente	Si
11	63239	AdoptanteTe...	0.037	0.329	0.460	0.174	M	47	S	Regular	Si
12	67603	MayoriaTemp...	0	0	0	1	F	62	S	Regular	Si
13	65685	AdoptanteTe...	0.037	0.329	0.460	0.174	M	32	C	Regular	Si
14	77373	MayoriaTemp...	0.083	0.028	0	0.889	M	17	C	Escasa	Si
15	54239	AdoptanteTe...	0.037	0.329	0.460	0.174	M	36	S	Regular	Si

Prediction **prediction(AdopcionEReader)**	Polynominal	0	Least Innovador (25)	Most AdoptanteTemprano (2...	Values AdoptanteTemprano (253), MayoriaTardia (146), ...[2 more]
Confidence_MayoriaTardia **confidence(MayoriaTardia)**	Real	0	Min 0	Max 1	Average 0.290
Confidence_Innovador **confidence(Innovador)**	Real	0	Min 0	Max 1	Average 0.150
Confidence_AdoptanteTemprano **confidence(AdoptanteTempran...**	Real	0	Min 0	Max 0.889	Average 0.292
Confidence_MayoriaTemprana **confidence(MayoriaTemprana)**	Real	0	Min 0	Max 1	Average 0.268

Se observan los rangos dentro de los cuales varía la confianza de los predictores.

Confianza	Mínimo	Máximo	Average	Desviación
Mayoría tardía	0	1	0,290	0,356
Innovador	0	1	0,150	0,190
Adoptante temprano	0	0,889	0,292	0,235
Mayoría temprana	0	1	0,268	0,247

Row No.	ID	prediction(AdopcionERe...	confidence(MayoriaTardia)	confidence(Innovador)	confidence(AdoptanteTemprano)	confidence(MayoriaTemprana)
4	93666	MayoriaTemprana	0.250	0	0	0.750
5	72282	MayoriaTardia	0.846	0.038	0.038	0.077
6	64466	MayoriaTemprana	0.250	0	0	0.750
7	76655	MayoriaTardia	0.842	0.018	0.061	0.079
8	48465	MayoriaTardia	0.500	0.500	0	0
9	19889	AdoptanteTemprano	0.037	0.329	0.460	0.174
10	63570	MayoriaTemprana	0	0	0	1
11	63239	AdoptanteTemprano	0.037	0.329	0.460	0.174
12	67603	MayoriaTemprana	0	0	0	1
13	65685	AdoptanteTemprano	0.037	0.329	0.460	0.174
14	77373	MayoriaTemprana	0.083	0.028	0	0.889
15	54239	AdoptanteTemprano	0.037	0.329	0.460	0.174
16	55781	AdoptanteTemprano	0.037	0.329	0.460	0.174
17	19854	MayoriaTemprana	0	0	0	1

En la fila 14 es posible observar que se estima que la instancia pertenecerá a la clase Mayoría Temprana con 0,889 de confianza , mientras que 0,083 y 0,028 a mayoría tardía e Innovador respectivamente, para Adoptante Temprano la confianza es 0 por lo que podemos afirmar que según el modelo no pertenece a dicha clase.

8. Tabla comparativa métodos Gain ratio y Gini index

	Complejidad	Mayoría tardía	Innovador	Adoptante temprano	Mayoría temprana
Gain ratio	5	147	20	264	42
Gini index	5	137	70	147	119
Gain ratio	10	167	62	149	95
Gini index	10	188	77	128	80

Haciendo Split con los datos de entrenamiento para ver cuán bien predice el modelo obtenemos

	Complejidad	Mayoría tardía	Innovador	Adoptante temprano	Mayoría temprana
Valor real	-	52	29	61	56
Gain ratio	5	60	0	116	22
Gini index	5	68	13	83	44
Gain ratio	10	62	31	74	31
Gini index	10	72	28	59	39
Gain ratio	15	66	34	62	36

Si aumentamos Gini index a 15,20, 30 se mantiene el valor de la predicción.

Si aumentamos Gain ratio a 20, 30, 40 se mantiene el valor de la predicción.

Experimentar variando los valores de los parámetros de máxima profundidad, cantidad de elementos para dividir un nodo y máxima cantidad de elementos en las hojas, y registrar los diversos resultados en la tabla.

Leaf size disminuido a 1 y minimal size for split 2

	Complejidad	Mayoría tardía	Innovador	Adoptante temprano	Mayoría temprana
Valor real	-	52	29	61	56
Gain ratio	5	61	2	116	29
Gini index	5	57	14	81	46
Gain ratio	10	60	28	60	50
Gini index	10	70	29	60	39
Gain ratio	15				

lustrar cómo los cambios en los parámetros han afectado las predicciones para determinados clientes: por ejemplo, seleccionar y estudiar para el Cliente con ID = 56031

El cliente ID=56031 es clasificado en un 0,565 y 0,423 de confianza como adoptante temprano utilizando gain ratio y Gini index respectivamente con 5 de complejidad y un mínimo valor de división de nodo 2 y mínimas instancias por nodo 1.

Al pasar a 10 de complejidad, utilizando Gini index es clasificado como innovador con 0,5 mientras que también adoptando temprano asume el valor 0,5, por tanto la elección resulta ser aleatoria, cuando se selecciona Gain ratio con los mismos parámetros de Split y tamaño de nodo sigue seleccionando como categoría de salida Adoptante temprano.

Ejercicio 3

¿cómo se puede utilizar ahora este modelo en producción?

• Consultoría (describir)

• Software que recomiende / prediga (describir funcionamiento)

Este modelo podría utilizarse para recomendar por ejemplo en una estrategia de marketing a qué grupo dirigirse a la hora de publicitar un nuevo producto en qué etapa del tiempo. Así como también analizar que funcionalidades disponibilizar primero (las que más interesen a los innovadores dejando para luego a la mayoría tardía).

El software que recomiende o prediga debería solicitar las características de la persona en base a los atributos considerados relevantes para el análisis como ser edad, método de pago, sexo, estado civil, si compró alguna vez libros digitales y si compró electrónicos en el último año y utilizando dicha información generar un modelo de predicción de en qué categoría es probable que se encuentre esa persona.