

PRACTICAS IBM SPSS

LOS ARBOLES DE DECISION

GUIA DE PRÁCTICAS ARBOLES CHAID, CRT Y QUEST

SANTIAGO VASQUEZ A. Ms.C

Docente Facultad de Ingeniería Industrial

UTP

UNIVERSIDAD TECNOLÓGICA DE PEREIRA

Pereira – Risaralda - Colombia

MAYO DE 2018

PRACTICA 1. METODO CHAID

Para crear un árbol de decisión, después de cargar el conjunto de datos (tree_credit.sav) elija en los menús: **Analizar** - **Clasificar** - **Árbol** (figura 1). En la pantalla de entrada seleccionamos una variable dependiente y una o más variables independientes y como método de crecimiento elegimos CHAID (define el método de construcción de árbol) tal y como se indica en la figura 2 Se puede hacer clic en el botón categorías para seleccionar una o más categorías de interés fundamental en el análisis. Por ejemplo, en nuestro análisis conocer los clientes que no devuelven el crédito, por eso elegimos malo como categoría objetivo y hacemos clic en continuar.

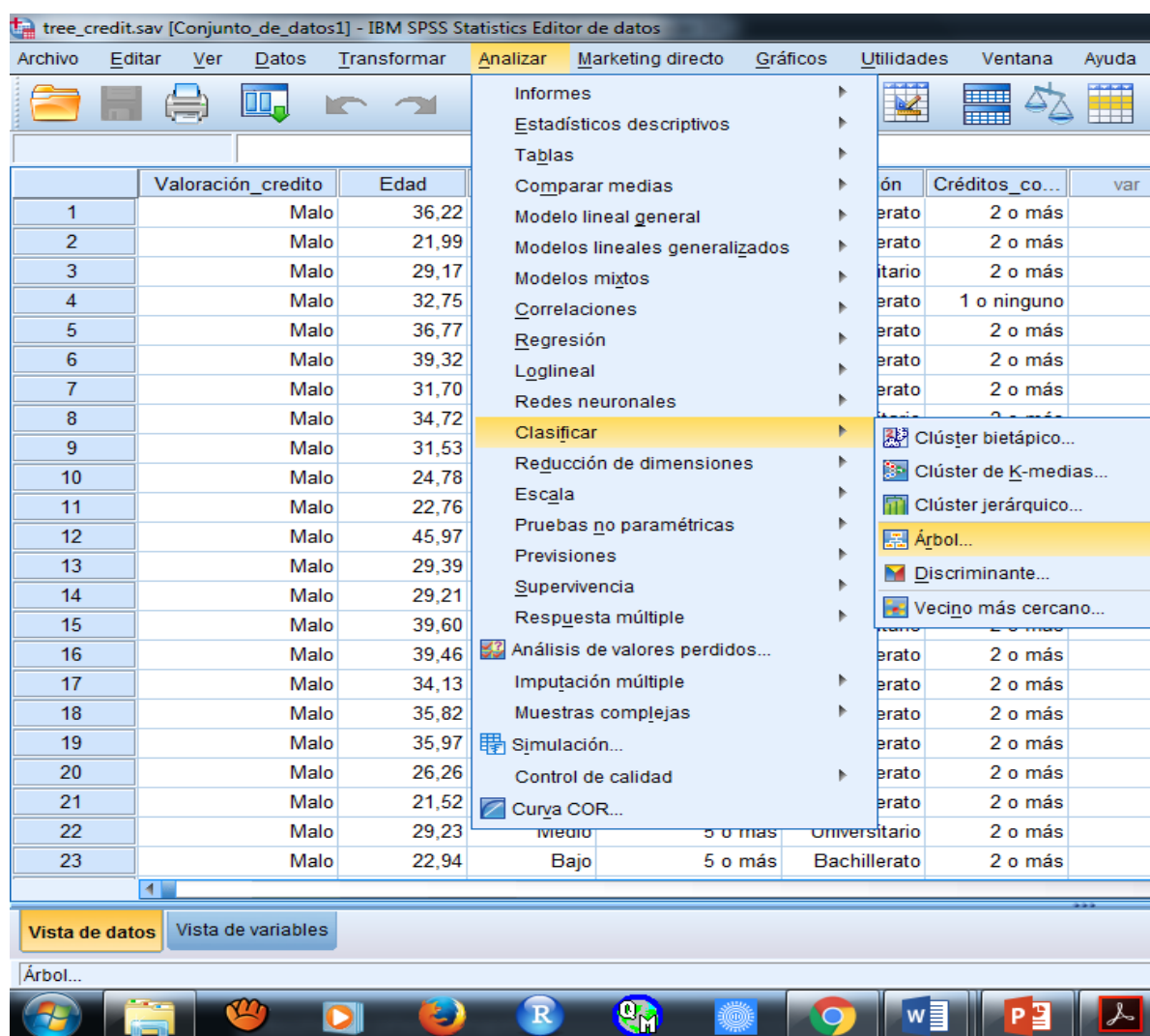


FIGURA1

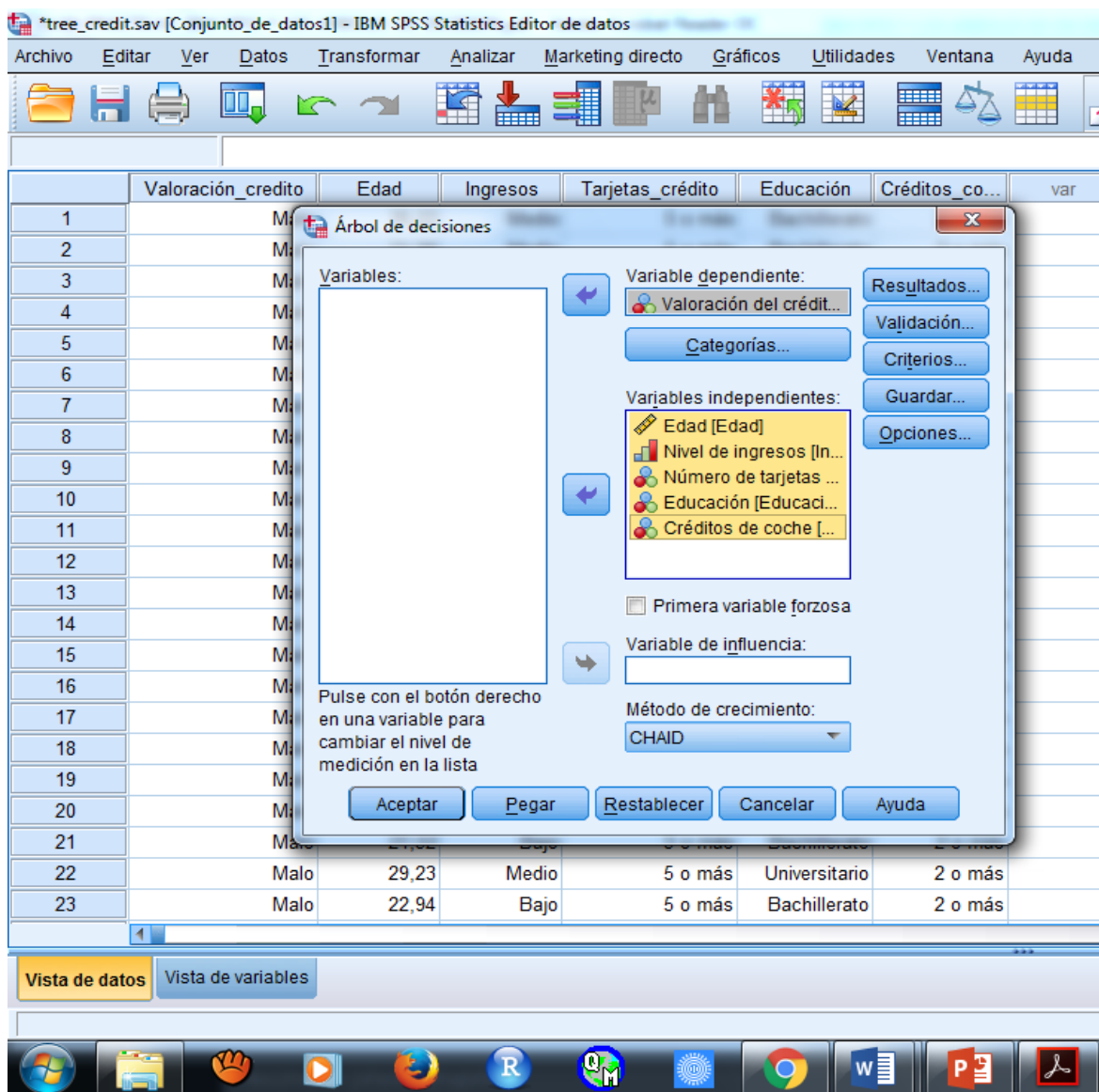


FIGURA 2

También se puede seleccionar una variable de influencia que defina cuanta influencia tiene un caso en el proceso de crecimiento de un árbol. Los casos con valores de influencia inferiores tendrán menos influencia, mientras que los casos con valores superiores tendrán más. Los valores de la variable en influencia deben ser valores positivos. Si se marca la casilla primera variable forzosa, se fuerza a que la primera variable en la lista de variables independientes en el modelo sea la primera variable de división. En el botón resultados de la figura 2 se selecciona la forma de representación del árbol (Figura 3), los estadísticos a obtener (figura 4) y las reglas (Figura 5). Pulse en continuar.

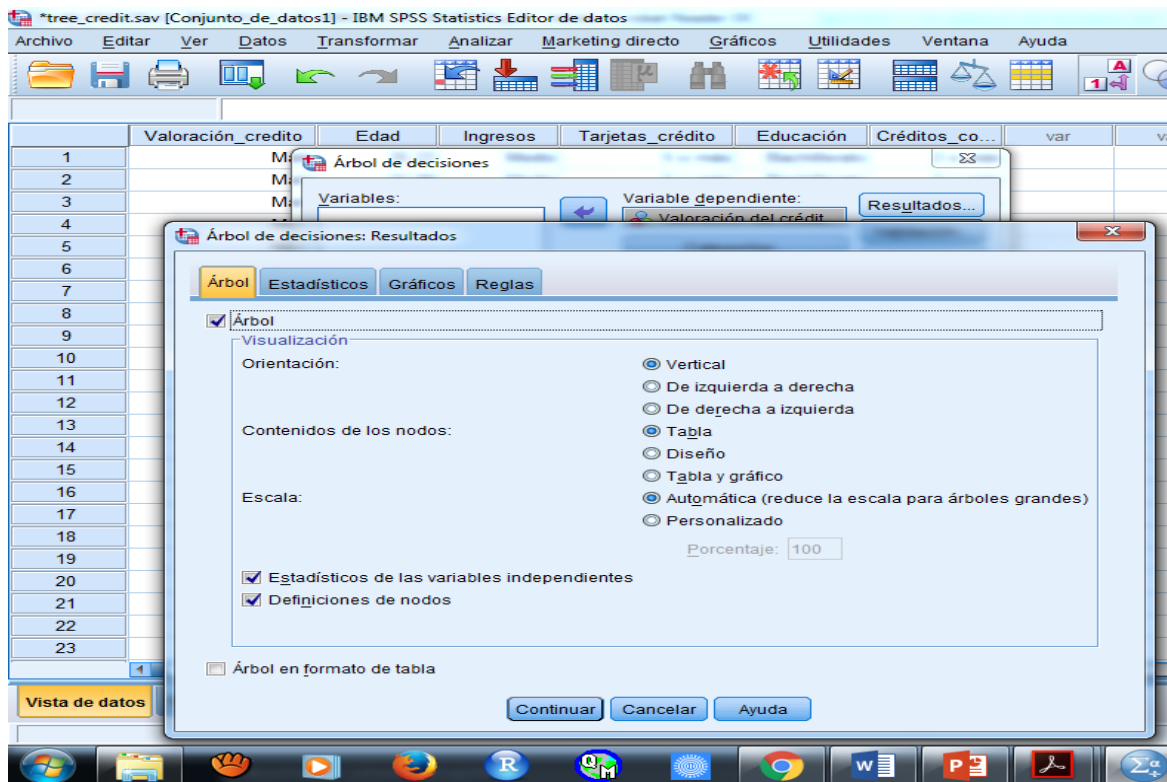


FIGURA 3

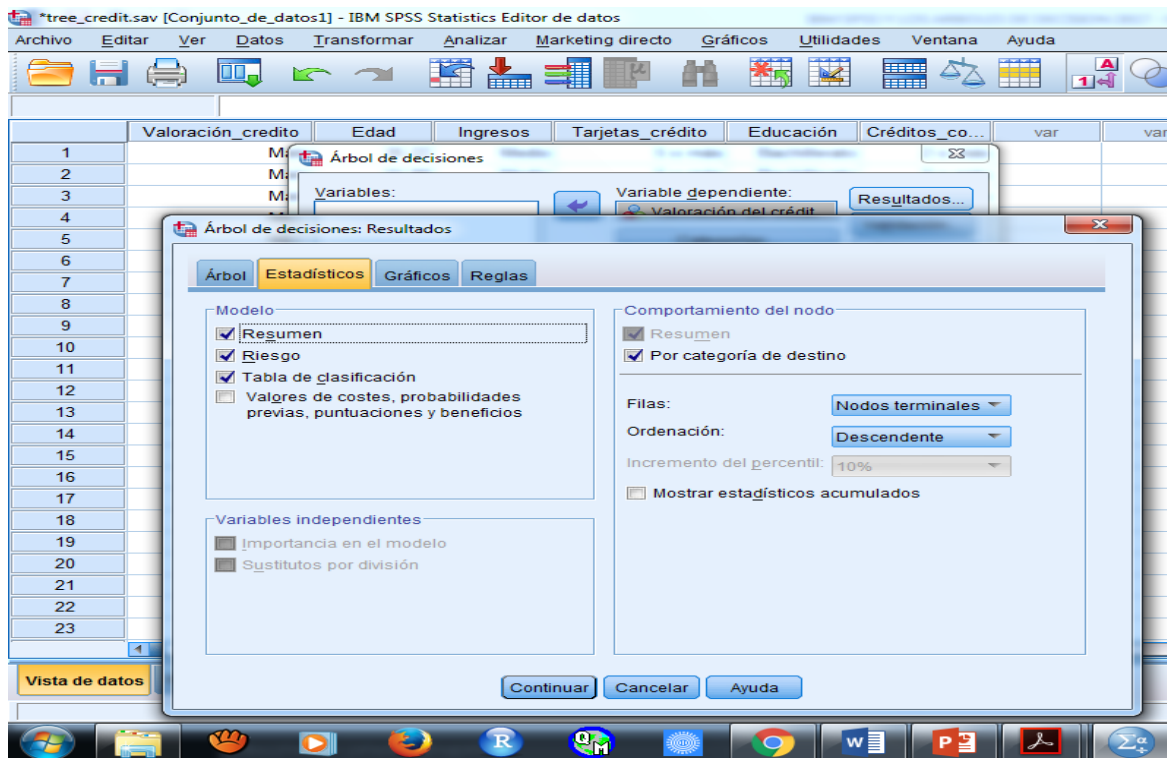


FIGURA 4

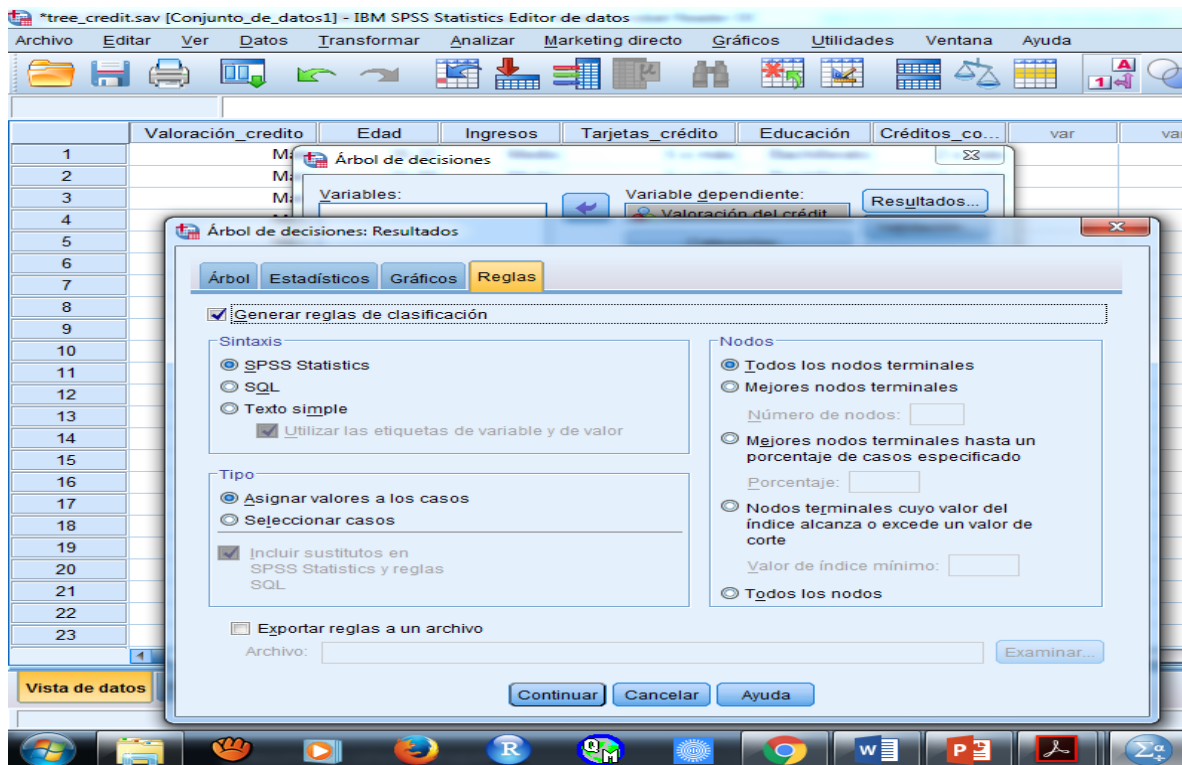


FIGURA 5

En el botón validación de la figura 2 se valida el árbol (Figura 6). La validación permite evaluar la bondad de la estructura de árbol cuando se generaliza para una mayor población. Hay dos métodos de validación disponibles: validación cruzada y validación por división muestral. La validación cruzada divide la muestra en un número de submuestras. A continuación, se generan los modelos de árbol, que no incluyen los datos de cada submuestra. El primer árbol se basa en todos los casos, excepto los correspondientes al primer pliegue de la muestra; el segundo árbol, se basa en todos los casos, excepto los correspondientes al primer pliegue de la muestra y así sucesivamente. Para cada árbol se calcula el riesgo de clasificación errónea, aplicando el árbol a la submuestra que se excluyó al generarse este. Se puede especificar un máximo de 25 pliegues de la muestra. Cuanto mayor sea el valor, menor será el número de casos excluidos de cada modelo de árbol. La validación cruzada genera un modelo de árbol único y final. La estimación de riesgos de todos los árboles. Con la validación por división muestral, el modelo se genera utilizando una muestra de entrenamiento y después pone a prueba ese modelo con una muestra de reserva. Puede especificar un tamaño de la muestra de entrenamiento, expresado como un porcentaje del

tamaño muestral total, o una variable que divida la muestra en muestras de entrenamiento, expresado como un porcentaje del tamaño muestral total, o una variable que divida la muestra en muestras de entrenamiento y de comprobación, los casos con un valor igual a 1 para la variable se asignaran a la muestra de comprobación. Dicha variable no puede ser ni la variable dependiente, ni la de ponderación, ni la de influencia, ni una variable independiente forzada. Los resultados se pueden mostrar tanto para la muestra de entrenamiento como para la de comprobación, o solo para esta última. La validación por división muestral se debe utilizar con precaución en archivos de datos pequeños (archivos de datos con un número pequeño de casos). Si se utilizan muestras de entrenamiento de pequeño tamaño, pueden generarse modelos que no sean significativos, ya que es posible que no haya suficientes casos en algunas categorías para lograr un adecuado crecimiento de árbol.

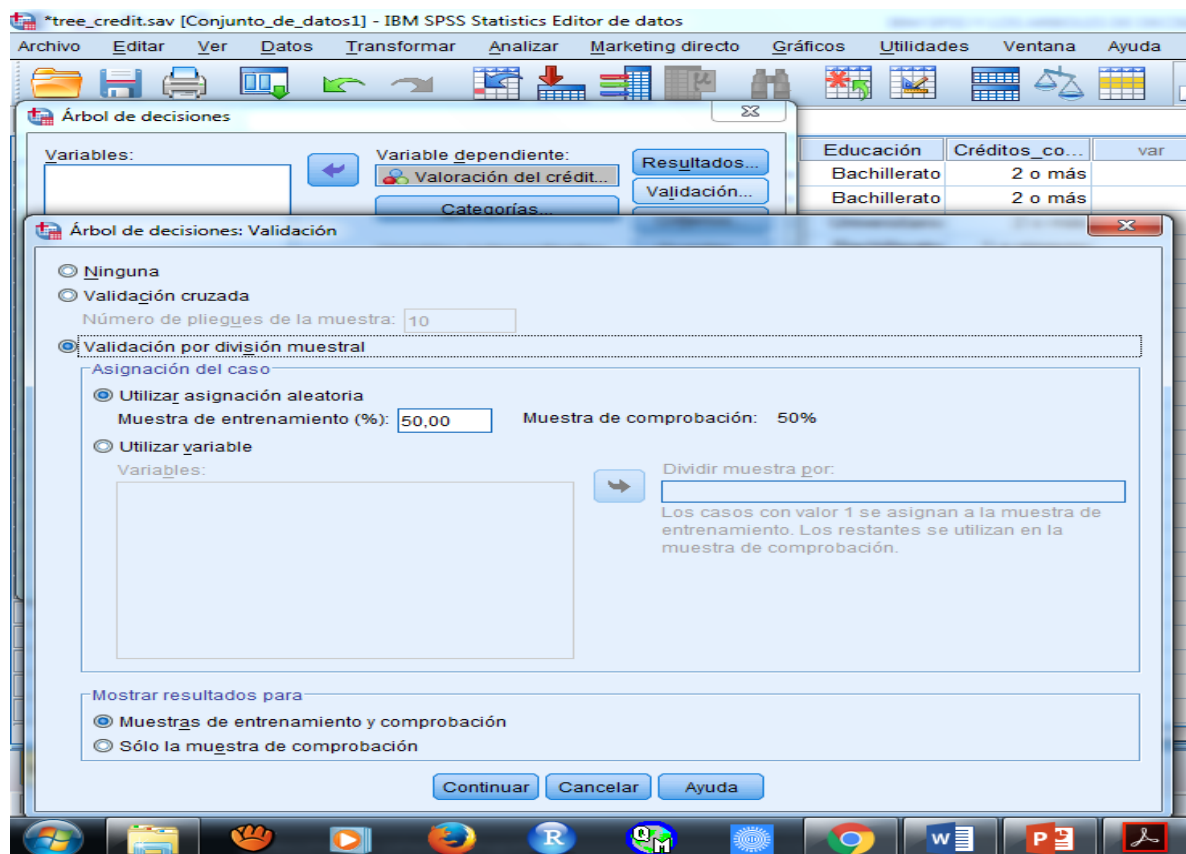


FIGURA 6

En el botón criterios de la figura 2 se personalizan los criterios de crecimiento del árbol. La pestaña límites de crecimiento (Figura 7) permite limitar el número de niveles del árbol y controlar el número de casos mínimo para nodos parentales y filiales. El campo máxima

profundidad de árbol controla el número máximo de niveles de crecimiento por debajo del nodo raíz. El ajuste automático limita el árbol a tres niveles por debajo del nodo raíz para los métodos CHAID y CHAID exhaustivo, y a cinco niveles para los métodos CRT y QUEST. El campo número e caos mínimo controla el número de casos mínimo para los nodos. Los nodos que no cumplen estos criterios no se dividen. El aumento de los valores mínimos tiende a generar arboles con menos nodos. La disminución de dichos valores mínimos generara arboles con más nodos.

Para archivos de datos con un número pequeño de casos, es posible que, en ocasiones, los valores por defecto de 100 casos para nodos parentales y de 50 casos para nodos filiales den como resultado arboles sin ningún nodo por debajo de la no raíz; en este caso, la disminución de los valores mínimos podría generar resultados más útiles.

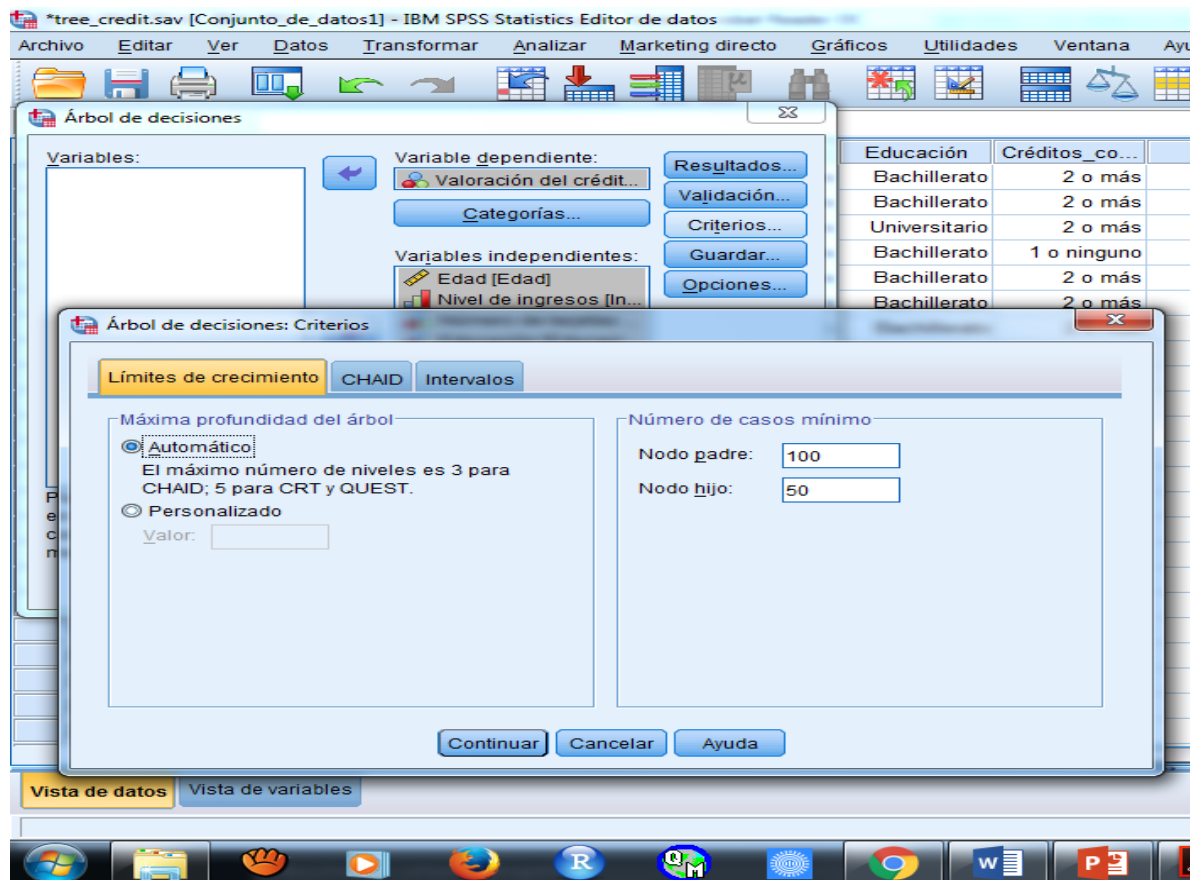


FIGURA 7

En la pestaña CHAID (Figura 8) se puede controlar para controlar para los métodos CHAID y CHAID efectivo el nivel de significación para la división de nodos y la función de

categorías, el estadístico chi-cuadrado a utilizar (Pearson para cálculos rápidos y muestras grandes o razón de verosimilitud si se quiere robustez o se trabaja con muestras pequeñas), en el método de estimación del modelo (para variables dependientes ordinales y nominales se puede especificar el número máximo de iteraciones, el cambio mínimo en las frecuencias esperados de las casillas), corregir los valores de significación mediante el método de Bonferroni (comparaciones múltiples, los valores de significación para los criterios de división y fusión se corrigen utilizando el método de Bonferroni que es el método por defecto), y permitir nueva división de las categorías fusionadas dentro de un nodo para que el procedimiento intente la fusión de las categorías de variables (predictoras) independiente entre sí para generar el árbol más simple posible.

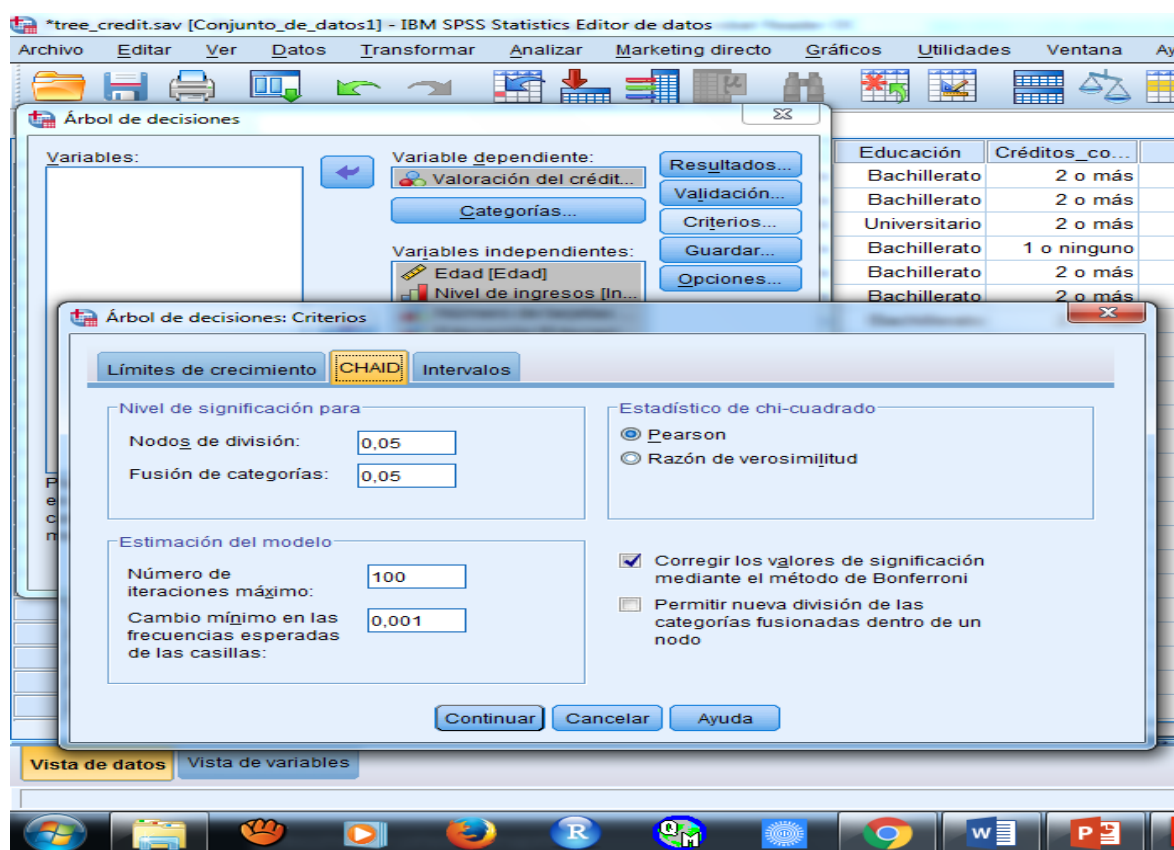


FIGURA 8

En la pestaña intervalos (Figura 9) se fijan intervalos de escala para el análisis CHAID. En el análisis CHAID, las variables (predictoras) independientes de escala siempre se categorizan en grupos discretos (por ejemplo, 0-10, 11-20, 21-30, etc.) antes del análisis.

El botón opciones de la figura 11 permite fijar opciones para tratamiento de valores perdidos (figura 11).

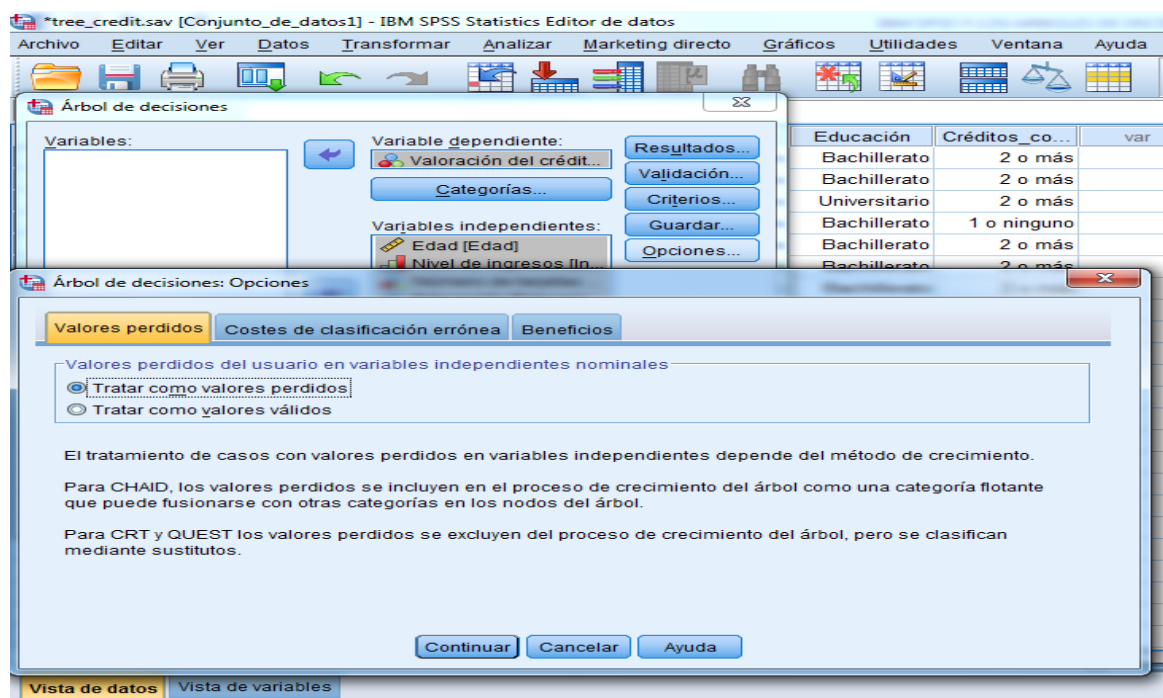


FIGURA 11

Se definen también los costes de clasificación errónea (figura 12).

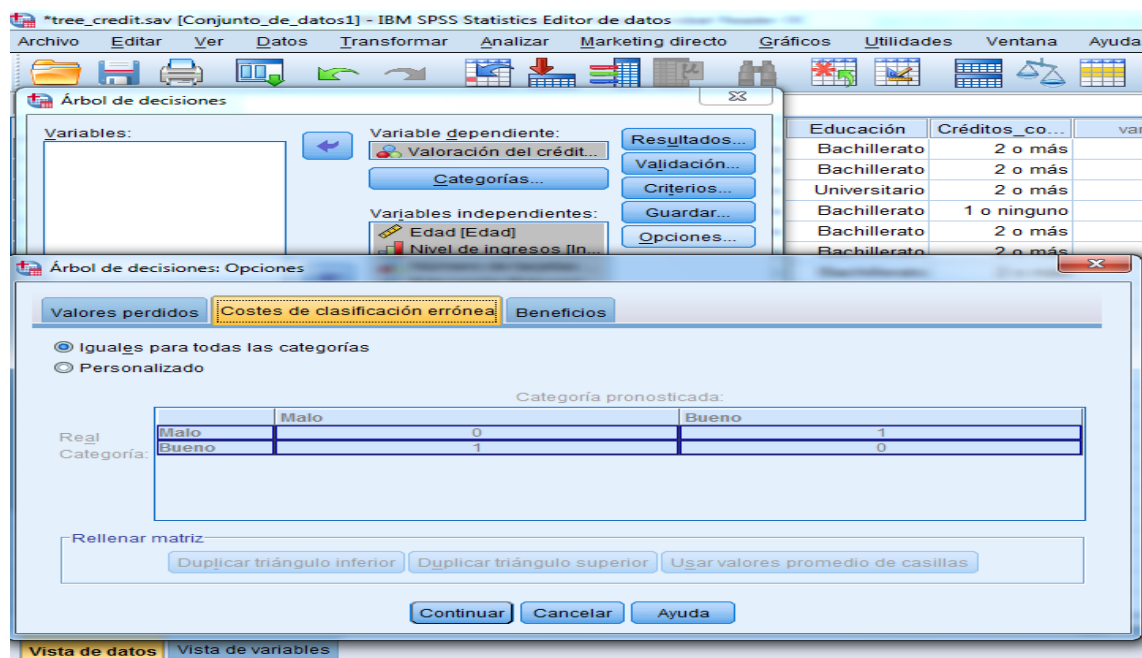


FIGURA 12

y beneficios por cada categoría (figura 13).

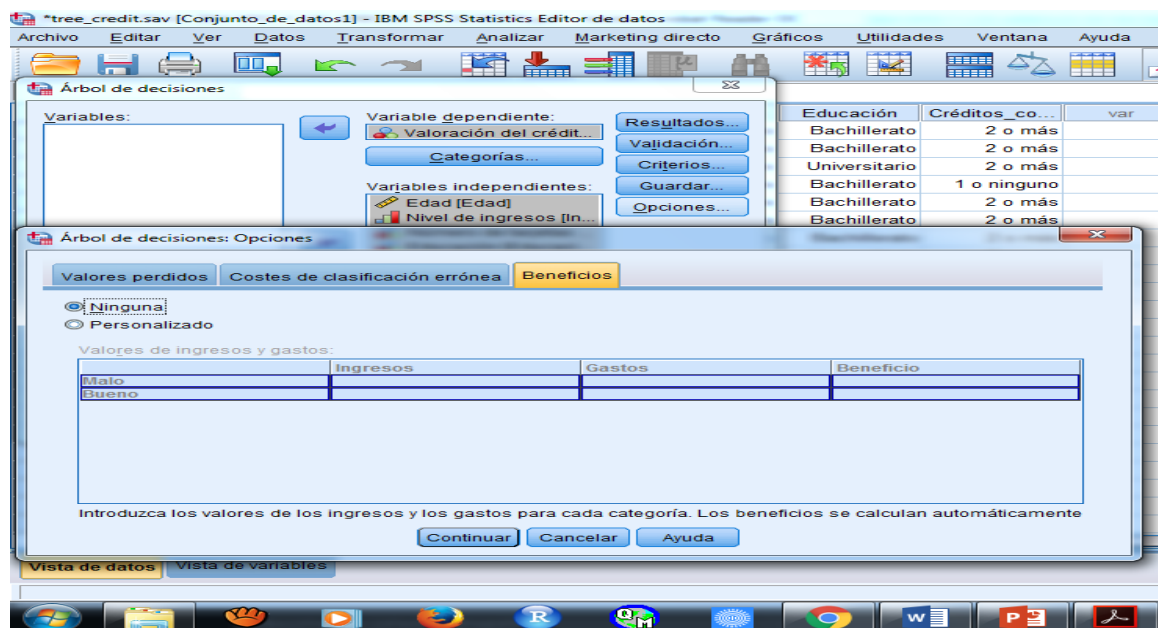


FIGURA 13

Al hacer clic en aceptar en la figura 2, se crea el árbol (figura 14). Lo primero que observamos en el árbol es que el 42,5% de los clientes presentara crédito fallido y el 57,5% presentara devolución de crédito en tiempo y forma. A continuación, se observa que el nivel de ingresos es el mejor predictor de la tasa de riesgo crediticio, ya que representa el primer nivel de ramificación en el árbol. Para el nodo 1 (nivel de ingreso bajo) el 82,4% de los clientes presentan crédito fallido y el 17,6% devuelven el crédito en tiempo y forma. Para el nodo 2 (nivel de ingresos medio) el 40.7% de los clientes presentan crédito fallido y el 59.3% devuelven el crédito en tiempo y forma. *(los valores obtenidos al hacer la practica varían un poco, pues las muestras generadas son diferentes en cada equipo)*

Para el nodo 3 (nivel de ingresos alto) el 14.7% de los clientes presentan crédito fallido y un 85.3% devuelven el crédito en tiempo y forma. El siguiente predictor en calidad de la tasa de riesgo crediticio es **crédito coches** para nivel de ingreso bajo (para la categoría de 2 o más, el 90,8% es malo y se reduce al 56,3% para 1 o ningún crédito), pero para nivel medio de ingreso la variable de clasificación es **número de tarjetas de crédito**. De igual forma se analizan los nodos restantes.

En la figura 15, 16 se presentan salidas de resumen del modelo

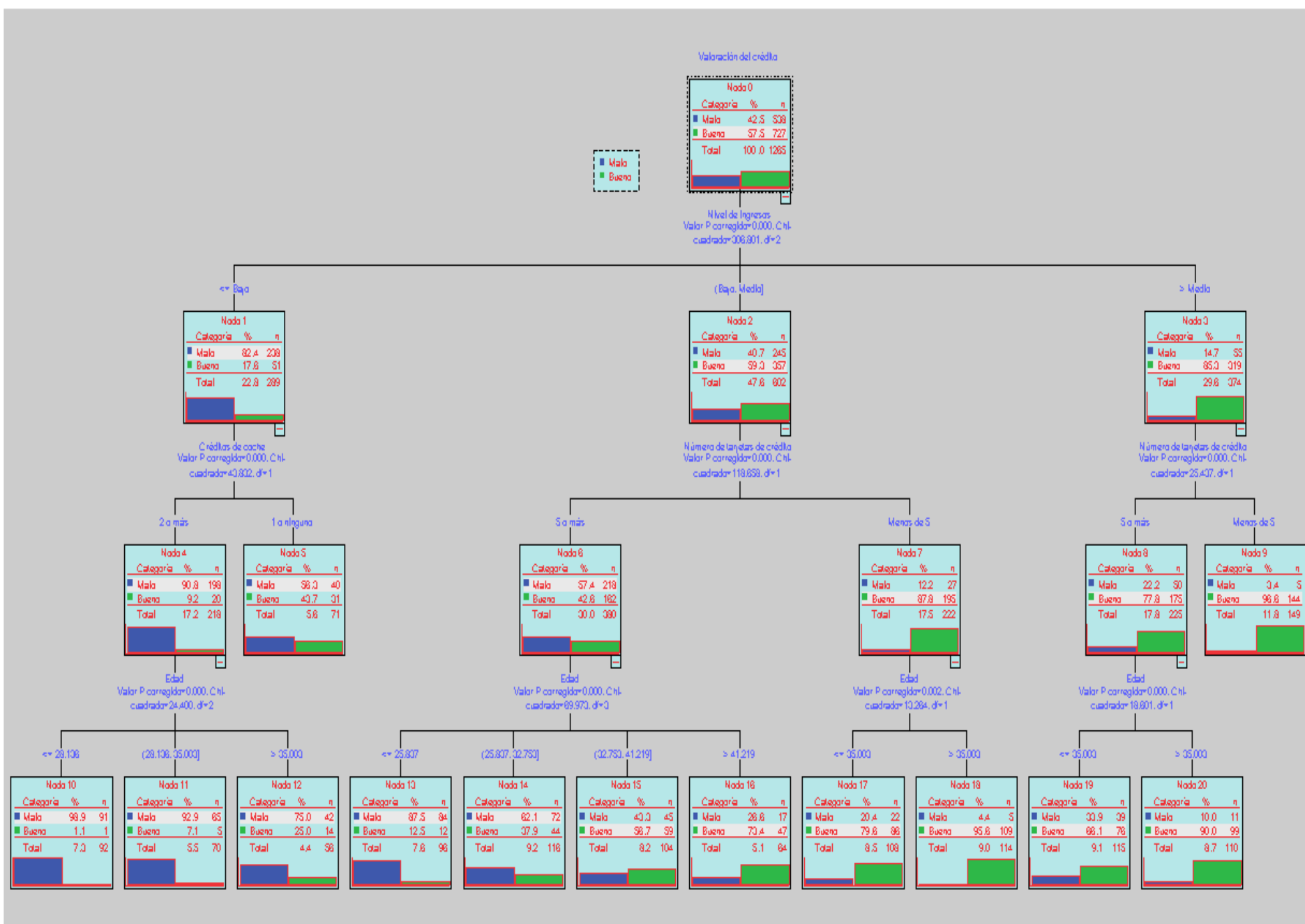


FIGURA 14

Resumen del modelo

Especificaciones	Método de crecimiento	CHAID
	Variable dependiente	Valoración del crédito
	Variables independientes	Edad, Nivel de ingresos, Número de tarjetas de crédito, Educación, Créditos de coche
	Validación	Dividir muestra
Resultados	Máxima profundidad de árbol	3
	Mínimo de casos en un nodo filial	100
	Mínimo de casos en un nodo parental	50
	Variables independientes incluidas	Nivel de ingresos, Créditos de coche, Edad, Número de tarjetas de crédito
	Número de nodos	21
	Número de nodos terminales	13
	Profundidad	3

FIGURA 15

Clasificación

Muestra	Observado	Pronosticado		
		Malo	Bueno	Porcentaje correcto
Entrenamiento	Malo	394	144	73,2%
	Bueno	107	620	85,3%
	Porcentaje global	39,6%	60,4%	80,2%
Contraste	Malo	350	132	72,6%
	Bueno	104	613	85,5%
	Porcentaje global	37,9%	62,1%	80,3%

Métodos de crecimiento: CHAID
Variable dependiente: Valoración del crédito

FIGURA 16

Riesgo

Muestra	Estimación	Típ. Error
Entrenamiento	,198	,011
Contraste	,197	,011

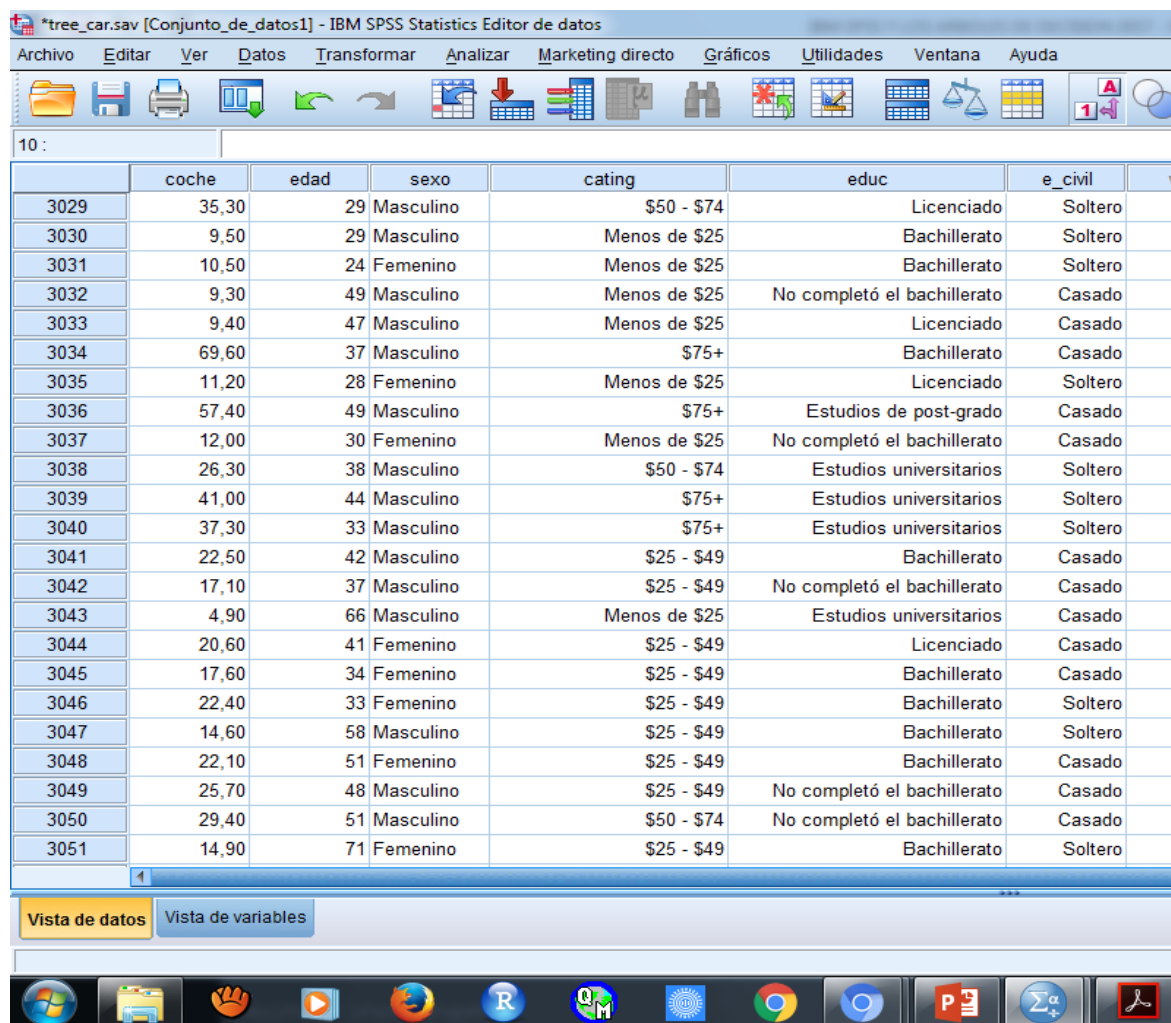
Métodos de crecimiento: CHAID
Variable dependiente: Valoración del crédito

PRACTICA 2. METODOS CRT. PODA DE ARBOLES

Entre los métodos de crecimiento para la creación de árboles de decisión tenemos los métodos CRT y QUEST con las características siguientes:

CRT. Árboles De Clasificación Y Regresión (classification and regression trees). Se trata de un método que divide los datos en segmentos para que sean lo más homogéneos posible respecto a la variable dependiente. Un nodo terminal en el que todos los casos toman el mismo valor en la variable dependiente es un nodo homogéneo y “puro”.

Partiendo del archivo tree_car.sav que contiene datos sobre coches, vamos a construir ahora un árbol de decisión en el que el precio del vehículo dependerá de la edad en años, sexo, categoría de ingresos, nivel de estudios y estado civil del cliente (figura 1).



	coche	edad	sexo	cating	educ	e_civil	v
3029	35,30	29	Masculino	\$50 - \$74	Licenciado	Soltero	
3030	9,50	29	Masculino	Menos de \$25	Bachillerato	Soltero	
3031	10,50	24	Femenino	Menos de \$25	Bachillerato	Soltero	
3032	9,30	49	Masculino	Menos de \$25	No completó el bachillerato	Casado	
3033	9,40	47	Masculino	Menos de \$25	Licenciado	Casado	
3034	69,60	37	Masculino	\$75+	Bachillerato	Casado	
3035	11,20	28	Femenino	Menos de \$25	Licenciado	Soltero	
3036	57,40	49	Masculino	\$75+	Estudios de post-grado	Casado	
3037	12,00	30	Femenino	Menos de \$25	No completó el bachillerato	Casado	
3038	26,30	38	Masculino	\$50 - \$74	Estudios universitarios	Soltero	
3039	41,00	44	Masculino	\$75+	Estudios universitarios	Soltero	
3040	37,30	33	Masculino	\$75+	Estudios universitarios	Soltero	
3041	22,50	42	Masculino	\$25 - \$49	Bachillerato	Casado	
3042	17,10	37	Masculino	\$25 - \$49	No completó el bachillerato	Casado	
3043	4,90	66	Masculino	Menos de \$25	Estudios universitarios	Casado	
3044	20,60	41	Femenino	\$25 - \$49	Licenciado	Casado	
3045	17,60	34	Femenino	\$25 - \$49	Bachillerato	Casado	
3046	22,40	33	Femenino	\$25 - \$49	Bachillerato	Soltero	
3047	14,60	58	Masculino	\$25 - \$49	Bachillerato	Soltero	
3048	22,10	51	Femenino	\$25 - \$49	Bachillerato	Casado	
3049	25,70	48	Masculino	\$25 - \$49	No completó el bachillerato	Casado	
3050	29,40	51	Masculino	\$50 - \$74	No completó el bachillerato	Casado	
3051	14,90	71	Femenino	\$25 - \$49	Bachillerato	Soltero	

FIGURA 1

Paso siguiente, rellenamos la pantalla de entrada del procedimiento árbol como se indica en la figura 2. Se observa que se va utilizar el método de crecimiento CRT.

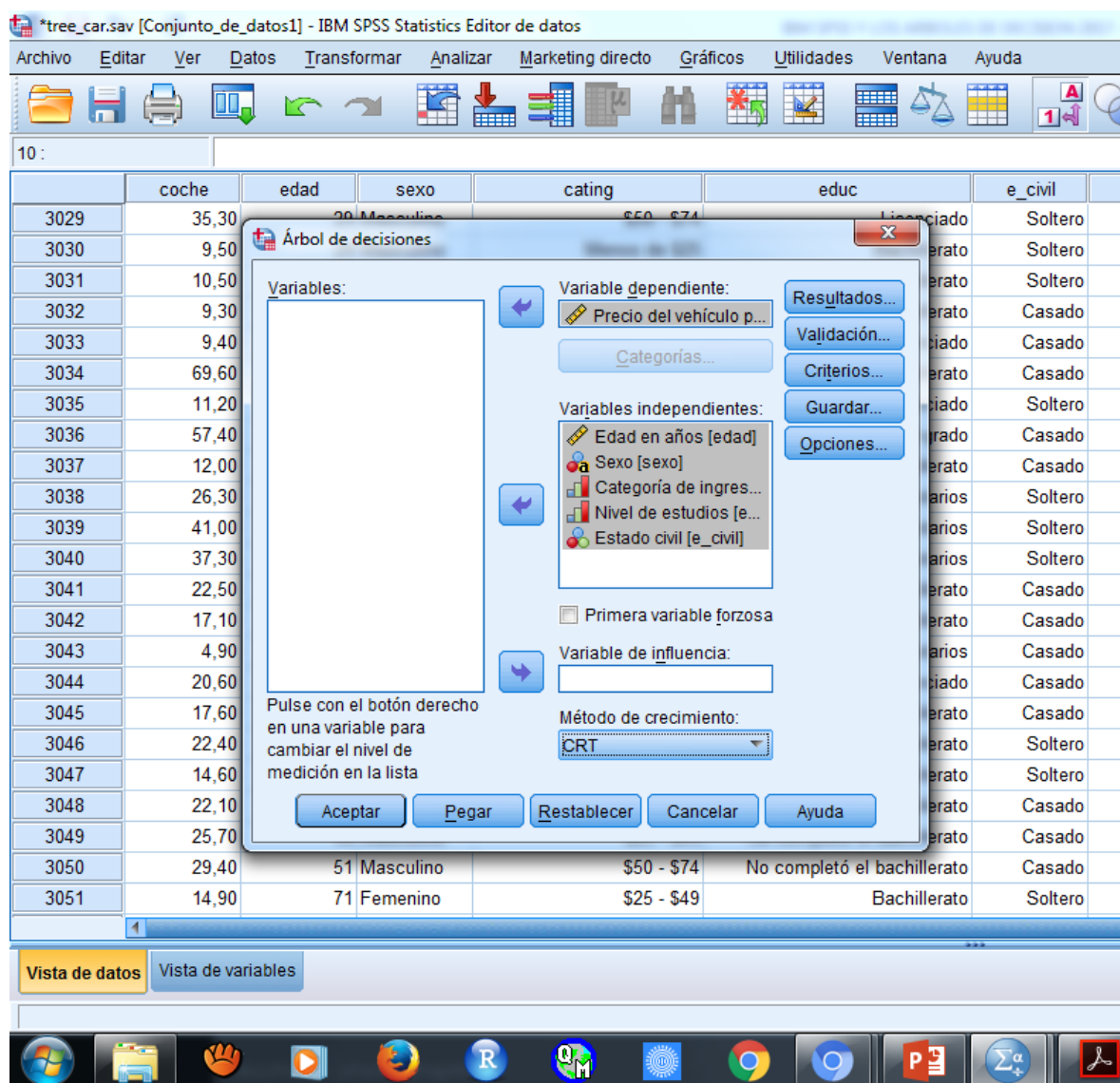


FIGURA 2

El método de crecimiento CRT (Figura 1 procura maximizar la homogeneidad interna de los nodos. El grado en el que un nodo no representa un subconjunto homogéneo de casos es una indicación de impureza. Por ejemplo, un nodo terminal en el que todos los casos tienen el mismo valor para la variable dependiente es un nodo homogéneo que no requiere ninguna división más, ya que es “puro”. Puede seleccionar el método utilizado para medir la impureza necesaria para dividir nodos. Ver figura 3.

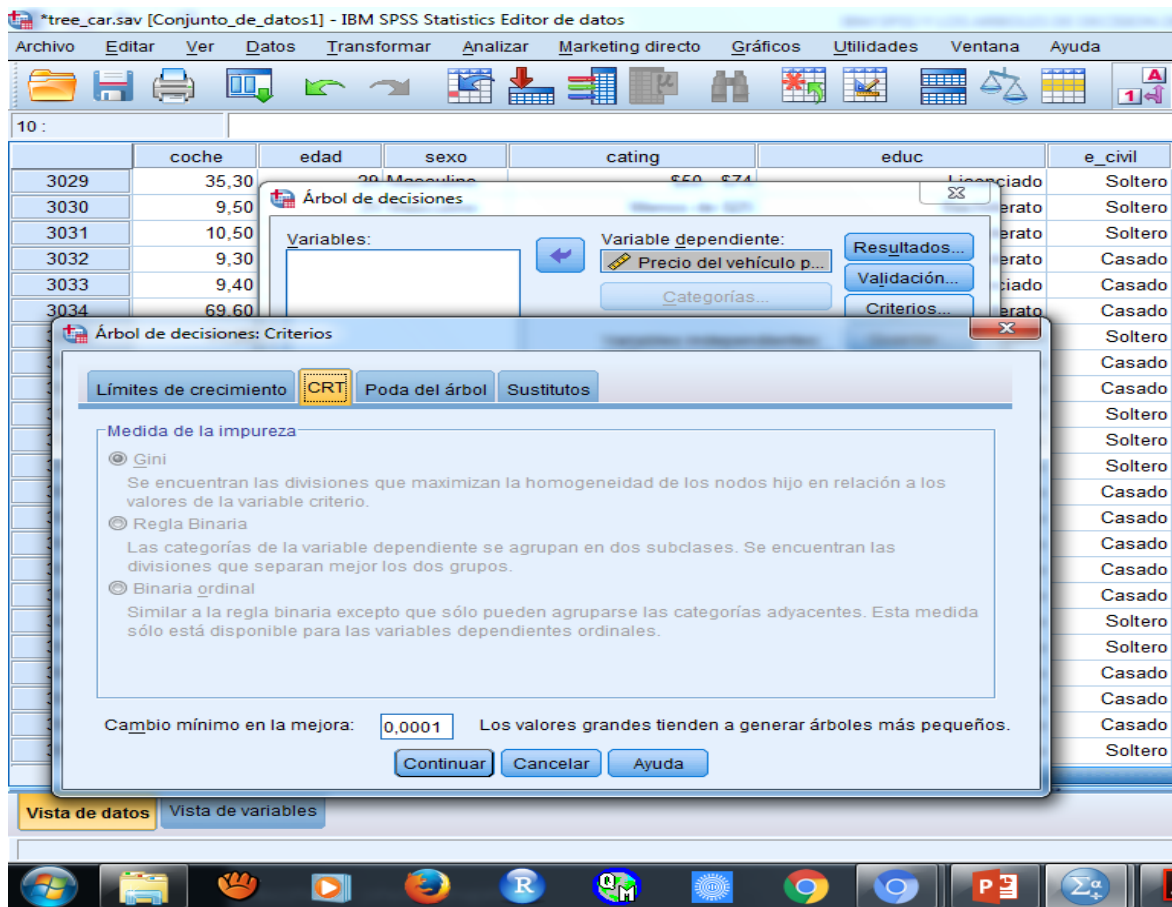


FIGURA 3.

En cuanto a medida de la impureza, para variables dependientes de escala, se utilizará la medida de impureza de desviación cuadrática mínima (LSD). Este valor se calcula como la varianza dentro del nodo, corregida para todas las ponderaciones de frecuencia o valores de influencia.

Para variables dependientes categorías (nominales, ordinales), puede seleccionar la medida de la impureza GINI (se obtienen divisiones que maximizan la homogeneidad de los nodos filiales con respecto al valor de la variable dependiente y se basa en el cuadrado de las probabilidades de pertenencia de cada categoría de la variable dependiente), Binaria (las categorías de la variable dependiente se agrupan en dos subclases y se obtienen las divisiones que mejor separan los dos grupos) y Binaria ordinal (similar a la regla binaria, con la única diferencia de que solo se pueden agrupar las categorías adyacentes). Esta medida solo se encuentra disponible para variables dependientes ordinales. En cuanto a cambio mínimo en la mejora, se trata de situar la reducción mínima de la impureza necesaria para dividir un

nodo. El valor por defecto es 0,0001. Los valores superiores tienden a generar arboles con menos nodos.

Puede evitarse el sobreajuste del modelo mediante la poda del árbol para los métodos CRT y QUEST. El árbol crece hasta que se cumplen los criterios de parada y, a continuación, se recorta de forma automática hasta obtener el subárbol más pequeño basado en la máxima diferencia en el riesgo especificada (figura 4). El valor del riesgo se expresa en errores típicos. El valor por defecto es 1. El valor debe ser no negativo. Para obtener el subárbol con el mínimo riesgo, especifique 0. Inicialmente no activaremos esta opción. Ver salida en la figura 5, de donde se puede observar un árbol muy complicado con muchas ramificaciones y difícil de interpretar.

Obtenido el árbol anterior, repetimos la instrucción activando ilustración de la figura 4 y el árbol podado se podrá observar en la figura 6.

En la figura 7, se presentan salidas del modelo para su análisis.

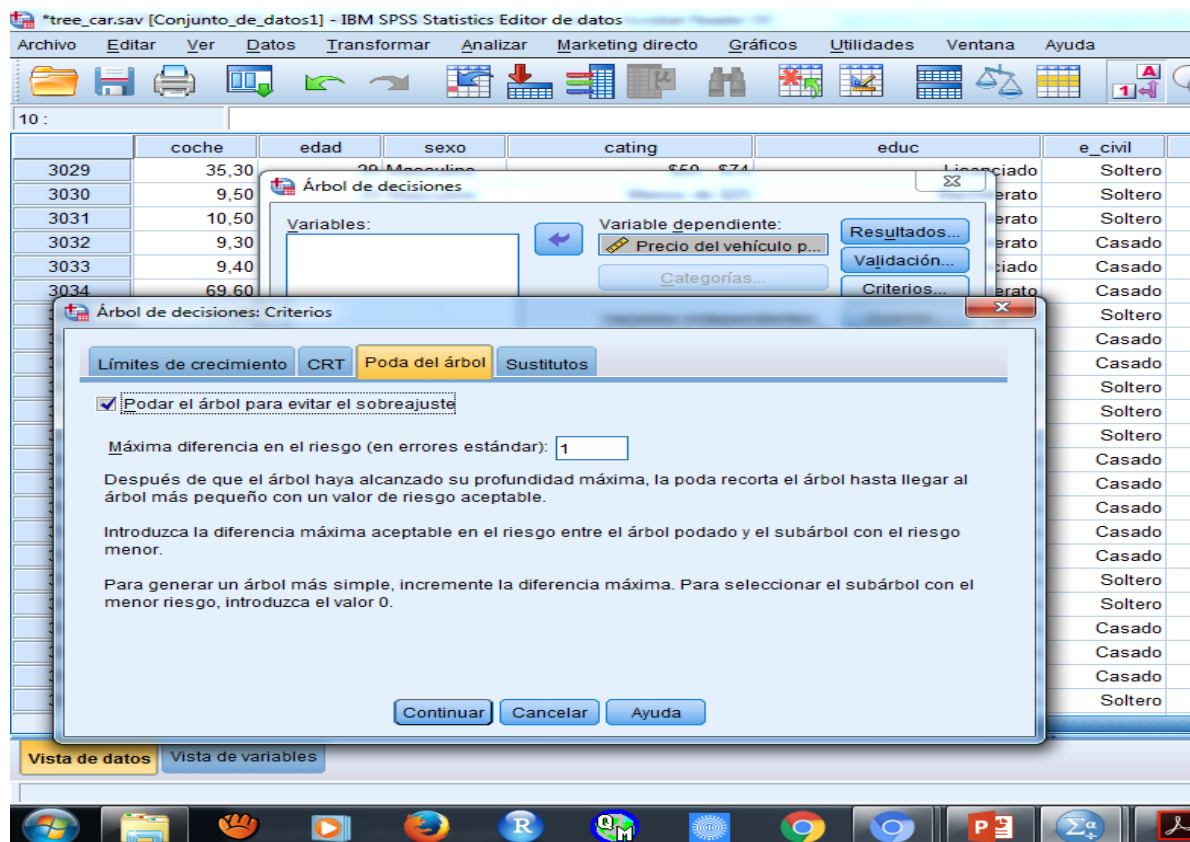


FIGURA 4.

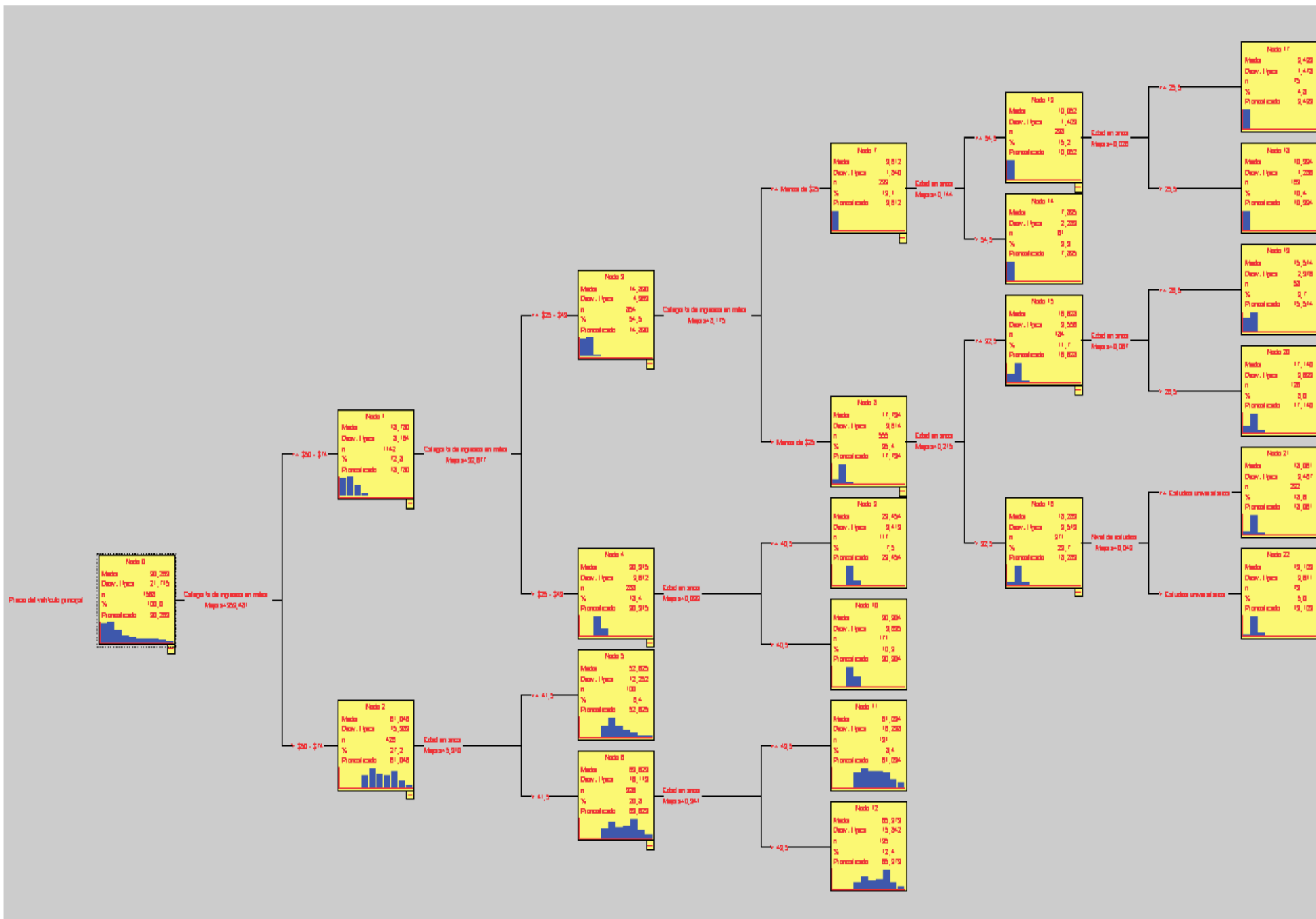
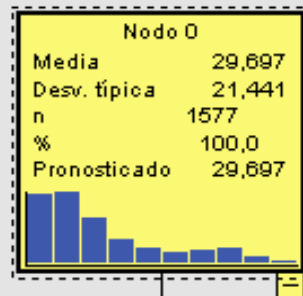


FIGURA 5.

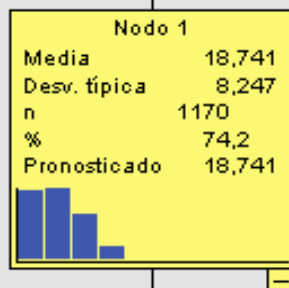
Precio del vehículo principal



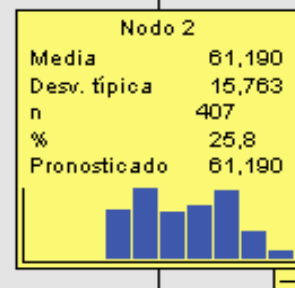
Categoría de ingresos en miles
Mejora=345,024

$\leq \$50 - \74

$> \$50 - \74



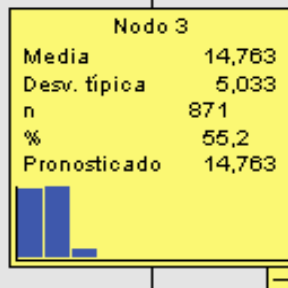
Categoría de ingresos en miles
Mejora=34,201



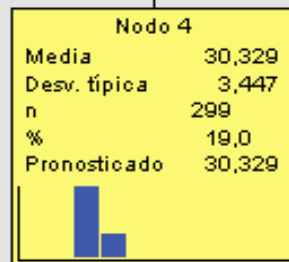
Edad en años
Mejora=5,415

$\leq \$25 - \49

$> \$25 - \49



Categoría de ingresos en miles
Mejora=8,753



\leq Menos de \$25

$>$ Menos de \$25

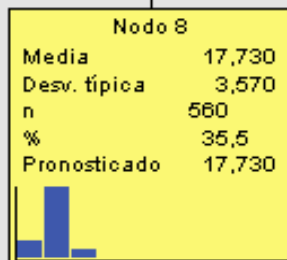
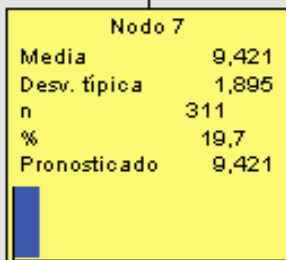


FIGURA 6.

Resumen de ganancia para los nodos

Muestra	Nodo	N	Porcentaje	Media
Entrenamiento	6	323	20,5%	63,5260
	5	84	5,3%	52,2083
	4	299	19,0%	30,3294
	8	560	35,5%	17,7300
	7	311	19,7%	9,4212
Contraste	6	312	20,4%	63,2554
	5	105	6,8%	52,9924
	4	268	17,5%	30,0821
	8	570	37,2%	17,5979
	7	278	18,1%	9,5097

Métodos de crecimiento: CRT

Variable dependiente: Precio del vehículo principal

Riesgo

Muestra	Estimación	Típ. Error
Entrenamiento	66,024	3,797
Contraste	76,931	4,858

Métodos de crecimiento: CRT

Variable dependiente: Precio del vehículo principal

FIGURA 7.

Dado que este modelo tiene variable cuantitativa dependiente, no arroja la salida tabla de clasificación que permita ver o medir la bondad de ajuste del modelo, por ello se utiliza el artificio matemático de la varianza explicada (No explicada) por el modelo en término de las variables relacionas en el árbol.

$$\text{VE}^* (\text{por el modelo}) = (1 - (\text{Estimación del riesgo/varianza estimada en el nodo parental})) * 100\%$$

$$\text{VE}^* = \text{VARIANZA EXPLICADA}$$

$$\text{VE} (\text{por el modelo}) = (1 - (76.93/21.441^2)) = 83.26\%$$

PRACTICA 3. METODOS QUEST. PODA DE ARBOLES

QUEST. Árbol Estadístico Rápido, Insesgado Y Eficiente (Quick, Unbiased, Efficient statistical tree). Se trata de un método que es rápido y que evita el sesgo que presentan otros métodos al favorecer los predictores con muchas categorías. Solo puede especificarse QUEST si la variable dependiente es *nominal*.

Para el método QUEST, puede especificar el nivel de significación para la división de nodos (figura 1). No se puede utilizar una variable independiente para dividir nodos a menos que el nivel de significación sea menor o igual que el valor especificado. El valor debe ser mayor que 0 y menor que 1. El valor por defecto es 0,05. Los valores más pequeños tendrán a excluir más variables independientes del modelo final.

Al pulsar ACEPTAR con las opciones por defecto, se obtiene un árbol muy complicado con demasiadas ramificaciones y difícil de interpretar (figura 3). Para solucionar este problema se hace clic en el botón criterios y se selecciona la pestaña poda de árbol con las opciones por defecto (figura 2). Se hace clic en continuar y aceptar y se obtiene el árbol ya podado que es más fácil de interpretar (figura 4).

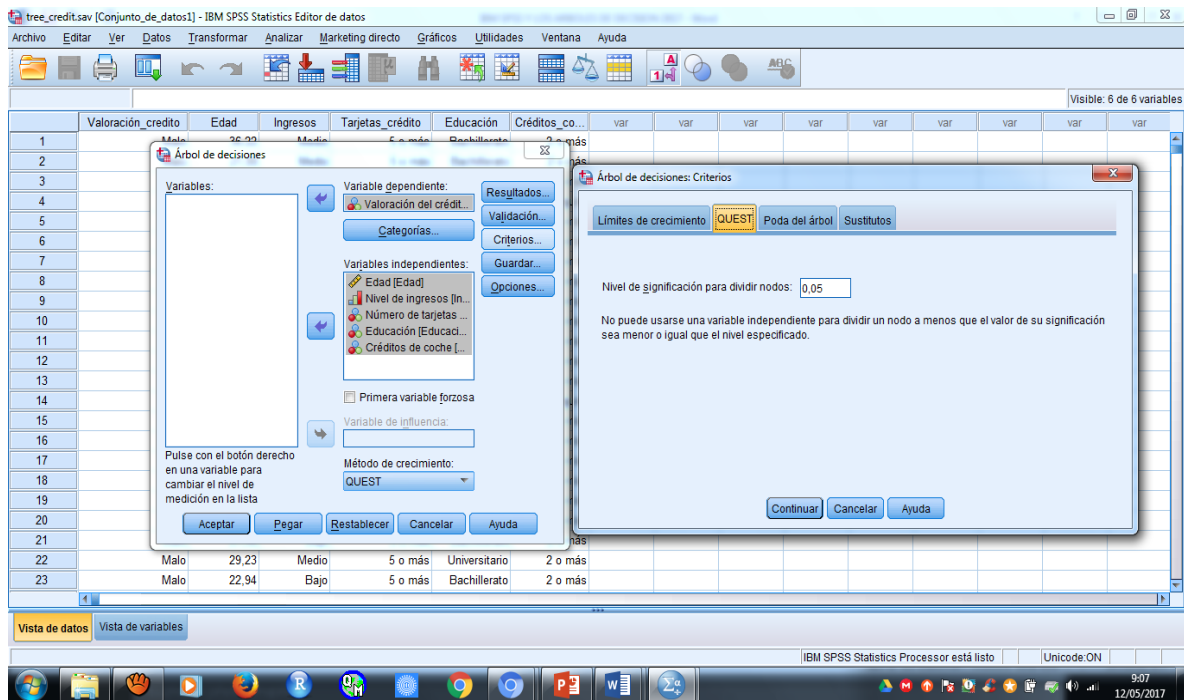


FIGURA 1.

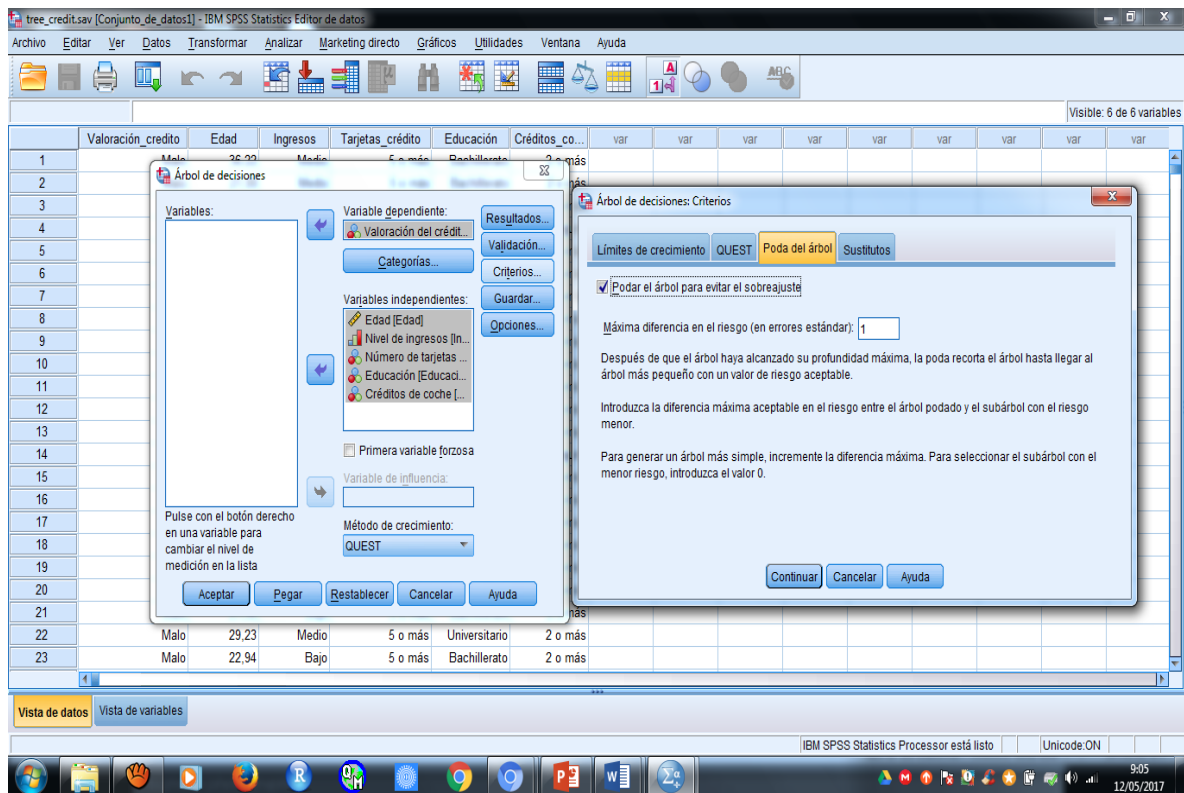


FIGURA 2.

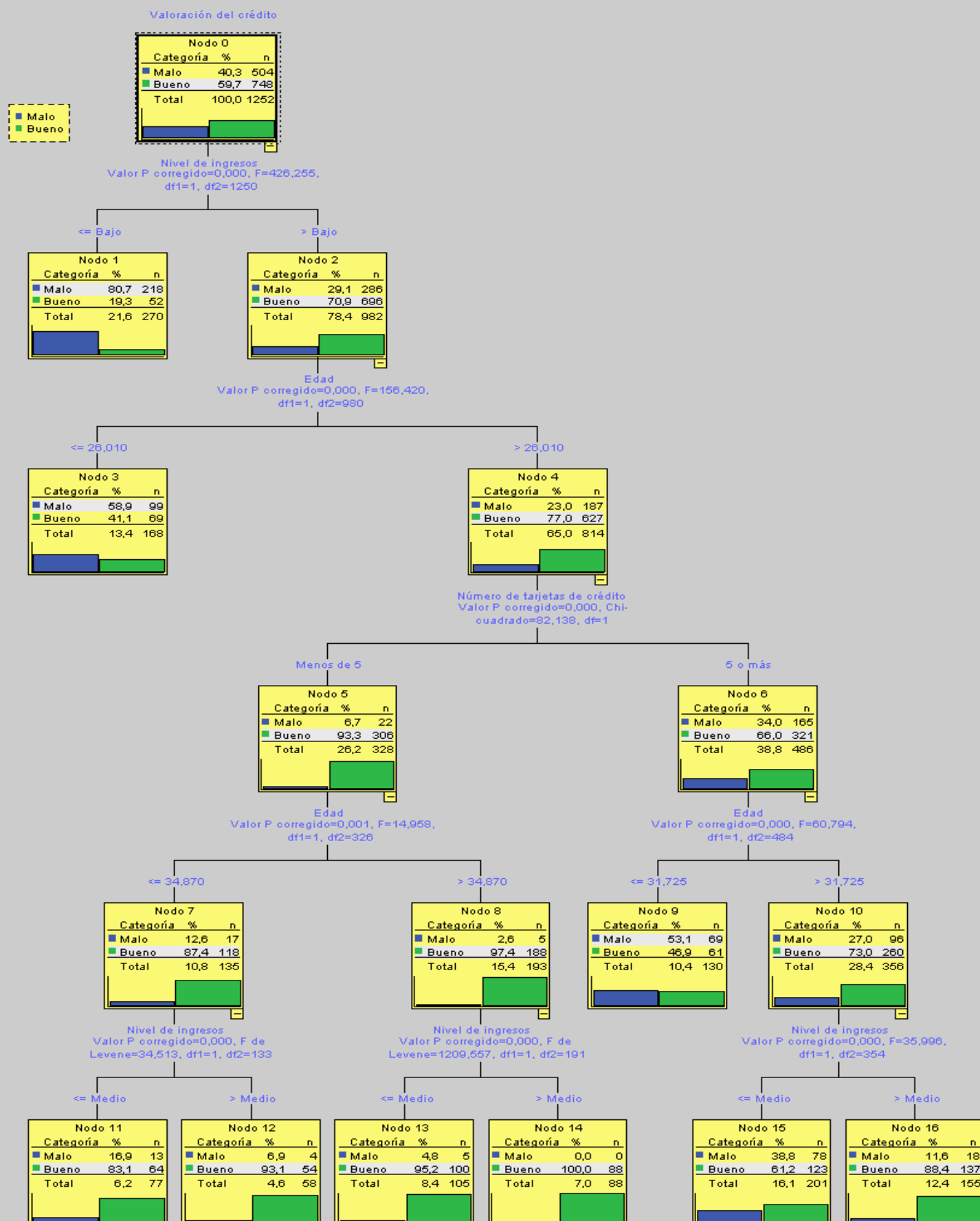


FIGURA 3.

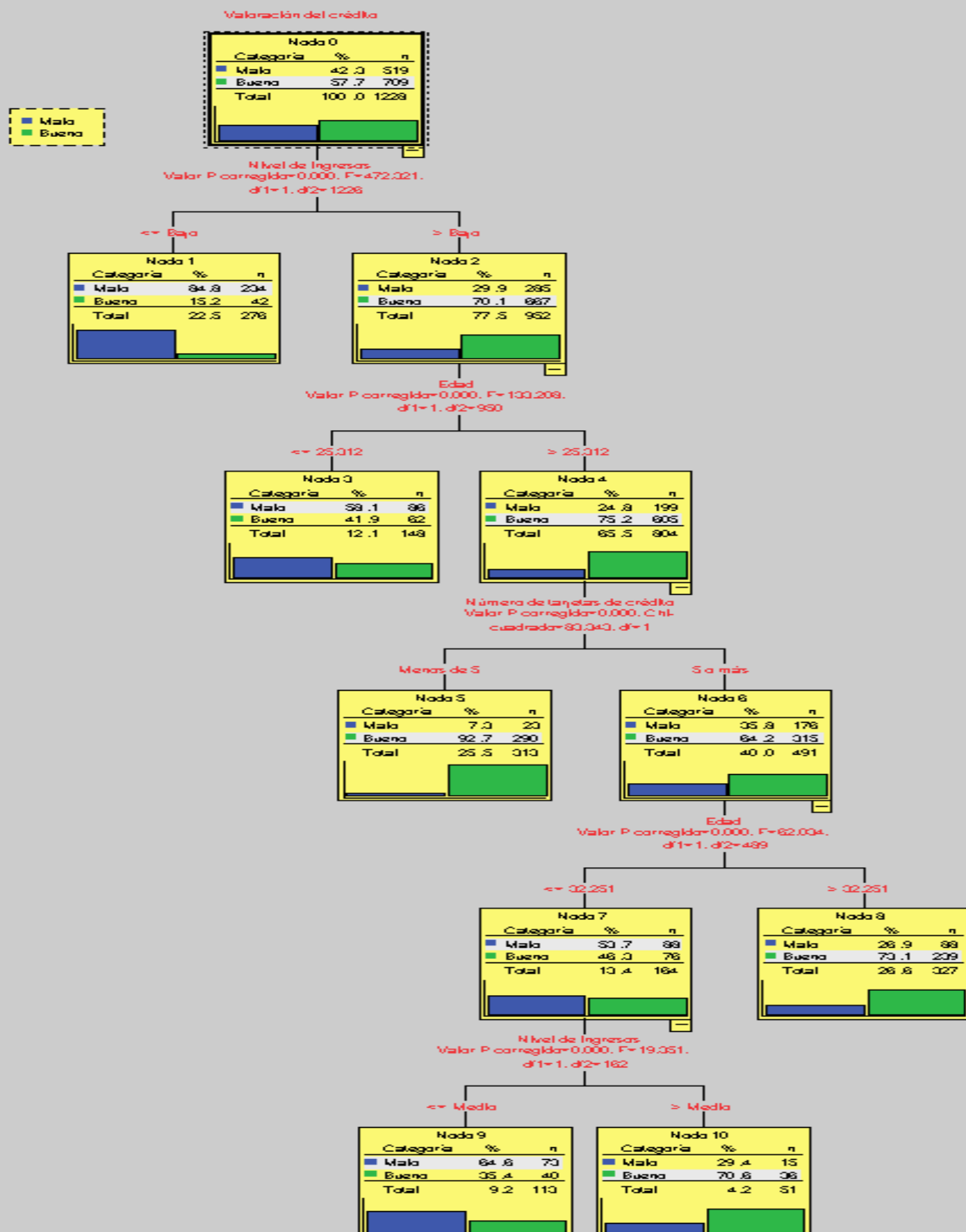


FIGURA 4.