UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA
FACULTAD DE ECONOMÍA Y PLANIFICACIÓN
TITULACIÓN POR EXAMEN PROFESIONAL

"ESTUDIO DE LAS PRINCIPALES VARIABLES QUE DETERMINAN EL CONSUMO DE UNA MARCA DE BEBIDA GASEOSA USANDO TÉCNICAS DISCRIMINANTES"

TRABAJO MONOGRÁFICO PRESENTADO POR

ROSARIO JESSICA ALCEDO ZAMBRANO

PARA OPTAR EL TÍTULO PROFESIONAL DE

INGENIERO EN ESTADÍSTICA E INFORMÁTICA

Lima – Perú

2019

UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA
FACULTAD DE ECONOMÍA Y PLANIFICACIÓN
TITULACIÓN POR EXAMEN PROFESIONAL
"ESTUDIO DE LAS PRINCIPALES VARIABLES QUE DETERMINAN EL CONSUMO DE UNA MARCA DE BEBIDA GASEOSA USANDO TÉCNICAS DISCRIMINANTES"

TRABAJO MONOGRÁFICO PRESENTADO POR

ROSARIO JESSICA ALCEDO ZAMBRANO

PARA OPTAR EL TÍTULO PROFESIONAL DE

INGENIERO EN ESTADÍSTICA E INFORMÁTICA

SUSTENTADO Y APROBADO ANTE EL SIGUIENTE JURADO:

……

Dr. Rino Nicanor Sotomayor Ruiz

Presidente

…….. ……

Ing. Julio Hugo Ángeles Olivera

Miembro

Mg. Clodomiro Fernando Miranda Villagómez

Miembro

Lima – Perú

2019

AGRADECIMIENTO

A mi madre y hermana

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

La investigación de mercados y el uso de la estadística básica y avanzada en los últimos tiempos han pasado a ser una de las herramientas más importantes para los Jefes y Gerentes de Producto, un claro ejemplo de ello es el presente estudio, que a través del análisis multivariante busca resolver algunas inquietudes.

El presente estudio, tiene como finalidad construir un modelo que permita predecir si un consumidor de gaseosas consume la Marca X. El modelo será utilizado para identificar a sus consumidores y conocer sus expectativas, evaluar sus gustos/ preferencias y medir el impacto de la publicidad y las promociones.

La data con la que se construyó el modelo, fue proporcionada por la Empresa X, ellos cuentan con información pasada y presente de una serie de variables que son de su interés. Lo que se buscó con este estudio, es aprovechar dicha información para realizar el modelo.

Para el análisis se utilizaron dos técnicas discriminantes, Análisis de Regresión Logístico y Arboles de decisión, ambos modelos se construyeron con uso del Enterprise Miner (SAS). El resultado final, fue un árbol de decisión con seis variables independientes, se trata de un modelo fácil de interpretar y aplicar.

I. INTRODUCCIÓN

En el pasado, el mercado de bebidas gaseosas se reducía a un número pequeño de marcas y la guerra entre ellas, se enfocaba principalmente en la publicidad, en los canjes y las promociones que ofrecían a sus consumidores, que en la mayoría de los casos eran consumidores fieles a su marca. El precio nunca fue un variable en discusión, ya que todas costaban casi lo mismo.

En los últimos años, el mercado de gaseosas sufrió cambios importantes con la aparición de nuevas marcas, las cuales se introdujeron en el mercado con precios bajos, diferentes presentaciones de tamaño, de tipos de envases y variedad de sabores. Ante estos cambios, surge la llamada guerra de precios, en la que las variables principales de diferenciación pasaron a hacer el precio de producto y la cantidad es decir la presentación.

Ante los hechos mencionados, la investigación de mercados jugó un papel importante y paso a ser una de las herramientas más útil para los jefes de producto. De este modo, los estudios cuantitativos de seguimiento de mercado se vuelven primordiales para la toma de decisiones porque a través de ellos se estima la participación de mercado, miden la recordación de marca, la preferencia de consumo así como la imagen y el posicionamiento, esta información la siguen con frecuencia mensual. Por otro lado, los estudios cualitativos son utilizados porque les permite evaluar las promociones, la publicidad que diseñan y conocer las necesidades de los consumidores.

La Empresa X, dedica a la elaboración de bebidas gaseosas tiene como marca principal a la Marca X; esta empresa en los últimos años viene realizando una encuesta de seguimiento a su mercado a través de un cuestionario estructurado, el cual les permite levantar información relevante del mercado de bebidas gaseosas, con la finalidad de observar la evolución en el tiempo de sus principales indicadores. Por otro lado, frecuentemente recopila información cualitativa de sus consumidores mediante entrevistas personales y focus groups con el

objetivo de evaluar sus campañas publicitarias y promociones.

La Empresa X, tiene la necesidad de conocer mejor a sus consumidores, esta interesada en saber que variables de las mide mensualmente son las que determinan o influyen significativamente en el consumo de la marca X y de esta manera poder tomar acciones comerciales sobre dichas variables.

También le interesa identificar y conocer lo mejor posible el perfil de sus consumidores, a través de los estudios cualitativos que realiza. Esta inquietud la podrían resolver preguntándole al consumidor ¿Qué marca de gaseosa consume?, sin embargo, como en la actualidad un consumidor de gaseosas tiende a consumir más de una marca de gaseosa, esto hace difícil identificar al verdadero consumidor de una marca determinada. Bajo este esquema se plantea lo siguiente; desarrollar un modelo estadístico que nos permita resolver los problemas planteados y crear una herramienta que permita aplicar el modelo de formar sencilla y práctica.

Dentro de las limitaciones que se ha tenido en el estudio, se puede mencionar que la Empresa X, busco aprovechar información que viene recopilando hace unos años (Estudio de seguimiento del mercado de bebidas gaseosas), y decidió utilizar esa información para cumplir con los objetivos de este estudio, por lo tanto se trabajó con información proporcionada por la Empresa X.

1.1. Objetivos generales

Construir un modelo que permita predecir si un consumidor de gaseosa consume la Marca X.

1.2. Objetivos específicos

- Construir un modelo de regresión logístico para predecir si un consumidor de gaseosa consume la Marca X.
- Construir un árbol de decisión para predecir si un consumidor de gaseosa consume

la Marca X.

- Comparar ambos modelos y elegir el mejor.
- Determinar cuáles son las variables que determinan el consumo de la Marca X.
- Elaborar una herramienta que permita la fácil aplicación del modelo.

II. METODOLOGÍA

2.1. Diseño y selección de la muestra

Como se mencionó anteriormente los datos provienen de una encuesta de seguimiento del mercado de bebidas gaseosas, la cual es aplicada mensualmente, el tamaño de la muestra mensual es de 400 casos, los cuales son distribuidos representativamente en función a la distribución que presentan los consumidores de bebidas gaseosas de Lima Metropolitana.

Se realizaron encuestas en hogares a hombres y mujeres de 12 a 45 años de edad, pertenecientes a los niveles socioeconómicos alto, medio, bajo y muy bajo.

El tamaño de muestra es de 400 casos (mensuales), lo cual permite trabajar con un error muestral de 4.90% con p=0.50 (asumiendo máxima dispersión en la muestra).

Se realizó, un muestreo estratificado por nivel socioeconómico, con selección aleatoria de manzanas por computadora y selección sistemática de viviendas al interior de cada manzana.

Para la construcción de los modelos se utilizó una muestra de 4800 casos, correspondiente a un año de historia (Enero – Diciembre 2004).

2.2. Definición y descripción de variables

Como mencionamos anteriormente el cuestionario utilizado, no fue elaborado específicamente para cumplir con los objetivos del estudio, sino se buscó aprovechar información ya disponible y proporcionada por la Empresa X.

La herramienta que utilizada en el estudio de seguimiento de mercado de bebidas gaseosa es un cuestionario estructurado que contienen preguntas de recordación, de consumo, de percepción de atributos, de preferencia de sabores, de recordación publicitaria y evaluación de presentaciones.

La Información proporcionada por la Empresa X se muestra en la Tabla 1.

Tabla 1: Información de las variables del cuestionario de seguimiento de mercado de bebidas gaseosas

Pregunta	Tipo variable	Tipo de respuesta	Nombre de Variable
Nivel Socio Económico	Nominal	Simple	Nse
Sexo	Nominal	Simple	Sexo
Edad	intervalo	Simple	Edad
¿Cuál es la primera marca que se le viene a la mente?	Nominal	Simple	Re_marca
¿Qué otras marcas le vienen a la mente?	Nominal	Múltiple	Re_marca
¿Aparte de las marcas que me ha mencionado, de la siguiente lista? ¿que otras marcas conoce?	Nominal	Múltiple	Re_marca
¿De cuáles de estas marcas ha visto u oído/escuchado publicidad recientemente?	Nominal	Múltiple	Re_publicidad
¿Cuál es la marca que toma usted con mayor frecuencia? (máximo 3 respuestas)	Nominal	Múltiple	Target
¿Cuál es la marca de bebida gaseosa que es su favorita	Nominal	simple	Favorita
De las siguientes marcas (mostrar tarjeta) ¿Cuál o cuales considera usted que posee los siguientes atributos?			
Es refrescante	Nominal	Múltiple	Atributo_1
Es de gran calidad	Nominal	Múltiple	Atributo_2
Es para alguien como tu	Nominal	Múltiple	Atributo_3
Tiene envases atractivos /adecuados	Nominal	Múltiple	Atributo_4
Produce orgullo consumir	Nominal	Múltiple	Atributo_5
Tiene una buena relación entre calidad y precio	Nominal	Múltiple	Atributo_6
Es una marca confiable y con prestigio	Nominal	Múltiple	Atributo_7
Quita la sed	Nominal	Múltiple	Atributo_8
Es cada vez mas popular/conocida	Nominal	Múltiple	Atributo_9

«continuación»

Tiene buen sabor	Nominal	Múltiple	Atributo_10
Tiene publicidad con la que te identificas	Nominal	Múltiple	Atributo_11
Es moderna	Nominal	Múltiple	Atributo_12
Es económica / barata	Nominal	Múltiple	Atributo_13
Se encuentra en todas partes	Nominal	Múltiple	Atributo_14
Trae mas cantidad	Nominal	Múltiple	Atributo_15
Buena para tomar con las comidas	Nominal	Múltiple	Atributo_16
Mejoró / cambio recientemente	Nominal	Múltiple	Atributo_17
Tiene mejor sabor que otras marcas	Nominal	Múltiple	Atributo_18
Tiene la cantidad apropiada de dulce	Nominal	Múltiple	Atributo_19
Para los que sacan lo mejor de cada situación	Nominal	Múltiple	Atributo_20
Te levanta el animo	Nominal	Múltiple	Atributo_21
Para compartir con amigos	Nominal	Múltiple	Atributo_22
Es una marca que esta innovando siempre	Nominal	Múltiple	Atributo_23
Es irreverente	Nominal	Múltiple	Atributo_24
¿Con qué frecuencia consume usted gaseosas?	Ordinal	Simple	Gaseosa
¿Con qué frecuencia consume usted agua mineral?	Ordinal	Simple	Mineral
¿Con qué frecuencia consume usted agua embotellada con gas?	Ordinal	Simple	Con_gas
¿Con qué frecuencia consume usted agua embotellada sin gas?	Ordinal	Simple	Sin _gas
¿Con qué frecuencia consume usted jugo líquido?	Ordinal	Simple	Jugo
¿Con qué frecuencia consume usted refresco en polvo?	Ordinal	Simple	Refresco
¿Con qué frecuencia consume usted bebidas deportivas?	Ordinal	Simple	Deportivas

FUENTE: Elaboración propia.

Para realizar los modelos estadísticos, las variables de la Tabla 1 fueron recodificadas como se muestra en la Tabla 2. En el caso de variables con más de dos categorías, el software se encargó de crear las variables ficticias.

Tabla 2: Re-codificación de variables

Nombre Variable	Descripción	Tipo	Valor	Etiquetas
Nse	Nivel Socio Económico	Ordinal	1	Alto
			2	Medio
			3	Bajo
			4	Muy bajo
Sexo	Sexo	Binaria	1	Masculino
			2	Femenino
Edad	Edad	Intervalo	12-45	
Re_marca	Recordación de marca		1	Recuerda la marca X espontáneamente en 1era mención.
			2	Recuerda la marca X espontáneamente en otras menciones.
			3	Recuerda la marca X con ayuda.
Re_publicidad	Recordación de Publicidad	Dicotómica	1	Si recuerda publicidad de marca X.
			0	No recuerda publicidad de marca X.
Target	Consume frecuentemente la marca	Dicotómica	1	Consumidor de la marca X
			0	No Consumidor de la marca X
Favorita	Marca favorita	Dicotómica	1	La marca X es su favorita.
			0	La marca X no es su favorita.
Evaluación de Atributos		Dicotómica	1	La marca X posee el atributo en mención.
			0	La marca X NO posee el atributo en mención.
Frecuencia Consumo de …		Ordinal	1	Más de una vez al día
			2	Una vez al día
			3	4/6 veces por semana
			4	2/3 veces por semana
			5	Una vez a la semana
			6	Una vez en quince días
			7	Una vez al mes
			8	Menos de una vez al mes
			9	Nunca

FUENTE: Elaboración propia.

2.3. Análisis estadístico

Con el uso de la herramienta Software Estadístico SAS (Enterprise Miner), se elaboraron los modelos usando dos diferentes técnicas arboles de decisión y análisis de regresión Logístico. La ventaja de la herramienta Enterprise Miner, es que nos permite de manera sencilla obtener lo siguiente:

- Explorar la data a través de estadísticas descriptivas.
- Permite dividir la muestra, en muestra de entrenamiento la cual se utiliza para elabora los modelos y muestra de validación la cual permite validar los resultados de la muestra de entrenamiento.
- Para modelos de regresión, permite aplicar las diferentes técnicas de selección de variables y compararlas con la finalidad de elegir la mejor opción.
- Para el árbol de decisión, permite comparar arboles con diferentes características tales como número de ramas, profundidad del árbol, métodos entre otros.
- Finalmente, ayuda a escoger el mejor modelo que se ajusta a los datos, mediante la comparación de los modelos seleccionados.

En la Figura 1, se muestra la interfase del Enterprise Miner que se utilizó para el estudio.

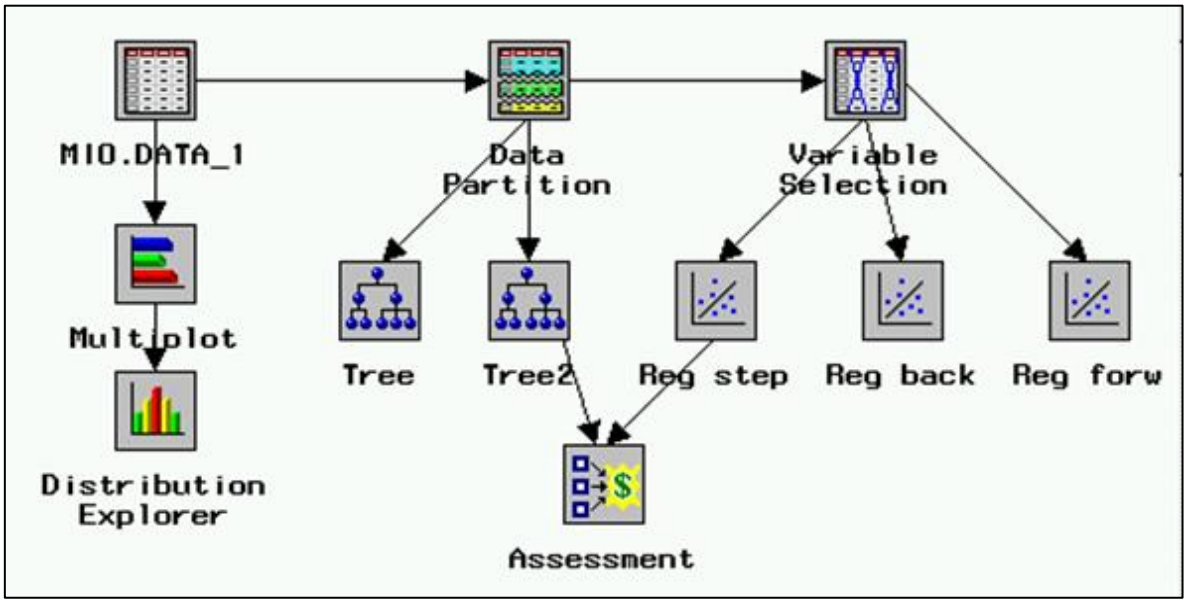


Figura 1: Interfase del *enterprise miner*

2.4. Marco teórico

A continuación se presenta los conceptos de las técnicas que vamos a utilizar.

La Regresión Logística es un método estadístico multivariante que describe la probabilidad de que un proceso ocurra o no ocurra en función de un número determinado de factores cualitativos y/o cuantitativos.

Los Arboles de decisión, es la representación gráfica de las diferentes opciones que se consideran en el análisis de decisión. Es una estructura en forma de árbol que representa un conjunto de decisiones. Estas decisiones generan reglas para la clasificación de un conjunto de datos.

Se utilizó, el método de Chaid (Detección de interacción automática de Chi cuadrado) técnica de árbol de decisión usada para la clasificación de un conjunto da datos. Provee un conjunto de reglas que se pueden aplicar a un nuevo conjunto de datos sin clasificar con el objetivo de predecir cuáles registros darán un cierto resultado.

III. RESULTADOS

A continuación se presenta los principales resultados.

3.1. Modelo de Regresión Logístico

Se buscó ajustar los datos a un modelo de regresión logístico debido a que la variable dependiente es dicotómica.

Con la finalidad de reducir el número de variables dependientes (37) y modelar con las variables más relevantes, el Enterprise Miner, brinda la opción de un nodo de selección de variables, el cual se recomienda usar antes del nodo de regresión (ver figura1).

La función del nodo, se basa en la reducción del número de variables usando el criterio R-square (coeficiente de determinación).

A continuación los resultados:

Tabla 3: Coeficiente de determinación de las variables independientes

R-Squares for Target Variable							
Effect	**DF**	**R-Square**		**Effect**	**DF**	**R-Square**	
lass: Atributo_3	1	0.3436		Class: Nse	3	0.0096	
Class: Favorita	1	0.1983		Class: Gaseosa	6	0.0088	
Class: Re_marca	2	0.1594		Group: Gaseosa	4	0.0086	
Class: Atributo_16	1	0.1165		AOV16: Edad	15	0.0079	
Class: Atributo_18	1	0.1063		Class: Re_publicidad	1	0.0053	
Class: Atributo_22	1	0.1048		Class: Sin__gas	9	0.0050	
Class: Atributo_10	1	0.1009		Group: Sin__gas	6	0.0049	R2 < MINR2
Class: Atributo_5	1	0.0888		Class: Atributo_2	1	0.0045	R2 < MINR2
Class: Atributo_14	1	0.0834		Class: Refresco	9	0.0042	R2 < MINR2
Class: Atributo_1	1	0.0823		Group: Refresco	7	0.0041	R2 < MINR2
Class: Atributo_21	1	0.0691		Class: Mineral	9	0.0040	R2 < MINR2
Class: Atributo_8	1	0.0656		Group: Mineral	6	0.0040	R2 < MINR2
Class: Atributo_19	1	0.0563		Class: Jugo	9	0.0038	R2 < MINR2
Class: Atributo_23	1	0.0546		Group: Jugo	6	0.0038	R2 < MINR2
Class: Atributo_11	1	0.0456		Class: Deportivas	9	0.0028	R2 < MINR2
Class: Atributo_6	1	0.0432		Class: Atributo_15	1	0.0028	R2 < MINR2
Class: Atributo_17	1	0.0427		Group: Deportivas	7	0.0028	R2 < MINR2
Class: Atributo_20	1	0.0410		Var: Edad	1	0.0022	R2 < MINR2
Class: Atributo_7	1	0.0376		Class: Sexo	1	0.0021	R2 < MINR2
Class: Atributo_4	1	0.0310		Class: Con_gas	9	0.0021	R2 < MINR2
Class: Atributo_9	1	0.0309		Group: Con_gas	7	0.0020	R2 < MINR2
Class: Atributo_13	1	0.0278		Class: p13	1	0.0011	R2 < MINR2
Class: Atributo_12	1	0.0106		Class: Atributo_24	1	0.0001	R2 < MINR2

FUENTE: Elaboración propia.

En la Tabla 3, se muestran los coeficientes de determinación (R-Square) para cada una de las variables independientes, esto se realizó con la finalidad de determinar qué porcentaje de la variabilidad de la variable dependiente (consumo de la Marca X) es explicada por cada una de las variables independientes.

En esta parte del análisis se establece lo siguiente, si el R-Square es menor a 0.005 la variable es excluida, mientras que las variables con R-Square iguales o mayores a 0.005 pasan a formar parte del análisis de variancia, con el objetivo de probar la significancia de cada una de las variables, los resultados se muestran a continuación:

Tabla 4: Variables significativas con alfa=0.10

Effects Chosen for Target						
Effect	**DF**	**R-Square**	**F Value**	**p-Value**	**Sum of Squares**	**Error Mean Square**
Class: Atributo_3	1	0.3436	1506.7349	<.0001	166.0081	0.1102
Class: Re_marca	2	0.0588	141.5615	<.0001	28.4179	0.1004
Class: Favorita	1	0.0251	125.8158	<.0001	12.1032	0.0962
Class: Atributo_4	1	0.0171	88.2189	<.0001	8.2366	0.0934
Class: Atributo_8	1	0.0084	43.8921	<.0001	4.0378	0.0920
Class: Atributo_16	1	0.0063	33.5772	<.0001	3.0542	0.0910
Class: Atributo_18	1	0.0038	20.0531	<.0001	1.8120	0.0904
Class: Atributo_7	1	0.0028	15.1821	<.0001	1.3651	0.0899
Class: Atributo_6	1	0.0031	16.6681	<.0001	1.4906	0.0894
Class: Atributo_22	1	0.0020	10.8984	0.001	0.9713	0.0891
Class: Atributo_21	1	0.0015	8.1983	0.0042	0.7288	0.0889
Group: Gaseosa	4	0.0014	1.9482	0.0998	0.6919	0.0888
Class: Atributo_1	1	0.0009	4.9514	0.0261	0.4390	0.0887
Class: Atributo_14	1	0.0006	3.4878	0.0619	0.3090	0.0886
Class: Nse	3	0.0005	0.9936	0.3948	0.2640	0.0886
Class: Atributo_12	1	0.0005	2.8658	0.0906	0.2537	0.0885

FUENTE: Elaboración propia.

En la Tabla 4, se muestra las variables más significativas con alfa=0.10. Estas fueron las variables que pasaron a forma parte del modelamiento en la regresión logística.

Tabla 5: Variables no significativas con alfa=0.10

Effects Not Chosen for Target					
Effect	**DF**	**R-Square**	**F Value**	**p-Value**	**Sum of Squares**
Class: Atributo_10	1	0.000	0.004	0.949	0.000
Class: Atributo_5	1	0.000	0.642	0.423	0.057
Class: Atributo_19	1	0.000	0.606	0.436	0.054
Class: Atributo_23	1	0.000	2.337	0.126	0.207
Class: Atributo_11	1	0.000	0.580	0.446	0.051
Class: Atributo_17	1	0.000	1.370	0.242	0.121
Class: Atributo_20	1	0.000	2.291	0.130	0.203
Class: Atributo_9	1	0.000	1.286	0.257	0.114
Class: Atributo_13	1	0.000	0.858	0.354	0.076
Class: Re_publicidad	1	0.000	0.132	0.716	0.012

FUENTE: Elaboración propia.

La Tabla 5, muestra las variables no significativas con alfa=0.10, estas variables fueron excluidas del proceso de modelamiento.

Con las 16 variables seleccionas por el criterio de selección de variables, se construyó el modelo de regresión logístico utilizando el método de Stepwise. El modelo final se obtuvo después de 12 pasos y considerando 12 variables dependientes. Los resultados se presentan a continuación:

- Prueba de hipótesis para todos los coeficientes de regresión:
  La Figura 2 hace inferencia sobre lo siguiente.

  Ho: $®_1=®_2=®_3=®_4=®_{15}=®_{16}=®_{17}=®_{18}=®_{19}=®_{10}=®_{11}=®_{12}=0$

  Ha: Al menos una es diferente 0

  $X^2_{calculada} = 1396.931 < X^2_{(13, 0.05)}$

  P_value = <0.0001

Se rechaza Ho. Por tanto se puede afirmar que a un nivel de significación de 0.05 existe evidencia estadística para afirmar que al menos uno de los ® es diferente de cero.

****

Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
-2 LOG L	2984.593	1587.662	1396.931 with 13 DF (p<.0001)

****

Figura 2: Prueba de hipótesis para todos los coeficientes de regresión

- Prueba de hipótesis para cada coeficiente de regresión:

  En esta parte se evalúa individualmente la significancia de cada uno de los parámetros de las variables dependientes. A continuación se interpreta los resultados de la variable Re_marca (Recordación la marca):

  Ho: $®_1=0$

  Ha: $®_1 \neq 0$

  Wald Chi-Square =106.19 $< X^2_{(2, 0.05)}$

  P_value = <0.0001

  Se rechaza Ho. Por lo tanto se puede afirmar que a un nivel de significación de 0.05 existe evidencia estadística para afirmar que ®1 es diferente de cero. Por tanto la variable recordación de marca contribuye significativamente al modelo.

  De manera similar se puede concluir para cada una de las variables que se presentan en la Figura 3; se rechaza la hipótesis planteada a un nivel de significación de 0.05.

****

Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
-2 LOG L	2984.593	1587.662	1396.931 with 13 DF (p<.0001)

****

Figura 3: Prueba de hipótesis para cada coeficiente de regresión

- Prueba de hipótesis para cada coeficiente de regresión:

  En esta parte se evalúa individualmente la significancia de cada uno de los parámetros de las variables dependientes. A continuación se interpreta los resultados de la variable Re_marca (Recordación la marca):

  Ho: $®_1=0$

  Ha: $®_1 \neq 0$

  Wald Chi-Square =106.19 $< X^2_{(2, 0.05)}$

  P_value = <0.0001

  Se rechaza Ho. Por lo tanto se puede afirmar que a un nivel de significación de 0.05 existe evidencia estadística para afirmar que ®1 es diferente de cero. Por tanto la variable recordación de marca contribuye significativamente al modelo.

  De manera similar se puede concluir para cada una de las variables que se presentan en la Figura 4; se rechaza la hipótesis planteada a un nivel de significación de 0.05.

****

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > Chi-Square
Re_marca	2	106.1942	<.0001
Atributo_7	1	16.2721	<.0001
Atributo_3	1	287.9600	<.0001
Favorita	1	24.4118	<.0001
Atributo_18	1	4.5643	0.0326
Atributo_16	1	6.2474	0.0124
Atributo_14	1	5.0137	0.0251
Atributo_8	1	32.6724	<.0001
Atributo_6	1	15.9975	<.0001
Atributo_4	1	34.2828	<.0001
Atributo_22	1	9.5054	0.0020
Atributo_21	1	5.1028	0.0239

****

Figura 4: Prueba de hipótesis para cada coeficiente de regresión

- Estimación e interpretación de parámetros:

  En la Figura 5, se muestran los resultados del análisis máximo verosimilitud en el cual se estiman los paramentos y se realizan las pruebas de Chi-square para cada una de las variables dependientes, lo cual nos permite conocer la significancia de cada una de las variables en el modelo final.

****

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-square	Pr > Chi-square	exp(Est)
Intercept		1	1.8517	0.3593	26.56	<.0001	6.371
Re_marca	1	1	1.5874	0.1891	70.49	<.0001	4.891
Re_marca	2	1	-0.0596	0.1180	0.25	0.6136	0.942
Atributo_7	0	1	-0.2820	0.0699	16.27	<.0001	0.754
Atributo_3	0	1	-1.2010	0.0708	287.96	<.0001	0.301
Favorita	0	1	-1.7133	0.3468	24.41	<.0001	0.180
Atributo_18	0	1	-0.1632	0.0764	4.56	0.0326	0.849
Atributo_16	0	1	-0.2009	0.0804	6.25	0.0124	0.818
Atributo_14	0	1	-0.1512	0.0675	5.01	0.0251	0.860
Atributo_8	0	1	-0.5686	0.0995	32.67	<.0001	0.566
Atributo_6	0	1	-0.2739	0.0685	16.00	<.0001	0.760
Atributo_4	0	1	-0.4739	0.0809	34.28	<.0001	0.623
Atributo_22	0	1	-0.2427	0.0787	9.51	0.0020	0.785
Atributo_21	0	1	-0.1633	0.0723	5.10	0.0239	0.849

****

Figura 5: Resultados del análisis máximo verosimilitud

Las interpretaciones de cada uno de los parámetros se presentan en la Tabla 6.

Tabla 6: Interpretación de parámetros

Parámetro	Valor	Descripción	Exp(Est)	Interpretación del coeficiente
Intercepto			6.371	
Re_marca	1	Recordación de marca	4.891	Es 4.891 veces más probable que un consumidor de gaseosa consuma la marca x si recuerda la marca en primera mención espontánea, frente a uno que no la recuerda espontáneamente.
Re_marca	2	Recordación de marca	0.942	Es 0.942 veces menos probable que un consumidor de gaseosa consuma la marca x si recuerda la marca espontánea pero no en primera mención, frente a uno que no la recuerda espontáneamente.
Atributo_7	0	Es una marca confiable y con prestigio	0.754	Es 0.754 veces menos probable que un consumidor de gaseosa consuma la marca X si no considera a la marca X como confiable y de prestigio, frente a un consumidor que sí considera que es confiable y de prestigio.
Atributo_3	0	Es para alguien como tú	0.301	Es 0.301 veces menos probable que un consumidor de gaseosa consuma la marca X si no considera a la marca X es para alguien como el/ella, frente a un consumidor que si considera que es para el/ella.
Favorita	0	Marca favorita	0.180	Es 0.180 veces menos probable que un consumidor de gaseosa consumo la marca X si no la considera su marca favorita frente, a un consumidor que si la considera su marca favorita.
Atributo_18	0	Tiene mejor sabor que otras marcas	0.849	Es 0.849 veces menos probable que un consumidor de gaseosa consuma la marca x si no considera que tiene mejor sabor que otras marcas frente, a un consumidor que si considera que tiene mejor sabor que otras marcas.
tributo_16	0	Buena para tomar con las comidas	0.818	Es 0.818 veces menos probable que un consumidor de gaseosa consuma la marca x si no considera que es buena para tomar con las comidas, frente a un consumidor que si considera que es buena para tomar con las comidas.
Atributo_14	0	Se encuentra en todas partes	0.860	Es 0.860 veces menos probable que un consumidor de gaseosa consuma la marca x si no considera que la marca X se encuentra en todas partes, frente a un consumidor que si considera que se encuentra en todas partes.
Atributo_8	0	Quita la sed	0.566	Es 0.566 veces menos probable que un consumidor de gaseosa consuma la marca x si no considera que la marca X quita la sed, frente a un consumidor que si considera que quita la sed.

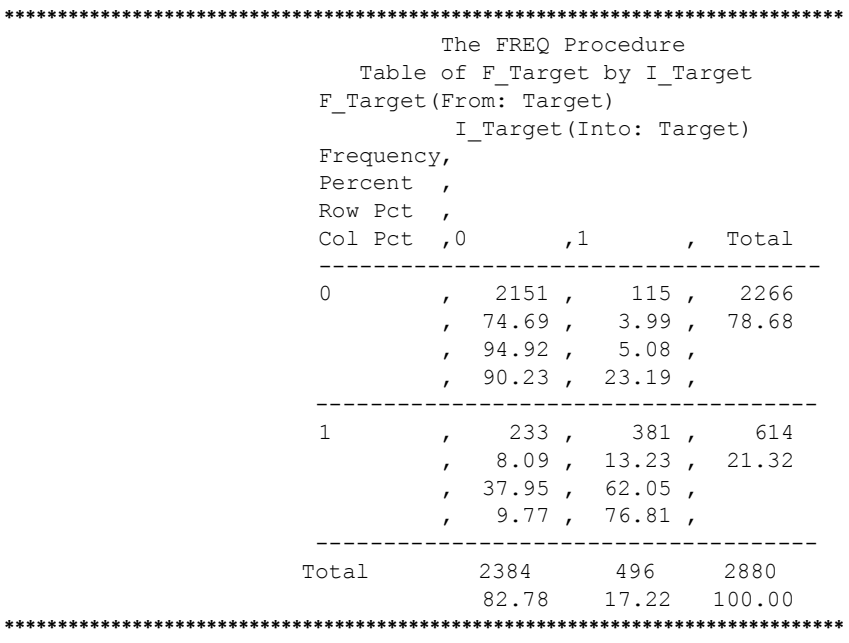
«continuación»

Atributo_6	0	Tiene una buena relación entre calidad y precio	0.760	Es 0.760 veces menos probable que un consumidor de gaseosa consuma la marca x si no considera que la marca X tiene buena relación calidad y precio, frente a un consumidor que sí considera que tiene buena relación calidad y precio.
Atributo_4	0	Tiene envases atractivos /adecuados	0.623	Es 0.623 veces menos probable que un consumidor de gaseosa consuma la marca x si no considera que la marca tiene envases atractivos /adecuados , frente a un consumidor que si considera que tiene envases atractivos/adecuados.
Atributo_22	0	Para compartir con amigos	0.785	Es 0.785 veces menos probable que un consumidor de gaseosa consuma la marca x si no considera que la marca X es para compartir con los amigos, frente a un consumidor que considera que es para compartir con los amigos.
Atributo_21	0	Te levanta el animo	0.849	Es 0.849 veces menos probable que un consumidor de gaseosa consuma la marca x si no considera que marca X levanta el animo, frente a un consumidor que si considera que la marca levanta el animo.

- Capacidad predictiva de aciertos:

  El porcentaje de aciertos que tiene el modelo de regresión Logístico es de 87.9%

  Este valor se obtiene de la Tabla 7, sumando la diagonal de la matriz y dividiendo el porcentaje entre el total de casos.



```
*******************************************************************************
                          The FREQ Procedure
                     Table of F_Target by I_Target
                  F_Target(From: Target)
                           I_Target(Into: Target)
                  Frequency,
                  Percent  ,
                  Row Pct  ,
                  Col Pct  ,0       ,1       ,  Total
                  ---------------------------------------
                  0        ,   2151 ,    115 ,   2266
                           ,  74.69 ,   3.99 ,  78.68
                           ,  94.92 ,   5.08 ,
                           ,  90.23 ,  23.19 ,
                  ---------------------------------------
                  1        ,    233 ,    381 ,    614
                           ,   8.09 ,  13.23 ,  21.32
                           ,  37.95 ,  62.05 ,
                           ,   9.77 ,  76.81 ,
                  ---------------------------------------
                Total        2384      496     2880
                             82.78    17.22   100.00
*******************************************************************************
```

Figura 6: Capacidad productiva de aciertos

El modelo final está dado por la siguiente función:

$$f\left(\frac{p}{1-p}\right) = e^{(1.851+1.58\,\text{Re\_}marca\_1-0.0596\,\text{Re\_}marca-0.2820\,Atributo\_7-1.2010\,Atributo\_3-}$$

$$1.713\,favorita-0.163\,Atributo\_18-0.201\,Atributo\_16-0.151\,Atributo\_14-0.568\,Atributo\_8$$

$$-0.2739\,Atributo\_6-0.4739\,Atributo\_4-0.2427\,Atributo\_22-0.163\,Atributo\_21)$$

3.2. Árbol de decisión

El árbol seleccionado, fue un árbol de 4 niveles con 9 nodos terminales y con 6 variables importantes las cuales se presentan en la Tabla 7.

Tabla 7: Variables selecciónadas para el árbol de decisión

Variable	*Descripción*	**Importancia**
Atributo_3	Es para alguien como tu	1
Re_marca	Recordación de marca	0.51
Atributo_1	Es refrescante	0.487
Atributo_22	Para compartir con amigos	0.388
Atributo_4	Tiene envases atractivos /adecuados	0.311
Atributo_2	Es de gran calidad	0.228

El árbol de decisión se muestra en la Figura 7 y la interpretación de cada uno de los nodos terminales se presenta en la Tabla 8.

Tabla 8: Interpretación de los nodos terminales

Node ID	Probabilidad (Target=1)	Regla de Desición	Descripción
4	0.94	Atributo_3=1 and Re_marca=1	Considera que la marca X es para alguien como el/ella (se identifica con la marca) y es la primera marca que se le viene a la mente cuando se habla de bebidas gaseosas.
5	0.64	Atributo_3=1 and Re_marca=2	Considera que la marca X es para alguien como el/ella (se identifica con la marca) y recuerda la marca no en primera mención pero la recuerda espontáneamente.
6	0.23	Atributo_3=1 and Re_marca=3	Considera que la marca X es para alguien como el/ella (se identifica con la marca) y pero no la recuerda espontáneamente.
23	0.53	Atributo_3=0 and Atributo_1=1 and Atributo_4=1 and Atributo_2=0	Considera que la marca X no es para alguien como el/ella (no se siente identificado con la marca) pero si la considera una marca refrescante, de envases adecuados/atractivos pero no de calidad.
24	0.29	Atributo_3=0 and Atributo_1=1 and Atributo_4=1 and Atributo_2=1	Considera que la marca X no es para alguien como el/ella (no se siente identificado con la marca) pero si la considera una marca refrescante, de envases adecuados/atractivos y calidad.
16	0.12	Atributo_3=0 and Atributo_1=1 and Atributo_4=0	Considera que la marca X no es para alguien como el/ella (no se siente identificado con la marca) pero si la considera una marca refrescante pero que no tiene envases adecuados/atractivos.
17	0.17	Atributo_3=0 and Atributo_1=1 and Atributo_22=1	Considera que la marca X no es para alguien como el/ella (no se siente identificado con la marca) pero si la considera una marca refrescante e ideal para compartir con los amigos.
18	0.03	Atributo_3=0 and Atributo_1=1 and Atributo_22=0	Considera que la marca X no es para alguien como el/ella (no se siente identificado con la marca) pero si la considera una marca refrescante pero no ideal para compartir con los amigos.

Un individuo será considerado consumidor de la Marca X, si pertenece a cualquiera de los siguientes nodos del árbol de decisión 4, 5 y 23, ya que son los que presentan probabilidades (*target*=1) mayores de 0.50.

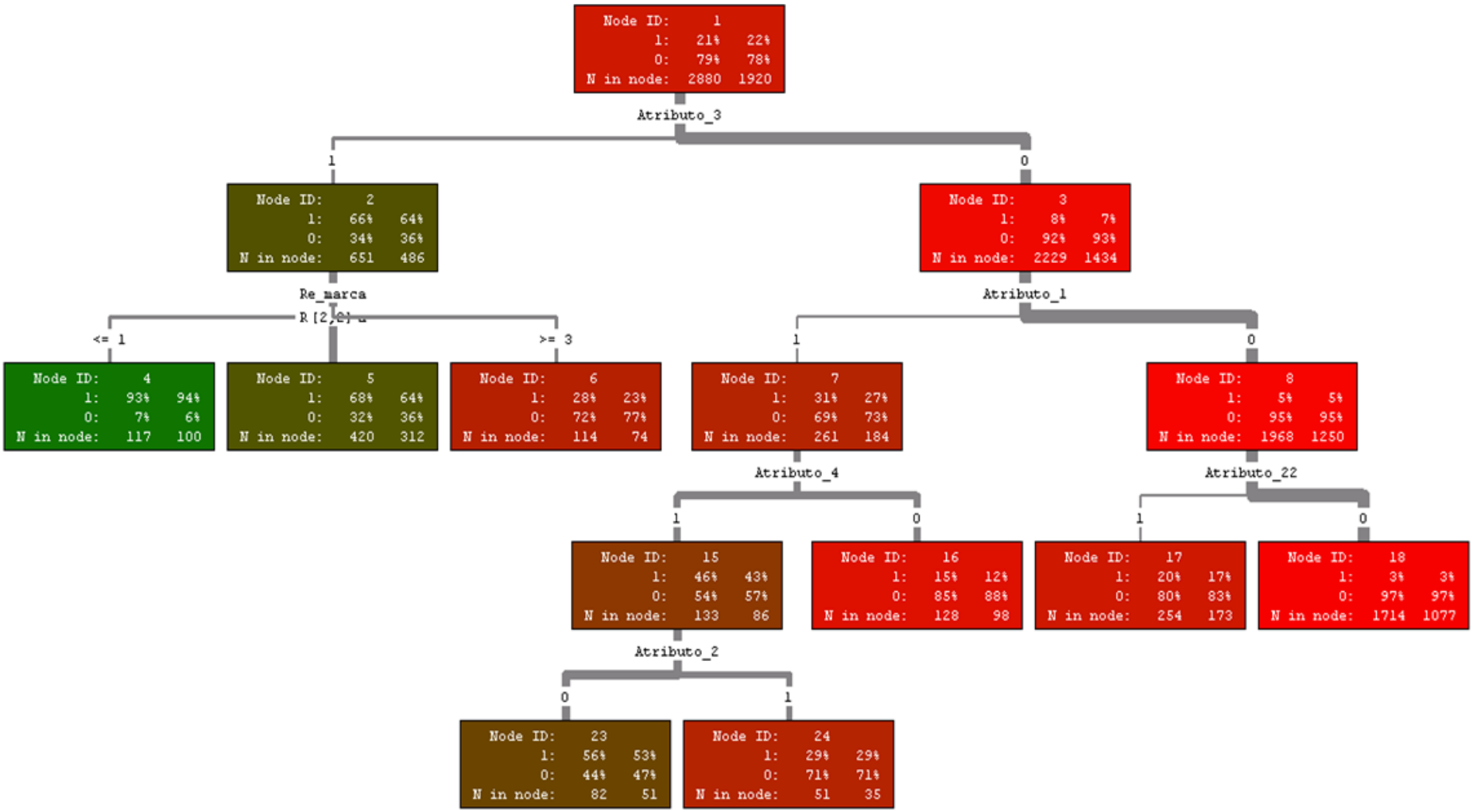


Figura 7: Árbol de decisión

Otra forma gráfica de presentar el árbol de decisión, es la que se presenta a continuación en la Figura 8, denominado *Tree Ring*, el cual muestra como la data es particionada en cada nivel, siendo el centro del diagrama la data total. El color indica la precisión en la predicción dentro de la partición. El color rojo indica la baja precisión con respecto a la variable *target* y el color amarillo indica alta precisión con respecto a la variable *target*.

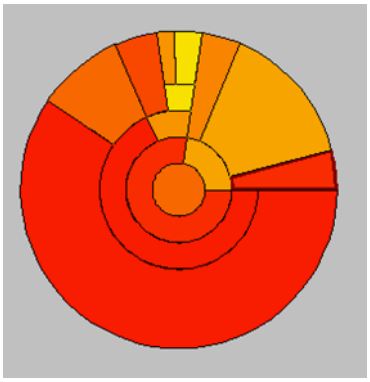


Figura 8: *Tree Ring*

3.3. Comparación entre el Modelo de Regresión y el Árbol de Decisión

El *Enterprise Miner* a través del nodo *Assessment* (ver figura1), nos permite comparar ambos modelos, con la finalidad de escoger el modelo que mejor se ajuste a los datos, para ello brinda una serie de gráficos que se muestra a continuación.

- Porcentaje de respuesta por percentil:

  La Figura 9 de obtiene de la siguiente manera, se calcula los percentiles de la muestra en función de la probabilidad de éxito; Target igual a 1 significa consumidor de la Marca X, de este modo se puede decir que en el percentil 10 estarán los individuos que según el modelo de árbol de decisión o regresión según corresponda, tienen mayor probabilidad de consumir la Marca X y en el percentil 100 estarán los individuos que según el modelo de árbol de decisión o regresión tienen menor probabilidad de consumir la Marca X.

  Entonces, la interpretación del gráfico en el percentil 10 es la siguiente: La proporción de consumidores de la Marca X que se encuentran en el percentil 10 es

80% según el modelo de árbol de decisión y según el modelo de regresión es 85%. Si el modelo se ajusta bien a los datos, lo lógico es que a medida que aumenta el percentil disminuya la proporción de consumidores de la Marca X.

Cabe mencionar, que la línea *baseline* corresponde al porcentaje de consumidores de la Marca X en una muestra aleatoria.

En general, podemos concluir que ambas curvas resultan ser importantes, sin embargo la curva de la regresión se encuentra ligeramente por encima a la curva del árbol decisión, por ello el modelo de regresión sería ligeramente mejor al modelo de árbol de decisión.

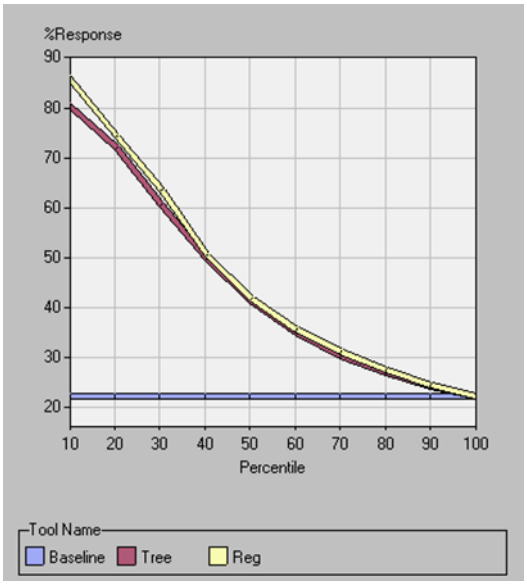


Figura 9: Porcentaje de respuesta por modelo

- Curva Lift

  La Figura 10, se obtiene comparando el porcentaje de respuesta en cada percentil con el porcentaje de respuesta del modelo *baseline* el cual se construye considerando datos aleatorios. Como resultado de esta comparación lo que obtiene es el número de veces que el modelo de regresión o el modelo de árbol de decisión es mejor o peor al modelo construido con datos aleatorios, esto es considerando porcentajes de respuesta.

El Lif para el modelo árbol de decisión es 3.67 y para el modelo de regresión es 3.92. Según el árbol de decisión, un individuo que pertenece al percentil 10 es 3.67 veces más probable que consuma la Marca X a un individuo que es seleccionado de una muestra aleatoria.

Según la regresión Logística, un individuo que pertenece al percentil 10 es 3.92 veces más probable que consuma la Marca X a un individuo que es seleccionado de una muestra aleatoria.

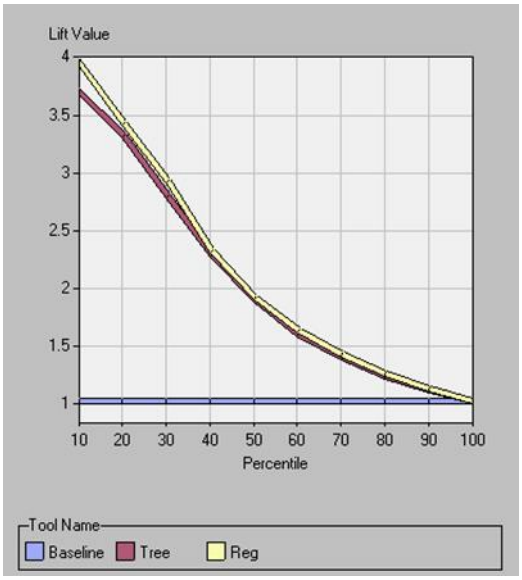


Figura 10: Curva lift por modelo

- Porcentaje de Captura de Respuesta:

  La Figura 11, muestra la proporción de aciertos del modelo con respecto al total de la muestra, para el modelo de árbol de decisión nos muestra que el 39% de todos consumidores de la Marca X se encuentran en el percentil 10, para el modelo de regresión nos muestra que el 41% de los consumidores de la Marca X se encuentran en el percentil 10.

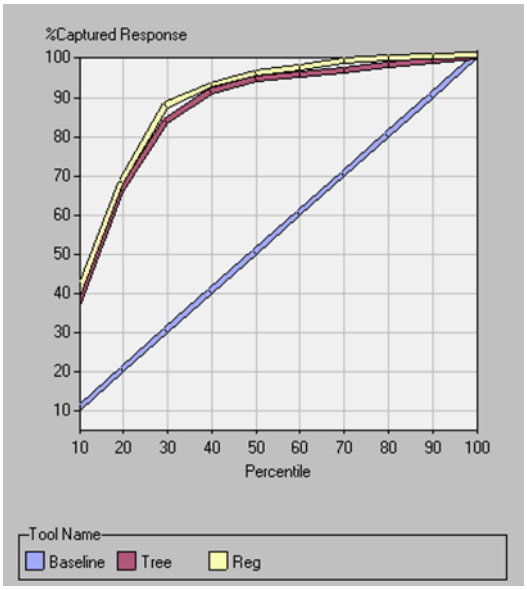


Figura 11: Porcentaje de captura de respuesta por modelo

De manera general, se puede concluir que el modelo de regresión Logístico es ligeramente superior al árbol de decisión, sin embargo la ventaja del árbol seria que solo utiliza 6 variables por lo cual la clasificación de nuevos individuos resulta sencilla, en otras palabras la aplicación de modelo es más sencilla si usamos el modelo de árbol de decisión. Por lo tanto, se eligió el modelo de árbol de decisión.

3.4. Variables que determinan el consumo de la Marca X

En la Tabla 9, se muestra las variables seleccionas por cada uno de los modelos desarrollados.

Sin embargo, como se ha concluido que el modelo adecuado es el árbol de decisión, se considera como necesarias y suficientes para predecir el consumo de la Marca X, las siguientes 6 variables seleccionas por el árbol de decisión: Es para alguien como tú, recordación de marca, es refrescante, es para compartir con amigos, tiene envases atractivos, adecuados y es de gran calidad. Es importante mencionar que cuatro variables de las seis que selecciona el árbol también fueron seleccionadas por el modelo de regresión.

Tabla 9: Variables seleccionas por modelo

Variables	Descripción	Arbol	Regresión
Atributo_3	Es para alguien como tu	x	X
Re_marca	Recordación de marca	x	X
Atributo_1	Es refrescante	x	
Atributo_22	Para compartir con amigos	x	X
Atributo_4	Tiene envases atractivos /adecuados	x	X
Atributo_2	Es de gran calidad	x	
Atributo_7	Es una marca confiable y con prestigio		X
Favorita	Marca favorita		X
Atributo_18	Tiene mejor sabor que otras marcas		X
Atributo_16	Buena para tomar con las comidas		X
Atributo_14	Se encuentra en todas partes		X
Atributo_8	Quita la sed		X
Atributo_6	Tiene una buena relación entre calidad y precio		X
Atributo_21	Te levanta el animo		X

3.5. Instrumento para la Aplicación del Modelo

Con la finalidad de que el modelo elegido sea utilizado por la Empresa X, para identificar a los consumidores de la Marca X cuando se realicen estudios cualitativos como entrevistas o focus grups, se elaboró una herramienta sencilla, de fácil aplicación y que rápidamente permita obtener resultados.

La herramienta se elaboró en excel, utilizando lógicas condicionales que determina a que nodo pertenece el individuo y la probabilidad de consumir la marca X.A continuación se muestra el formato de la herramienta.

Nombre : _____

Edad: _____

Sexo: 1.Fem 2.Masc

NSE: 1. Alto 2 Medio 3 Bajo 4. Muy Bajo

¿Qué marca de bedida gaseosa recuerda?

		Respuestas		
		1	2	3
Re_marca	Gaseosas	Primera mención Espontanea	Otras menciones Espontanea	Otras menciones Ayudada (Con tarjeta)
	1.Marca X	x		
	2.Marca A			x
	3.Marca B			x
	4.Marca C		x	
	5.Marca D			x
	6.Marca F			x
	7.Marca G		x	x
	8.Marca H			
	Otra_____			

¿Cuál o cuales de las siguientes marcas (Mostar tarjeta con las marca escogidas) considera usted que tiene o tienen las siguientes caracteristicas?

		Marca A	Marca B	Marca X	Marca C
Atributo_3	Es para alguien como tú	x		x	
Atributo_1	Es refrescante				
Atributo_22	Para compartir con amigos		x	x	
Atributo_4	Tiene envases atractivos /adecuados	x			x
Atributo_2	Es de gran calidad			x	

Identificación del consumidor de Marca X

	Puntuación
Re_marca	1
Atributo_3	1
Atributo_1	0
Atributo_22	1
Atributo_4	0
Atributo_2	1
Combinación	110101
Nodo al que pertenece	**4**
Probabilidad de consumir la marca X	**94%**
Consumidor de la Marca X	**SI**

IV. CONCLUSIONES

El modelo de regresión logístico que mejor se ajustó a los datos, utilizó 12 variables mientras que el árbol de decisión utilizó 6 variables. Ambos modelos coincidieron en seleccionar 4 variables.

Cuando se compararon ambos modelos considerando el porcentaje de respuesta por percentil, Lift y porcentaje de captura de respuesta, se observó que el modelo de regresión era ligeramente superior al modelo árbol de decisión. Sin embargo el modelo elegido para predecir el consumo de la marca X fue el árbol de decisión básicamente por las siguientes razones: buena capacidad predictiva, es fácil interpretar, utiliza pocas variables y su aplicación es sencilla.

Según el modelo elegido, las variables que determinan el consumo de la Marca X son la recordación de marca y los siguientes atributos: es para alguien como tú, es refrescante, tiene envases atractivos, adecuados, es de gran calidad y es para compartir con amigos. Estas serían la variable sobre las cuales la Empresa X debería tomar acción para mantener a sus consumidores actuales o captar nuevos consumidores, una forma sería utilizar los atributos mencionados en la publicidad de marca X.

La clasificación de nuevos individuos se realizará a través de la aplicación la herramienta elaborada.

V. REFERENCIAS BIBLIOGRÁFICAS

Hair, J.; Anderson, R.; Tatham, R.; Black, W. 1999. Análisis Multivariante. 5ed. Editorial Prentice Hall Hispanoamericana. México.

Lehmann, D. 1993. Investigación y Análisis de Mercado. 1 ed. Editorial Continental S.A.

Levin, R.; Rubin, D.; Balderas, M.; Del Valle, J.; Gómez, R. 2004. Estadística para administración y economía. 5 ed. Editorial Pearson.

SAS Institute Inc. Cary. 2003. Predictive Modeling Using Enterprise Miner.

SAS Institute. 2001. Predictive Modeling Using Logistic Regression.