FACULTAD DE ESTUDIOS ESTADÍSTICOS

GRADO EN ESTADISTICA APLICADA

Curso 2024/2025

Trabajo de Fin de Grado

TÍTULO: **Análisis multivariante y modelización predictiva del precio de diamantes: un enfoque comparativo con y sin extracción de factores**

***Alumno*: Cristina Ordaz Solera**

***Tutor*: Eduardo Ortega Castelló**

Junio de 2025

UNIVERSIDAD COMPLUTENSE
MADRID

A quienes me ofrecieron su apoyo y comprensión
en cada etapa de este proceso, incluso en los
momentos en que necesité detenerme para seguir
adelante con más fuerza.

Mi agradecimiento especial a los profesores
Eduardo Ortega, Aida Calviño y José Luis Valencia,
por su acompañamiento, orientación y confianza.

Índice

Índice de Gráficos

Índice de Tablas

1. Introducción

Este estudio busca predecir el precio de los diamantes en función de sus características físicas, utilizando diversas técnicas estadísticas para evaluar la precisión y utilidad de cada enfoque. Para ello, primero se realizará un análisis factorial con el fin de identificar factores latentes que resuman la información de las variables originales y faciliten la interpretación de los resultados.

Con estos factores, se construirá un árbol de regresión que permitirá modelar el precio como una variable continua. Paralelamente, se categorizarán los precios en cuatro niveles (bajo, medio-bajo, medio-alto y alto) para aplicar técnicas de clasificación, incluyendo un análisis discriminante, una regresión logística multinomial y un árbol de clasificación, con el objetivo de evaluar su capacidad predictiva.

Posteriormente, se repetirá el estudio sin la extracción de factores para comparar los resultados y determinar qué enfoque ofrece una mejor interpretación y precisión en la predicción del precio de los diamantes.

Los datos con los que vamos a trabajar fueron extraídos de un mercado online de diamantes utilizando la biblioteca Selenium de phyton. Este dataset recopila información exhaustiva de más de 6,400 diamantes, abarcando sus dimensiones físicas (longitud, ancho, altura y peso en quilates), así como aspectos de calidad, como el corte, color, claridad y fluorescencia, junto con sus precios. Además, incluye detalles sobre el tipo de certificación, proporciones y simetría, lo que lo hace una valiosa fuente para estudiar el impacto de distintos factores en el valor de los diamantes. Las variables son las siguientes:

Forma: Tipo de figura geométrica del diamante.
Corte: Calidad de la talla del diamante.
Color: Clasificación del color del diamante, de D a H.
Claridad: Grado de pureza basado en la presencia de imperfecciones.
Peso en quilates: Peso del diamante medido en quilates.
Relación largo/ancho: Proporción entre la longitud y el ancho del diamante.
Profundidad %: Profundidad del diamante expresada como porcentaje de su ancho.
Tabla %: Ancho de la faceta superior expresado como porcentaje.
Pulido: Calidad del acabado superficial del diamante.
Simetría: Precisión en la forma del diamante.
Faja: Grosor del borde del diamante.
Culet: Tamaño de la faceta inferior.
Longitud: Medida del largo del diamante en milímetros.
Ancho: Medida del ancho del diamante en milímetros.
Altura: Medida de la altura del diamante en milímetros.
Precio: Valor del diamante en dólares estadounidenses ($).
Tipo: Certificación o tipo de origen del diamante.
Fluorescencia: Nivel de fluorescencia del diamante bajo luz ultravioleta.

Como suele ocurrir en el análisis de datos reales, se presentaron algunos desafíos al procesar la información, lo que requirió un enfoque meticuloso para garantizar la calidad y fiabilidad de los resultados. Para comenzar, se realizó un análisis exploratorio descriptivo con el objetivo de comprender la estructura de los datos, evaluar su distribución y detectar posibles irregularidades.

Uno de los primeros pasos fue la identificación de valores atípicos en cada variable. Se consideró que, en ciertas variables con alta dispersión, algunos valores que inicialmente podrían parecer atípicos no necesariamente lo eran. Para evitar la eliminación indebida de información relevante, se emplearon técnicas gráficas y estadísticas, como diagramas de caja (boxplots), con el fin de contextualizar mejor la presencia de estos valores.

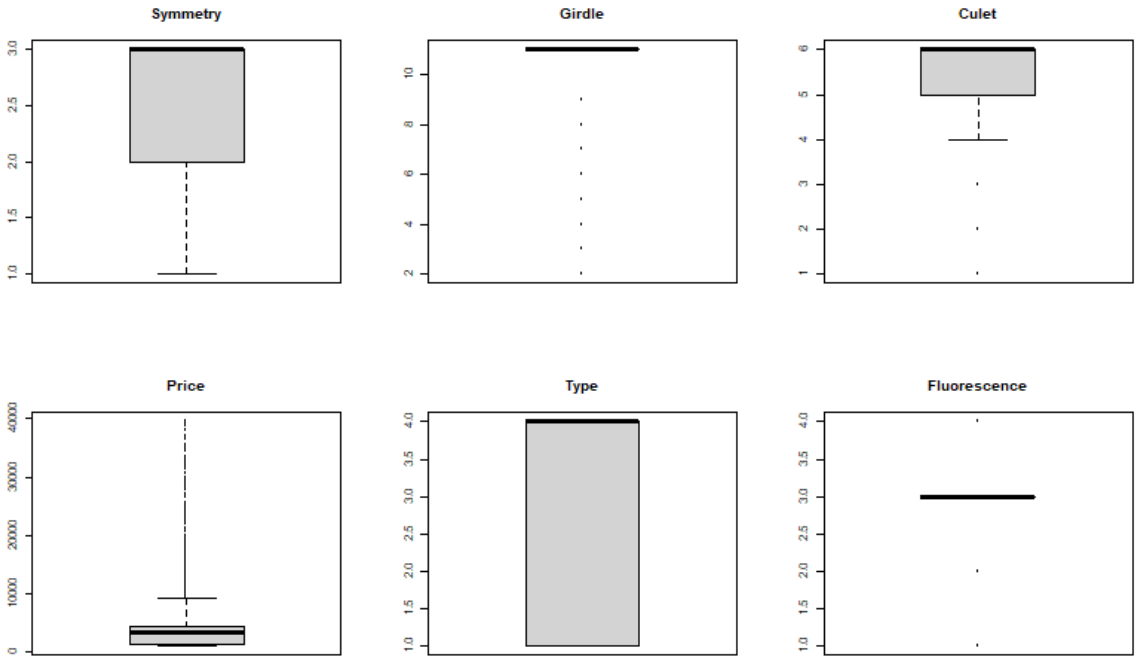


Gráfico 1: Boxplots primera agrupación de variables.

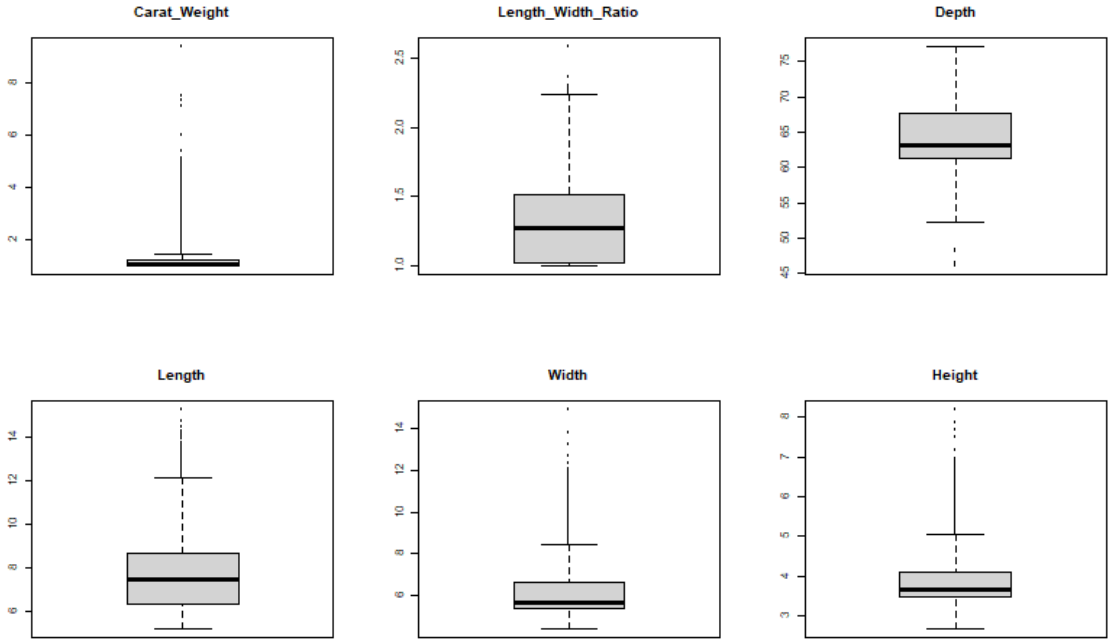


Gráfico 2: Boxplots segunda agrupación de variables.

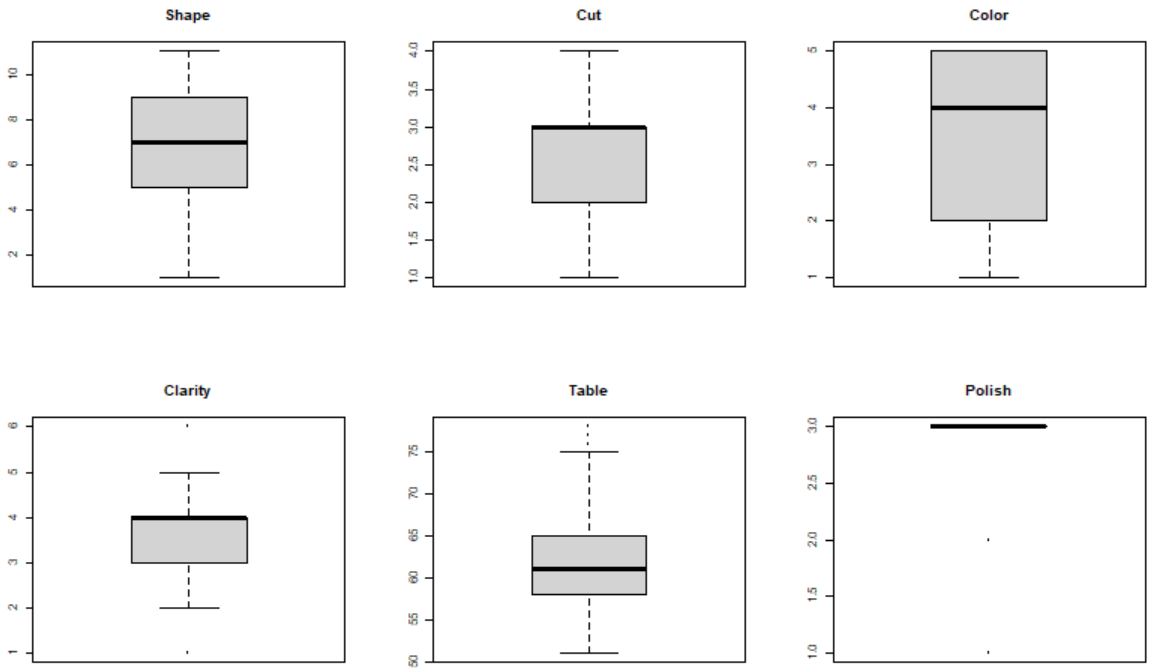


Gráfico 3: Boxplots tercera agrupación de variables.

A la vista de los resultados de los boxplots, se revisa la desviación estándar de las variables Clarity, Polish, Carat_Weight y Girdle (aquellas que cuentan con más de un 5% de datos atípicos), ya que los datos atípicos podrían originarse debido a una gran amplitud en los valores. Todas cuentan con una alta desviación estándar lo que sugiere que los datos están dispersos y podrían incluir valores extremos que no necesariamente son errores. Además, se llevó a cabo un análisis de los datos ausentes para determinar su distribución y magnitud dentro del conjunto de datos.

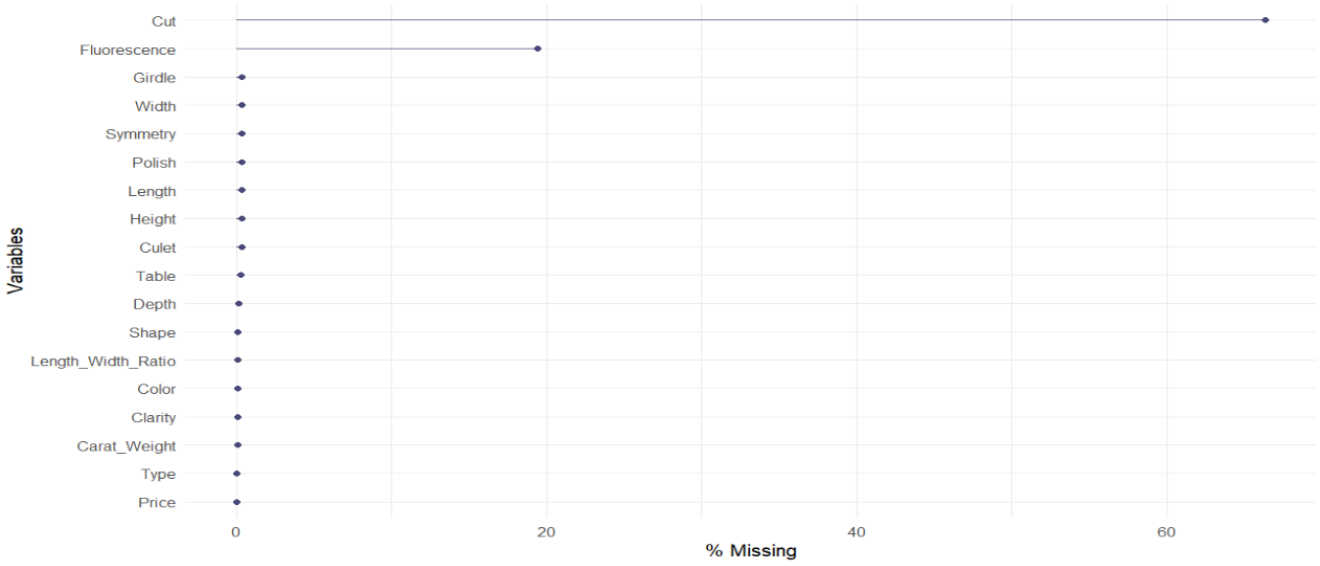


Gráfico 4: Datos missing de la base.

Se identificó que las variables Corte y Fluorescencia presentaban una cantidad significativa de valores faltantes, con tasas de ausencia del 60% y 20%, respectivamente.

Debido a la elevada proporción de datos ausentes en Corte, se optó por excluir esta variable del análisis, ya que una imputación extensiva podría introducir un sesgo considerable. En el caso de Fluorescencia, aunque el porcentaje de valores faltantes era menor, también se decidió excluirla para evitar afectar la calidad de los resultados.

Eliminando ambas variables, se vuelve a graficar los datos ausentes. Se aprecia que ninguna variable cuenta con más de un 0.5% de ausentes.

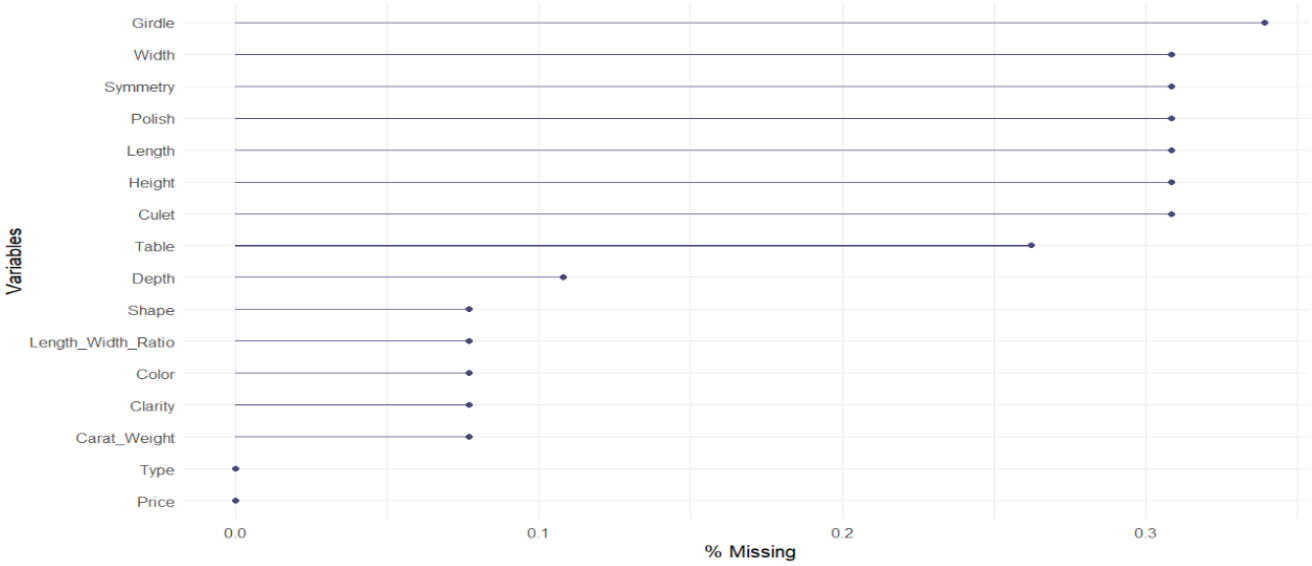


Gráfico 5: Datos missing de la base tras la eliminación de variables.

Para estas variables, se aplicó un procedimiento de imputación basado en el método CART (Classification and Regression Trees). Este método utiliza árboles de clasificación y regresión para estimar los valores faltantes en función de la relación existente entre las demás variables. Se eligió este enfoque debido a su capacidad para manejar tanto variables numéricas como categóricas, así como su flexibilidad para capturar relaciones no lineales en los datos.

La imputación con CART permitió completar el conjunto de datos de manera estructurada, minimizando la pérdida de información y reduciendo el impacto de los valores ausentes en los análisis posteriores. Esta estrategia aseguró que los modelos desarrollados posteriormente, tanto de regresión como de clasificación, contaran con una base de datos más robusta y representativa de la realidad.

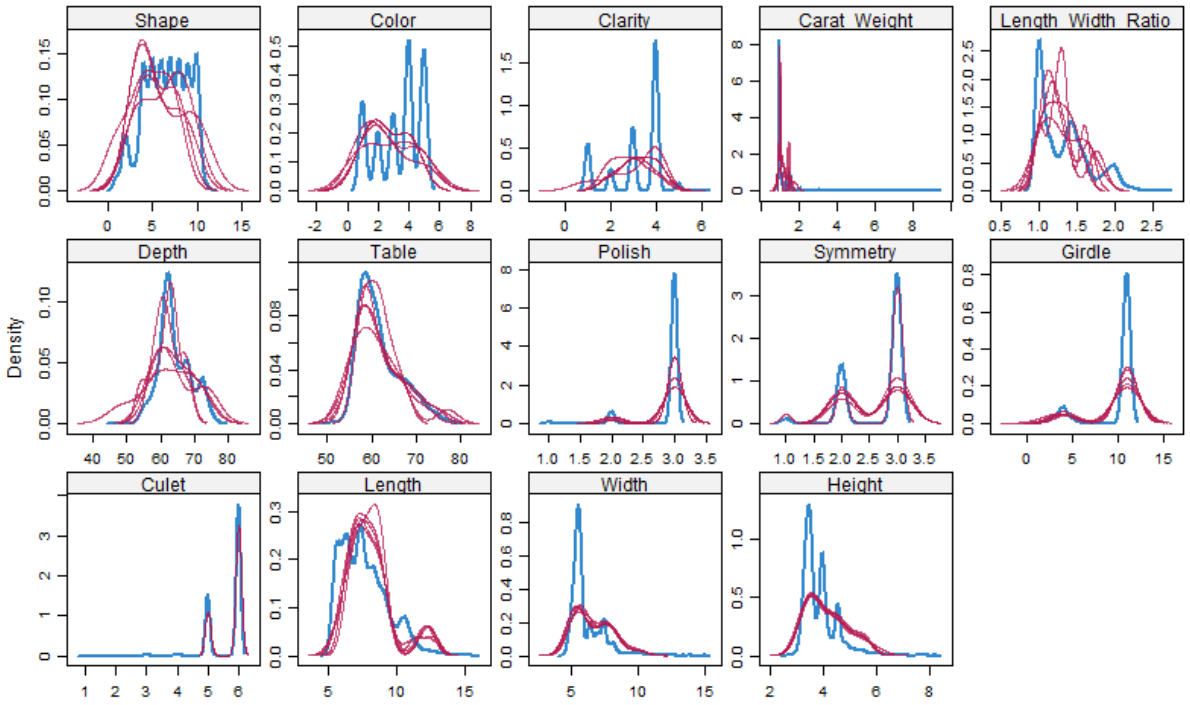


Gráfico 6: Imputación de las variables.

Los datos graficados en color azul son los datos originales de la base mientras que los datos graficados en color rojo son aquellos que se han imputado. Se observa que al haber realizado la imputación con el método CART se ajustó correctamente las observaciones missing a la naturaleza de las variables.

Tal como se ha evidenciado previamente en los gráficos, es previsible la aparición de problemas relacionados con la normalidad de los datos al aplicar determinadas técnicas estadísticas, como el análisis discriminante. No obstante, un aspecto favorable es que existen otros métodos estadísticos que no requieren el supuesto de normalidad para su correcta aplicación, lo que permite continuar con el análisis sin comprometer la validez de los resultados.

2. Estudio a partir del Análisis factorial

A partir de la matriz de covarianzas o de correlaciones de un conjunto de variables, es posible identificar las relaciones que existen entre ellas. El objetivo principal del análisis factorial es representar dichas relaciones mediante un número reducido de variables subyacentes e inobservables, conocidas como factores (Valencia Delfa & Vicente Hernanz, 2015).
Sabiendo esto, lo primero que se debe hacer es examinar la matriz de correlaciones. Se saca por pantalla siendo:

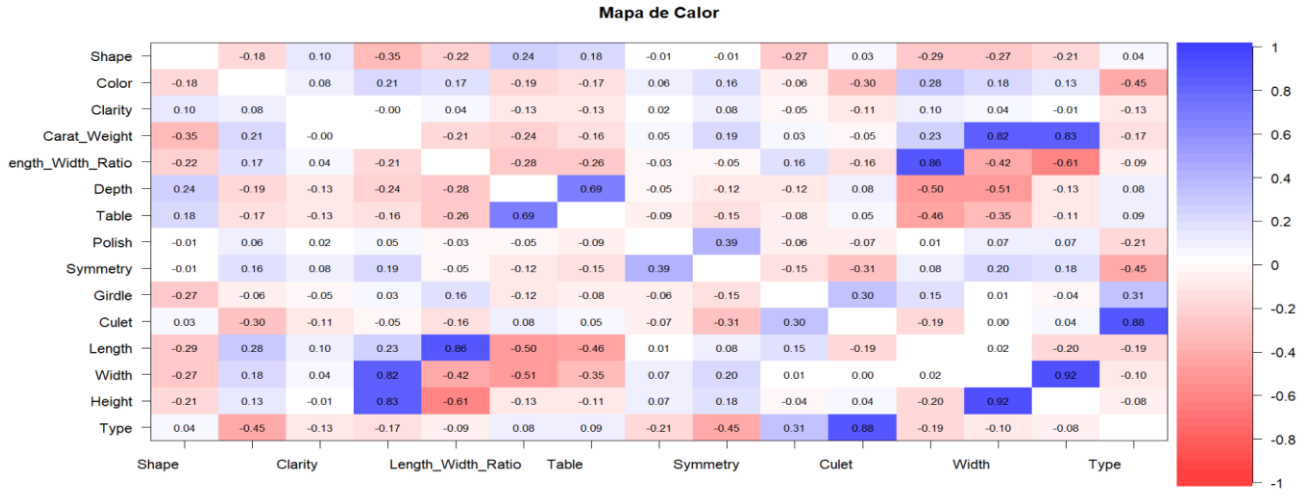


Gráfico 7: Matriz de correlaciones.

Como se puede observar, se cuenta con varias correlaciones altas y esto puede ser un buen inicio para comprobar si procede el análisis factorial. A continuación, se realiza el test de esfericidad de Bartlett. Con este test comprobamos si la matriz de correlaciones se ajusta a la matriz identidad. Si esto pasara, significaría una ausencia de correlación significativa entre las variables y el análisis factorial no procedería (Valencia Delfa & Vicente Hernanz, 2015). Las hipótesis planteadas son las siguientes:

$$H0: La\ matriz\ de\ correlaciones\ se\ ajusta\ a\ la\ matriz\ identidad$$
$$H1: La\ matriz\ de\ correlaciones\ no\ se\ ajusta\ a\ la\ matriz\ identidad$$

El estadístico en este caso tiene el valor de 3,18357e-118, el cual es prácticamente 0, por lo que rechazamos la hipótesis nula. Se puede afirmar que la matriz de correlaciones no se ajusta a la matriz identidad, por tanto, el análisis factorial por este criterio procede.

A continuación, se comprueba el Índice de Kaiser-Meyer-Olkin. Este índice mide el grado de correlación parcial entre las variables para determinar si los datos son apropiados para extraer factores significativos. En este caso:

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = datos)
Overall MSA =  0.51
```

Se cuenta con un KMO de 0,51. No es un valor excelente, se trata más bien de un valor bajo, pero con el que se puede proceder con el análisis factorial.

El siguiente paso será elegir el número de factores con los que agruparemos nuestras variables. Para ello, se procede a crear un gráfico para comparar por diversos métodos cuál sería el número óptimo de factores a retener.

Eigenvalues: representan los autovalores de la matriz de correlaciones. El criterio más común es considerar factores con autovalores mayores a 1.

Parallel Analysis: es un método que compara los autovalores observados con autovalores generados aleatoriamente. Se seleccionan aquellos cuyos autovalores sean mayores que los de la simulación.

Optimal Coordiantes: es un método que ajusta una curva a los autovalores y define un punto óptimo donde seleccionar los factores.

Acceleration Factor: identifica el punto de mayor cambio en la pendiente de los autovalores.

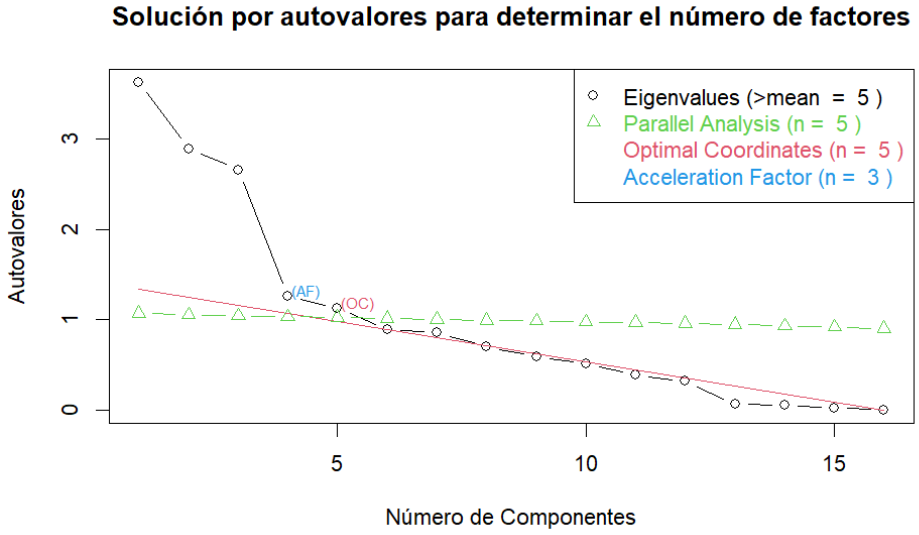


Gráfico 8: Factores a retener.

Se observa como en todos los criterios (menos en Acceleration Factor) es preferible retener 5 factores. Para mejorar la interpretabilidad de los factores, procedemos con una rotación. Dada la naturaleza de los datos, la rotación elegida es la Varimax. La rotación Varimax es una técnica de rotación ortogonal empleada en el análisis factorial que tiene como objetivo facilitar la interpretación de los factores obtenidos. En este caso, se obtiene:

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Shape	-0,217		-0,2	0,112	0,688
Color	0,139	-0,434	0,2		-0,11
Clarity		-0,123		-0,138	0,191
Carat_Weight	0,901	-0,139	0,158		-0,13
Length_Width_Ratio	-0,419		0,848	-0,182	-0,179
Depth	-0,139		-0,165	0,996	
Table	-0,152		-0,249	0,64	
Polish		-0,206			
Symmetry	0,145	-0,442			0,104
Girdle		0,321	0,134		-0,301
Culet	0,102	0,881			
Length		-0,115	0,928	-0,334	
Width	0,874	-0,11	-0,184	-0,419	-0,104
Height	0,951		-0,276		
Type		0,993			

Tabla 1: Factores.

Con estos resultados, se procede a caracterizar los factores de la siguiente manera:

Factor 1: Tamaño.
Factor 2: Calidad.
Factor 3: Geometría visual.
Factor 4: Proporción y brillo.
Factor 5: Forma del diamante.

Por último, se revisa cuanta proporción de la varianza explicada se consigue con esta transformación. La varianza explicada refleja cuánta información del conjunto de datos original está siendo capturada por los factores extraídos. Como se puede apreciar, se cuenta con una proporción de varianza explicada superior a 0.6 por lo que el análisis ha sido adecuado.

	Factor1	Factor2	Factor3	Factor4	Factor5
SS loadings	2.800	2.374	1.923	1.728	0.709
Proportion Var	0.187	0.158	0.128	0.115	0.047
Cumulative Var	0.187	0.345	0.473	0.588	0.636

Tabla 2: Factores retenidos y varianza explicada.

2.1. Análisis discriminante con las clases creadas

La discriminación y la clasificación son técnicas multivariantes que envuelven la separación de observaciones y la posterior ubicación de nuevas observaciones en alguno de los grupos previamente definidos (Valencia Delfa & Vicente Hernanz, 2015).

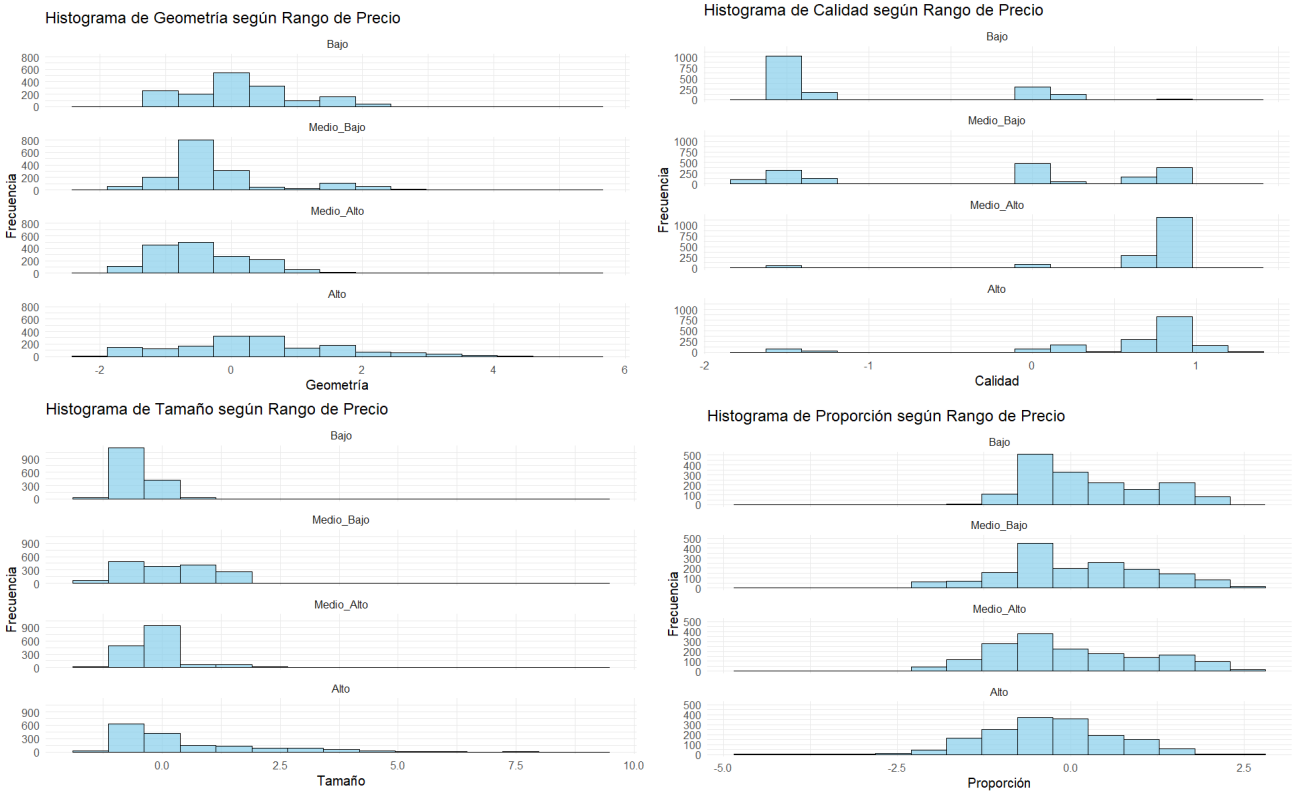
Para llevar a cabo este análisis, recodificamos nuestra variable en 4 niveles, teniendo en cuenta los cuartiles. Por tanto, por debajo de 1210 se considerará clase X1 (la que cuenta con el menor precio), entre este valor y 3320 se considerará precio medio-bajo siendo X2, entre 3320 y 4390 se considerará precio medio-alto y mayor que 4390 se considerará precio alto.

0%	25%	50%	75%	100%
1010	1210	3320	4390	39460

Tabla 3: Cuartiles de la variable precio.

Para proceder con el análisis discriminante se deben cumplir varios supuestos. El primero de todos, es el supuesto de Normalidad. Las variables que intervienen en el análisis deben al menos recordar a una distribución normal en cada una de las categorías de la variable dependiente (en nuestro caso precio). Las hipótesis son las siguientes:

$$H0: Los\ datos\ provienen\ de\ una\ distribución\ Normal$$
$$H1: Los\ datos\ no\ provienen\ de\ una\ distribución\ Normal$$



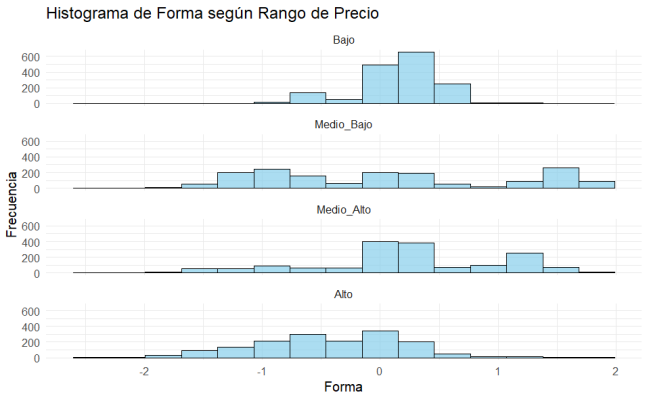


Gráfico 9: Histogramas de las variables independientes.

Podemos confirmar que, en casi todas las variables y categorías, recuerdan un poco a una Normal. La única variable que no cumple la hipótesis es Calidad. Se lleva a cabo una transformación de la variable Calidad para lograr normalidad. Como cuneta con valores negativos, la transformación Box-Cox no será eficiente en este caso. Por esto, se lleva a cabo una función de R llamada bestNormalize que ayuda a encontrar que transformación es más apropiada para cada variable. En este caso, aplicamos la función a la variable Calidad y se aconseja usar el Cuantil Ordenado. La función Cuantil Ordenado combina dos procesos:

- Transformación de Rangos: Ordena los datos y les asigna puntuaciones basadas en su posición relativa dentro del conjunto.

- Transformación Probit Inversa: Aplica una transformación que mapea los percentiles obtenidos en la primera etapa a los valores correspondientes de una distribución normal estándar, aproximando así los datos a una distribución normal con media cero y desviación estándar uno.

La transformación de normalización denominada Cuantil Ordenado es un procedimiento basado en rangos en el cual los valores de un vector se asignan a su correspondiente percentil. Se consigue como resultado:

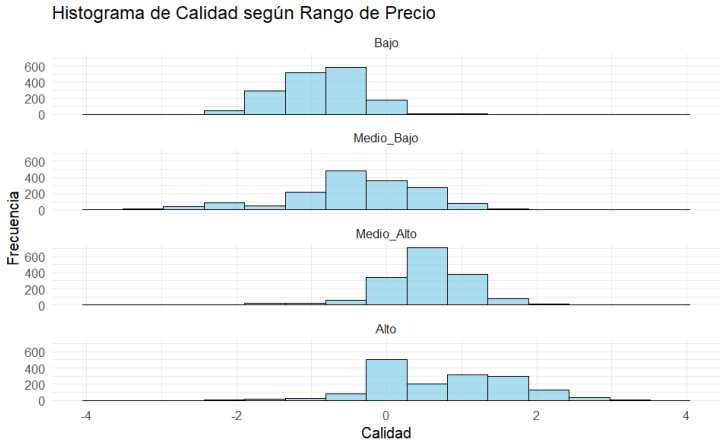


Gráfico 10: Histograma tras la normalización de una variable.

Con todas las variables ya similares a normales, continuamos con las comprobaciones para poder proceder con el análisis. Se lleva a cabo un test de homogeneidad de la varianza. Las hipótesis son:

$$H0: Las\ varianzas\ no\ son\ significativamente\ diferentes$$
$$H1: Las\ varianzas\ son\ significativamente\ diferentes$$

En caso de aceptar H0 será conveniente hacer un análisis lineal, mientras que si rechazamos H0 será apropiado realizar un análisis cuadrático. El test de Levenne nos proporciona un p_valor de 2.2e-16 en todas las categorías rechazando la hipótesis nula. Con este resultado, se procede a hacer un análisis discriminante cuadrático, obteniendo las siguientes ecuaciones:

$$X1 = -0.506 * \text{Tamaño} - 0.895 * \text{Calidad} + 0.165 * \text{Geometría} + 0.272 * \text{Proporción} + 0.149 * \text{Forma} + 0.252$$

$$X2 = 0.177 * \text{Tamaño} - 0.370 * \text{Calidad} - 0.199 * \text{Geometría} + 0.77 * \text{Proporción} + 0.052 * \text{Forma} + 0.249$$

$$X3 = -0.079 * \text{Tamaño} + 0.523 * \text{Calidad} - 0.388 * \text{Geometría} + 0.013 * \text{Proporción} + 0.221 * \text{Forma} + 0.250$$

$$X4 = 0.419 * \text{Tamaño} + 0.750 * \text{Calidad} + 0.434 * \text{Geometría} - 0.281 * \text{Proporción} - 0.431 * \text{Forma} + 0.249$$

A continuación, nos fijamos en la matriz de confusión, donde en verde se señalan los aciertos y en rojo los fallos:

	X1	X2	X3	X4
X1	315	60	0	0
X2	0	167	13	7
X3	3	73	256	96
X4	5	27	55	219

Tabla 4: Tabla de contingencia discriminante.

Para comprender mejor esto, utilizamos el Hit Ratio. Esta medida pondera los casos acertados entre el número total de observaciones.

$$Hit\ Ratio = \frac{Número\ de\ aciertos}{Número\ de\ observaciones}$$

$$Hit\ Ratio = \frac{957}{1296} = 0.738$$

Acertamos en un 73.8% utilizando las reglas discriminantes creadas. También podemos conseguir el Índice de Significación Práctica para cada grupo. Representa la mejora de la proporción de observaciones acertadas de la clasificación mediante las reglas creadas con respecto al conseguido al azar (Valencia Delfa & Vicente Hernanz, 2015).

$$ISP = \frac{o - e}{n - e}$$

Siendo:
O: bien clasificados
E: bien clasificados al azar
N: tamaño muestral

$$ISP_{X1} = \frac{315-(0.25^2*375)}{375-(0.25^2*375)} = 0.829 \qquad ISP_{X2} = \frac{167-(0.25^2*187)}{187-(0.25^2*187)} = 0.886$$

$$ISP_{X3} = \frac{256-(0.25^2*428)}{428-(0.25^2*428)} = 0.571 \qquad ISP_{X4} = \frac{219-(0.25^2*306)}{306-(0.25^2*306)} = 0.697$$

$$ISP = \frac{(315+167+256+219)-(0.25^2*1296)}{1296-(0.25^2*1296)} = 0.721$$

Se observa que todos los valores del Índice de Significación Práctica superan el umbral de 0.25, que corresponde al rendimiento esperado en un escenario aleatorio, es decir, el número de aciertos que se obtendrían por puro azar. Este resultado sugiere que las reglas generadas a través del análisis discriminante tienen una capacidad significativa para clasificar correctamente los casos, ya que los índices obtenidos son considerablemente más altos que los que se esperarían sin ningún modelo predictivo.

Por último, se evalúa la sensibilidad y especificidad del modelo a partir de la tabla de contingencia anteriormente mostrada. La sensibilidad es la capacidad del modelo para identificar correctamente las observaciones que pertenecen a un grupo específico mientras que la especificidad es la capacidad del modelo para identificar correctamente las observaciones que no pertenecen a un grupo particular Valencia Delfa & Vicente Hernanz, 2015).

$$Sensibilidad = \frac{Verdaderos\ positivos}{Total\ de\ positivos\ reales}$$

$$Especificidad = \frac{Verdaderos\ negativos}{Total\ de\ negativos\ reales}$$

Aplicando las fórmulas se consiguen los siguientes resultados:

	Sensibilidad	Especificidad
X1	0.840	0.992
X2	0.893	0.859
X3	0.598	0.924
X4	0.716	0.899

Tabla 5: Sensibilidad y especificidad discriminante.

El modelo discriminante presenta un rendimiento general satisfactorio, con una adecuada capacidad para clasificar a los individuos en sus respectivas categorías. Las clases X1 y X2 muestran altos niveles de sensibilidad y especificidad, lo que indica una correcta identificación de casos y una baja tasa de errores desde otras categorías. Esto sugiere una buena utilidad discriminativa de las variables en estos grupos.

En contraste, las clases X3 y X4 presentan una sensibilidad moderada, lo que dificulta la detección precisa de los casos reales en estas categorías. Aunque la especificidad se mantiene en niveles aceptables, la menor sensibilidad afecta la discriminación equilibrada entre todas las clases.

2.2. Regresión Logística Multinomial con las clases creadas

Los modelos de Regresión Logística Multinomial son aquellos que tienen como variable objetivo una de tipo cualitativa con varios niveles. (Alonso Revenga & Calviño Martínez, 2025). Para ello, se vuelve a crear una partición de los datos en entrenamiento (80% de la muestra) y prueba (20% de la muestra) para poder crear el modelo y posteriormente validarlo. Hay que tener en cuenta que los modelos creados a partir de regresión logística multinomial toman como referencia la primera clase, en este caso, precio bajo para los diamantes.
Se empieza creando un modelo con todas las variables y se comprueba que el proceso converja (nos indica que el modelo resultante es óptimo). Se lleva a cabo un análisis de tipo II para saber qué variables son significativas, a partir de la función Anova:

```
Analysis of Deviance Table (Type II tests)

Response: Precio
           LR Chisq Df Pr(>Chisq)
Tamaño       1011.8  3  < 2.2e-16 ***
Calidad      3290.5  3  < 2.2e-16 ***
Geometría     518.6  3  < 2.2e-16 ***
Proporción    505.5  3  < 2.2e-16 ***
Forma         765.5  3  < 2.2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tabla 6: Anova regresión Logística Multinomial.

Se puede ver que todos los parámetros son significativos a un nivel de confianza del 0%. Las variables más importantes en el modelo se corresponden con aquellas que tienen un mayor aumento en la verosimilitud, siendo estas Tamaño y Calidad. A continuación, se obtiene el ODDS-ratio de los efectos del modelo para poder interpretarlo:

	Tamaño	Calidad	Geometría	Proporción	Forma
X2	7.179641	3.159624	0.6340489	0.4921650	0.3364156
X3	11.157150	50.878533	0.4254275	0.3273194	0.3120059
X4	18.992507	47.052322	1.2226444	0.2043584	0.0946153

Tabla 7: ODDS-ratio regresión logística multinomial.

- Tamaño: En todos los casos, el odds ratio es superior a 1, lo que indica que un incremento en el tamaño del diamante aumenta la probabilidad de que su precio se eleve. En particular, el odds ratio es 7.18 para los diamantes de precio medio-bajo respecto a los de precio bajo, 11.16 para los de precio medio-alto y 18.99 para los de precio alto, en comparación con los de precio bajo.

- Calidad: Al igual que con la variable Tamaño, una mejora en la calidad del diamante incrementa significativamente la probabilidad de un mayor precio. El odds ratio es 3.15 para los diamantes de precio medio-bajo, 50.88 para los de precio medio-alto, y 47.05 para los de precio alto, todos en relación con los de precio bajo.

- Geometría: En este caso, los odds ratios son menores a 1, lo que implica que una mejora en la geometría del diamante disminuye la probabilidad de que su precio aumente. Para los diamantes de precio medio-bajo, esta reducción es del 36.3%, para los de precio medio-alto del 57.5%, mientras que en los de precio alto, una mejor geometría se asocia con un aumento del 22.3% en la probabilidad de obtener un precio mayor.

- Proporción: Se observa un odds ratio inferior a 1 en todos los niveles de precio, lo que sugiere que una mejor proporción del diamante reduce la probabilidad de un mayor precio. La disminución es del 51.8% para los diamantes de precio medio-bajo, del 67.3% para los de precio medio-alto y del 79.3% para los de precio alto.

- Forma: De manera consistente con las variables anteriores cuyo odds ratio es menor a 1, una mejora en la forma del diamante también reduce la probabilidad de un aumento en su precio. En términos porcentuales, esto implica una reducción del 66.4% en los diamantes de precio medio-bajo, del 68.8% en los de precio medio-alto y del 90.5% en los de precio alto.

Por último, se evalúa este modelo a través de una tabla de contingencia:

	X1	X2	X3	X4
X1	1009	197	1	2
X2	257	565	59	62
X3	10	387	923	340
X4	30	146	316	886

	X1	X2	X3	X4
X1	241	60	0	0
X2	71	140	14	15
X3	7	89	231	72
X4	7	34	79	235

Tabla 8: Tablas de contingencia regresión logística multinomial. Entrenamiento y prueba.

La tabla de la izquierda corresponde con la tabla de contingencia de la base de datos de entrenamiento mientras que la tabla de la derecha corresponde a la base de datos de prueba. La tasa de acierto indica la proporción de observaciones correctamente clasificadas por el modelo respecto al total de observaciones. Se calcula para ambas bases:

$$Tasa\ de\ Acierto\ Entrenamiento = \frac{3383}{5190} = 0.652 \qquad Tasa\ de\ Acierto\ Prueba = \frac{847}{1295} = 0.654$$

No suele ser lo habitual, pero en este caso, se consiguen mejores resultados en la base de prueba, donde se aciertan un 65.4% de las veces al clasificar siguiendo el modelo creado.

El Índice de Kappa de Cohen es una métrica que ajusta el Accuracy para tener en cuenta el azar, es decir, la probabilidad de que el modelo haya acertado por casualidad. Se realiza evaluando si las predicciones son mejores que lo que cabría esperar por azar. Para la base de prueba, conseguimos un valor de 0.539 que implica que el modelo tiene una concordancia decente, teniendo un desempeño mejor que el azar, aunque mejorable.

Posterior a esto, se saca por pantalla la sensibilidad y la especificidad del modelo. Evalúan la capacidad del modelo de detectar los eventos y no eventos, respectivamente (Alonso Revenga & Calviño Martínez, 2025). Mientras que la sensibilidad indica cuántos casos positivos reales han sido correctamente clasificados como positivos por el modelo (pertenecer a una clase y ser clasificado en esa clase), la especificidad indica cuántos casos negativos reales han sido correctamente clasificados como negativos por el modelo (no pertenecer a una clase y ser clasificado como que no pertenece a esa clase).

	Sensibilidad	Especificidad
X1	0.739	0.938
X2	0.433	0.897
X3	0.713	0.827
X4	0.730	0.877

Tabla 9: Sensibilidad y especificidad regresión logística multinomial.

Se puede apreciar que prácticamente en todas las categorías se consiguen buenos valores menos es la sensibilidad de la clase X2, correspondiente a precio medio-bajo. Esto implica que los pertenecientes a la clase X2 están teniendo dificultades de ser clasificados bien. Pese a esto, no se trata de un mal modelo.

Por último, para finalizar la regresión logística multinomial con factores retenidos, se grafica la curva ROC. Se representa la sensibilidad frente a 1 - especificidad para todas las posibles clasificaciones que se pueden derivar de las probabilidades predichas, por lo que permite obtener una medida de evaluación del modelo sin necesidad de clasificar cada observación (Alonso Revenga & Calviño Martínez, 2025). El área bajo la curva es de 0.876, lo cual implica que se trata de un modelo de calidad buena. Observando el gráfico, las categorías mejor modelizadas son X1 (precio bajo) y X4 (precio alto), mientras que X3 (precio medio-alto) y X2 (precio medio-bajo) son sutilmente peores. Esto se debe también a que ambas categorías no presentan una diferencia considerable entre sí, a diferencia de las categorías en los extremos.

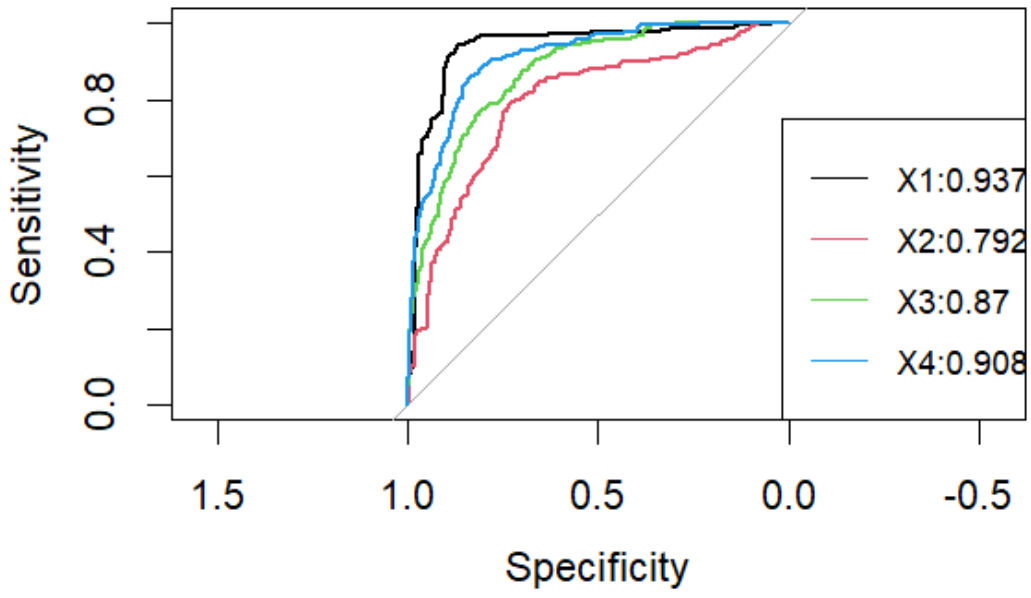


Gráfico 11: Curva ROC regresión logística multinomial.

2.3. Árbol de regresión con los valores originales de la variable precio

Para poder comprobar los resultados que se obtengan con el árbol de regresión, se realiza una partición de los datos, donde el 80% de estos servirán para entrenar el modelo (data_train) y el 20% restantes servirán para comprobar la eficacia de este (data_test).

Los árboles de regresión constituyen una herramienta útil para la predicción de variables cuantitativas de una manera sencilla y sin la necesidad de que los datos cumplan hipótesis teóricas como la normalidad o las relaciones entre ellos sean de tipo lineal (Calviño Martínez, 2024).

Se construye un árbol con la base de datos de entrenamiento consiguiendo lo siguiente:

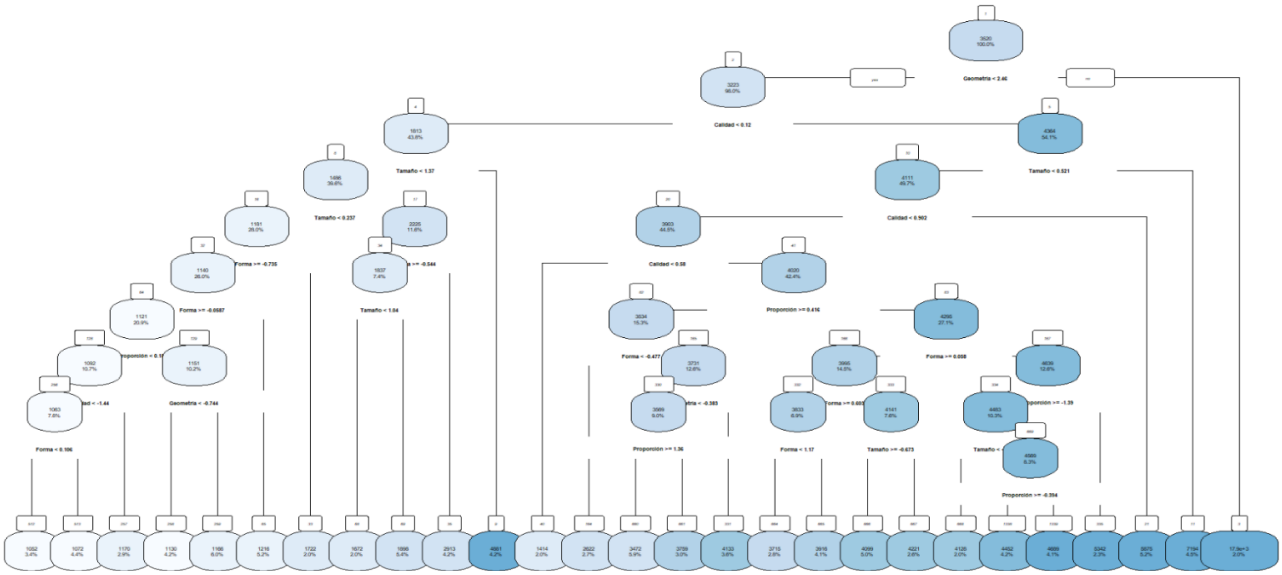


Gráfico 12: Árbol de regresión.

El árbol tiene una profundidad de 10 niveles y un total de 27 hojas. Por sí solo, resulta ambiguo y difícil de interpretar, por lo que no será el modelo final elegido. Para construir un modelo más adecuado, será necesario comprobar su fiabilidad y aplicar ciertos criterios que permitan simplificar su estructura y mejorar su interpretabilidad. Para evaluar la calidad del modelo, se analiza su R^2, que en este caso es 0.6890 en la base de entrenamiento y 0.6601 en la base de prueba. El R^2 mide qué tan bien un modelo de regresión explica la variabilidad de la variable dependiente. En este caso, los resultados son adecuados.

La importancia de las variables enseña qué variables son las que más aportan a la predicción del modelo. Como se puede apreciar, Geometría será la variable más predictora, seguida por Calidad y Tamaño mientras que Forma y Proporción son las variables menos predictoras.

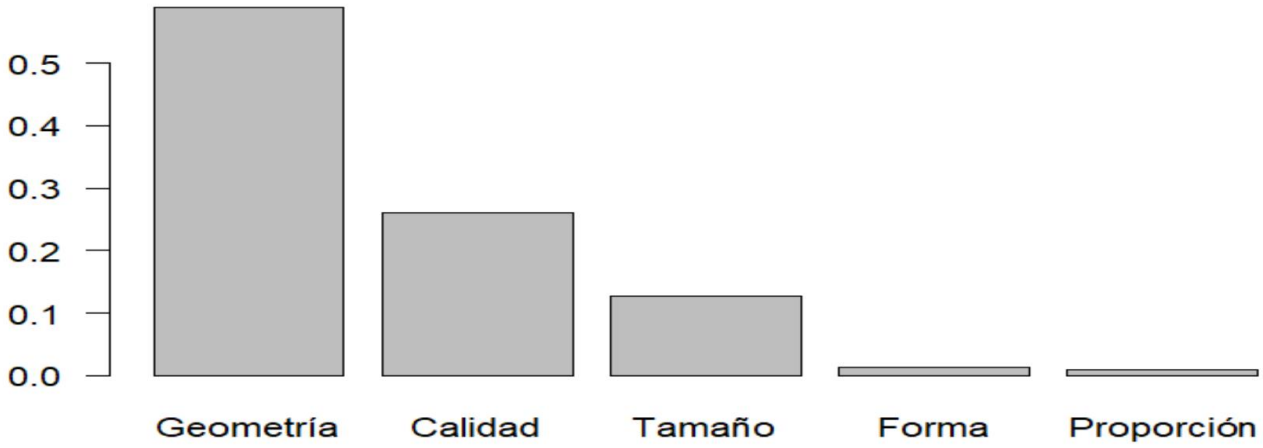


Gráfico 13: Importancia de las variables.

Con este modelo, se procede a la poda. Se entiende por poda el proceso mediante el cual se estudia la posibilidad de eliminar algunas hojas del árbol construido con el objetivo de reducir el sobreajuste (Alonso Revenga & Calviño Martínez, 2025). Con el siguiente gráfico, se encuentra el punto óptimo para podar el árbol, donde minimizamos el error que se comete sin perder información relevante del modelo. En este caso, se elige un valor de $cp = 0.01$.

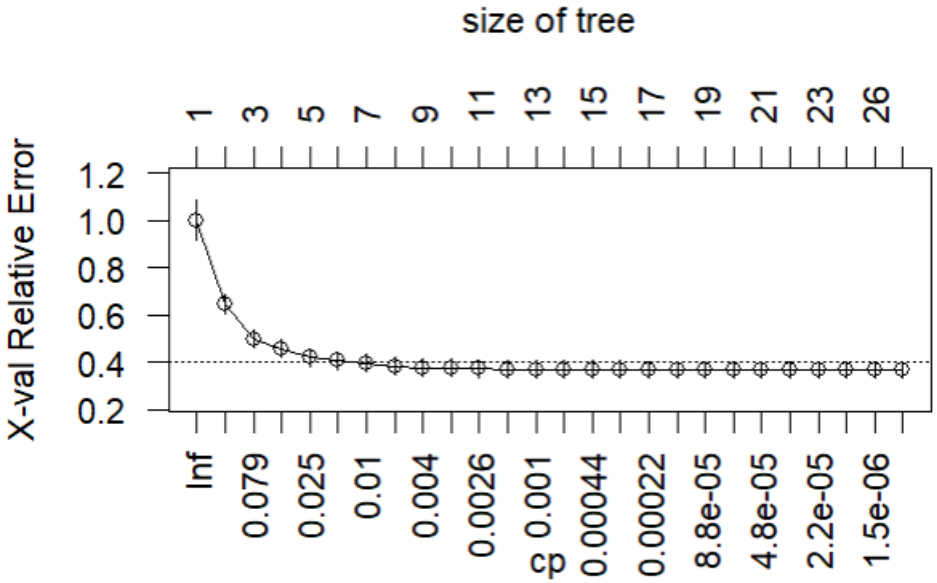


Gráfico 14: Corte óptimo del árbol de regresión.

Como resultado se consigue el siguiente árbol, que cuenta con una profundidad de 5 niveles y un total de 8 hojas. Con esto, se ha logrado reducir enormemente el tamaño del árbol y mejorar su interpretabilidad.

Asimismo, el árbol de regresión ofrece una representación visual de los nodos finales, en la cual la intensidad del color azul refleja la magnitud del precio de los diamantes: los tonos más oscuros indican valores más elevados, mientras que los más claros se asocian con precios inferiores.

Una posible manera de interpretarlo sería, por ejemplo: los diamantes que cuentan con una Geometría menor de 2.46, una Calidad menor de 0.12 y un Tamaño mayor de 1.37 tendrán un precio aproximado de 4881$, constituyendo un 4.2% de la muestra.

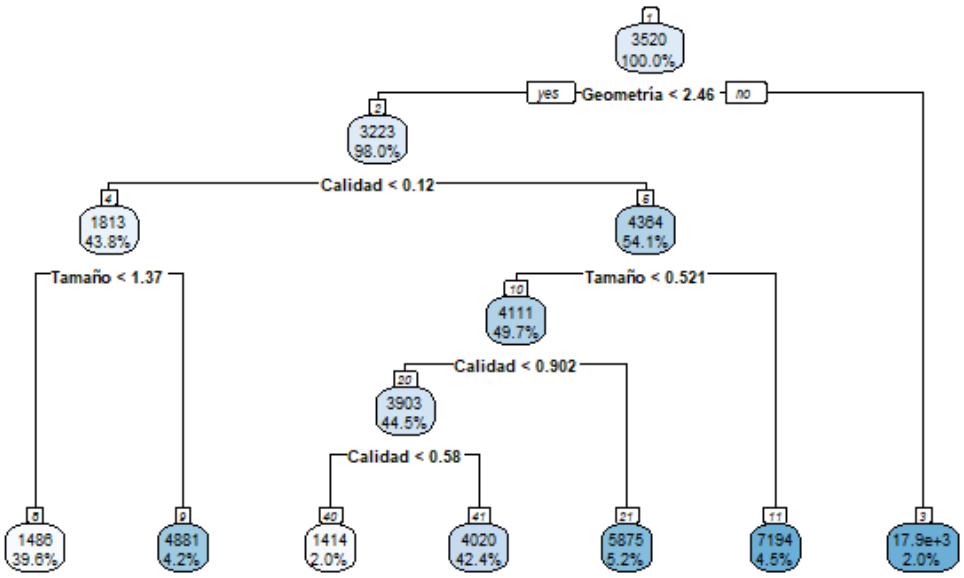


Gráfico 15: Árbol de regresión.

Por último, es necesario verificar el valor del R^2, ya que este indicador nos permite evaluar la viabilidad del modelo. Se obtuvo un valor de 0.6631 en la base de entrenamiento y de 0.6382 en la base de prueba. A pesar de la drástica reducción en el tamaño del árbol, la pérdida de variabilidad explicada de la variable dependiente ha sido mínima. Estos resultados indican que el modelo resultante presenta un buen equilibrio entre simplicidad e interpretabilidad, sin comprometer significativamente su capacidad predictiva.

2.4. Árbol de clasificación con las clases creadas

Para evaluar los resultados obtenidos con el árbol de clasificación, se realiza una partición de los datos, asignando el 80% para entrenar el modelo (data_train) y el 20% restante para validar su eficacia (data_test).

Un árbol de clasificación es un modelo de aprendizaje supervisado utilizado para predecir categorías o clases dentro de un conjunto de datos. Mientras que un árbol de regresión predice una variable continua, un árbol de clasificación se emplea para predecir una variable categórica (es decir, asignar una clase a cada observación.

Construimos un árbol de clasificación con múltiples niveles para más tarde proceder a la poda, aunque este sea difícil de interpretar. El árbol obtenido es el siguiente:

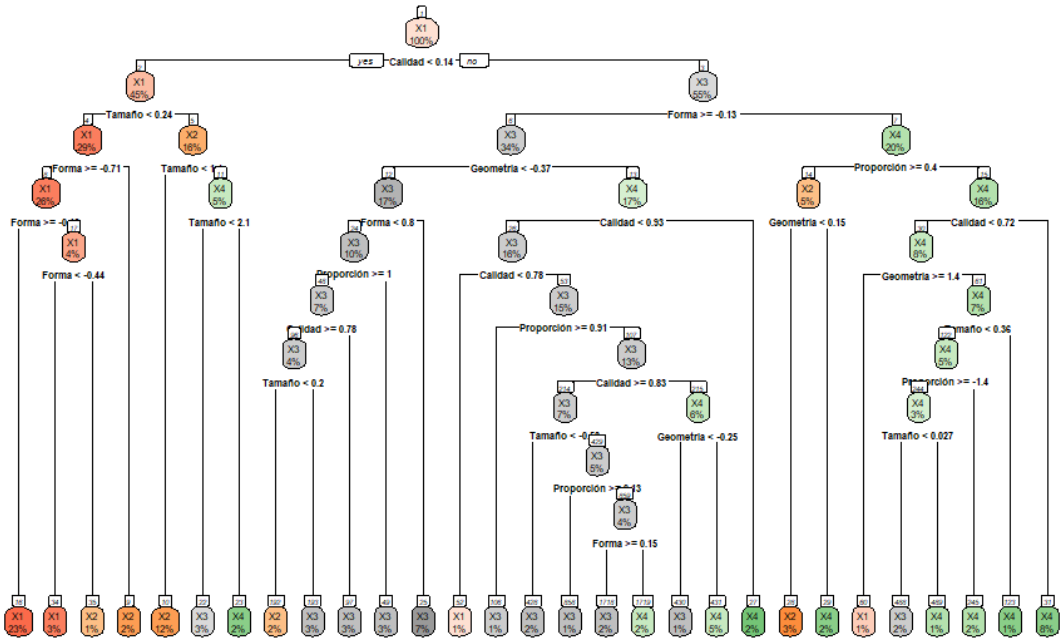


Gráfico 16: Árbol de clasificación.

Este árbol, con una profundidad de 10 niveles y 29 hojas, presenta una estructura relativamente compleja.

Las variables más importantes en este modelo, Calidad y Tamaño indican que estas características son fundamentales para predecir la clase de los diamantes, probablemente porque tienen una relación más directa con las categorías de precio. La alta importancia de estas variables sugiere que el modelo utiliza eficazmente la información contenida en ellas para tomar decisiones.

En cambio, las variables Forma, Geometría y Proporción, que tienen una menor relevancia en el árbol de regresión, contribuyen de manera más significativa en este modelo que en los modelos anteriores, aportando información adicional que mejora su capacidad predictiva.

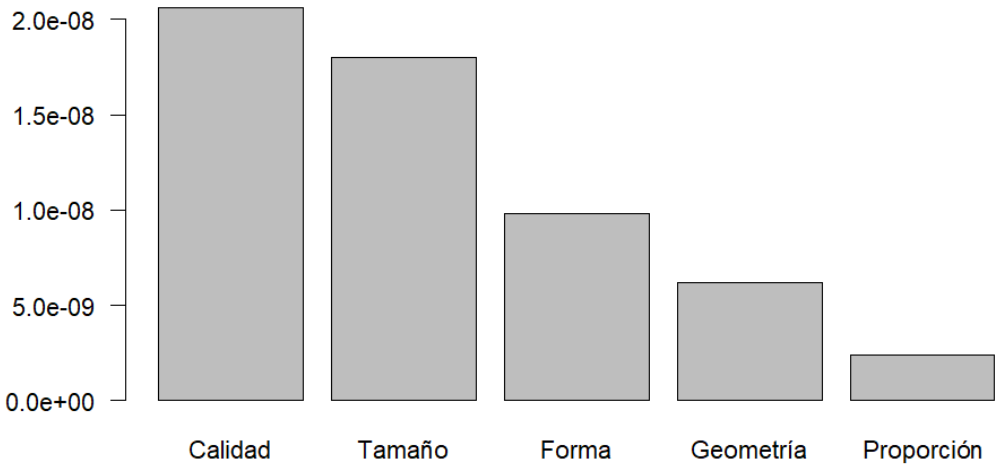


Gráfico 17: Importancia de las variables.

Para determinar el tamaño óptimo del árbol, se emplea la técnica de poda con el fin de reducir el sobreajuste, asegurando que no se pierda información relevante. Se elige un valor de $cp = 0.006$

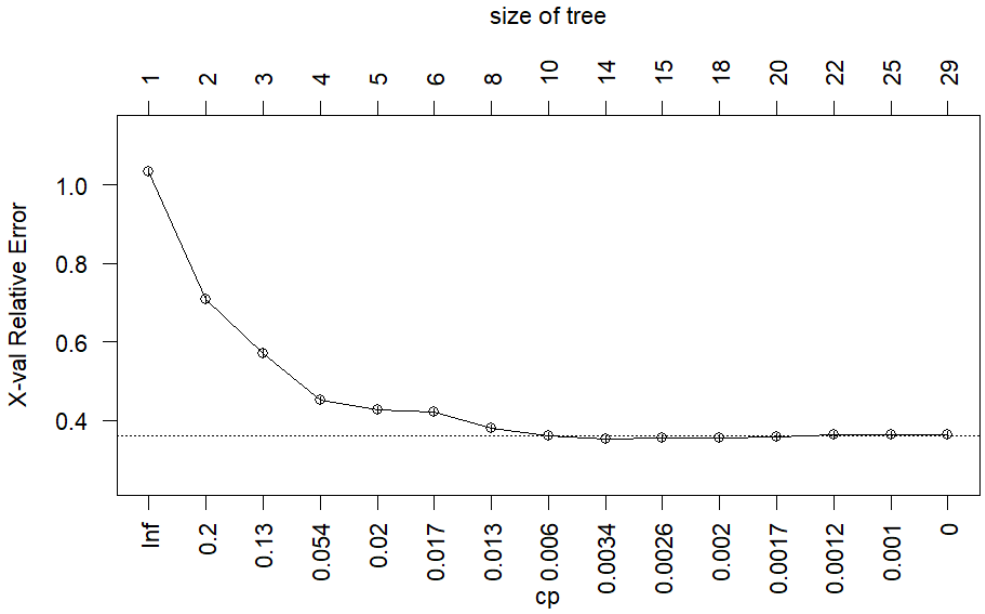


Gráfico 18: Corte óptimo del árbol de clasificación.

El árbol resultante presenta una estructura de 4 niveles y 10 hojas, lo que mejora su interpretabilidad. Dado que se trata de un árbol de clasificación, no es posible calcular su R^2; sin embargo, es factible obtener la matriz de confusión, la tasa de acierto, el índice de Kappa, así como las métricas de sensibilidad y especificidad.

Este árbol presenta una mayor variedad de colores, cada uno de los cuales indica una clase específica. El color rojo corresponde a la clase 1, asociada al precio bajo; el naranja a la clase 2, que representa precio medio-bajo; el gris a la clase 3, vinculada al precio medio-alto; y, finalmente, el verde a la clase 4, indicando precio alto.

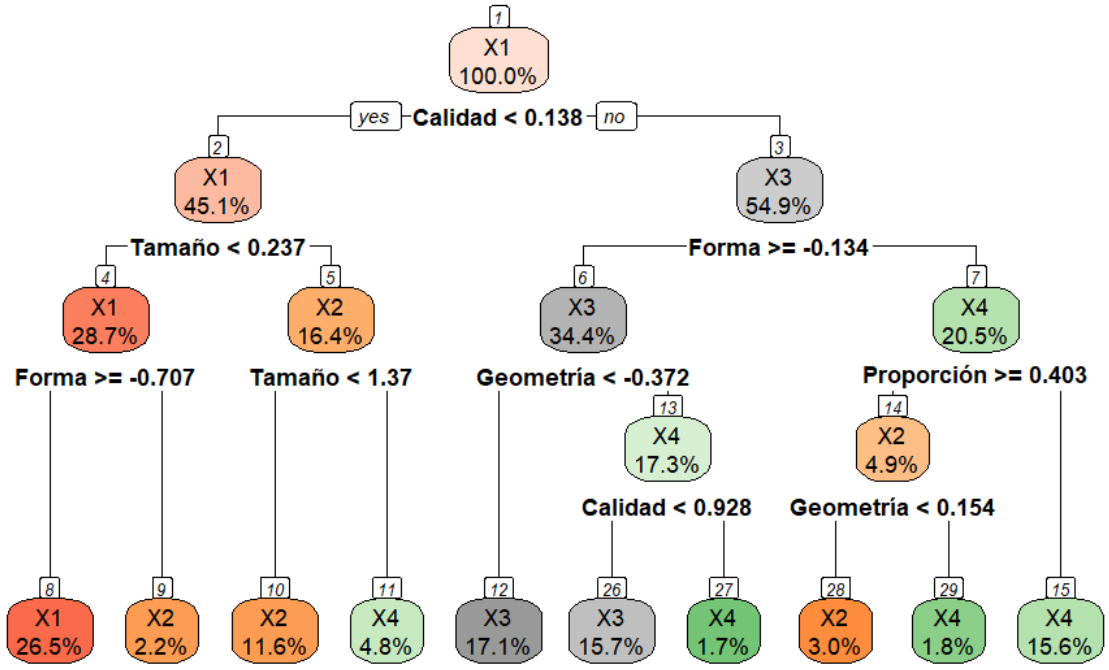


Gráfico 19: Árbol de clasificación.

Una interpretación posible sería la siguiente: los diamantes con una calidad igual o superior a 0.138, una forma inferior a -0.134 y una proporción menor a 0.403 representan el 15.6% de la muestra, correspondiendo a diamantes de precio alto.
Finalmente, es necesario evaluar las tasas de acierto y los indicadores de calidad del modelo.

	X1	X2	X3	X4
X1	1201	169	1	2
X2	50	741	51	32
X3	24	275	1013	388
X4	31	110	234	868

	X1	X2	X3	X4
X1	293	58	0	0
X2	16	188	11	8
X3	10	51	251	106
X4	7	26	62	208

Tabla 10: Tablas de contingencia árbol de clasificación. Entrenamiento y prueba.

La tasa de acierto obtenida es de 0.737, lo que se considera un valor alto, dado que los aciertos al azar en un escenario con cuatro clases serían de 0.25. El Índice Kappa es de 0.649, lo cual se considera un valor bueno, para la base de entrenamiento. En la base de prueba, la tasa de acierto es de 0.726 y el Índice Kappa es de 0.634. Dado que ambos valores no difieren significativamente de los obtenidos en la base de entrenamiento, podemos concluir que no hay evidencia de sobreajuste.

Para la sensibilidad y especificidad obtenemos en ambas bases:

	Sensibilidad	Especificidad
Base de entrenamiento	0,92	0,572
Base de prueba	0,899	0,582

Tabla 11. Sensibilidad y especificidad árbol de clasificación.

En ambos casos se consiguen buenos valores para la sensibilidad (captamos bien a los que pertenecen a cada grupo) pero no se obtienen buenos valores de especificidad (no asignamos bien a los que no pertenecen a los grupos).

3. Estudio sin el Análisis factorial

Desde este punto en adelante, se procederá a realizar nuevamente los mismos análisis, con la diferencia de que, en esta ocasión, no se retendrán factores. Esta modificación nos permitirá trabajar con un conjunto más amplio de variables, lo que podría contribuir a obtener resultados más robustos y representativos, al incluir una mayor diversidad de información en el modelo. Sin embargo, al aumentar el número de variables, también se incrementa la complejidad del modelo, lo que podría llevar a un mayor riesgo de sobreajuste. Este fenómeno ocurre cuando el modelo se ajusta demasiado a los datos de entrenamiento, capturando patrones específicos y ruidos que no se generalizan bien a nuevos datos. Por lo tanto, es esencial monitorear de cerca el desempeño del modelo en datos de prueba y aplicar técnicas como la validación cruzada o la regularización para mitigar este riesgo.

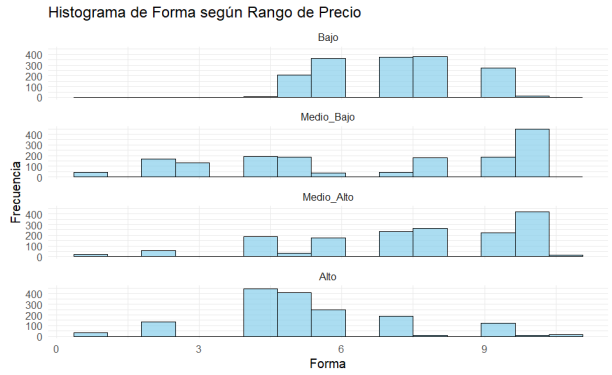
3.1. Análisis discriminante con las clases creadas

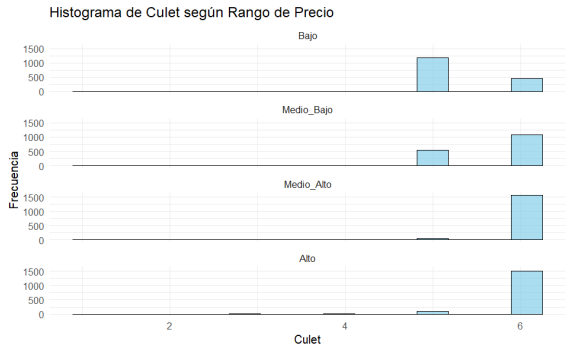
Para llevar a cabo este análisis, recodificamos nuestra variable en 4 niveles, teniendo en cuenta los cuartiles. Por tanto, por debajo de 1210 se considerará clase X1 (la que cuenta con el menor precio), entre este valor y 3320 se considerará precio medio-bajo siendo X2, entre 3320 y 4390 se considerará precio medio-alto y mayor que 4390 se considerará precio alto.

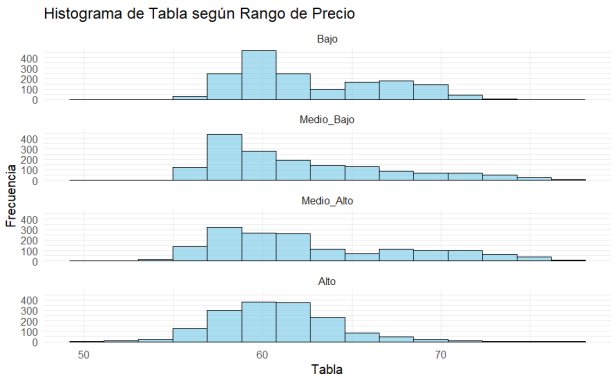
Como se hizo anteriormente, lo primero que se debe hacer para comprobar que es viable hacer un análisis discriminante con los datos es estudiar su normalidad. No es necesario que se asemeje de manera precisa a una normal, siempre y cuando, la forma del histograma de cada variable recuerde al histograma clásico de una normal.

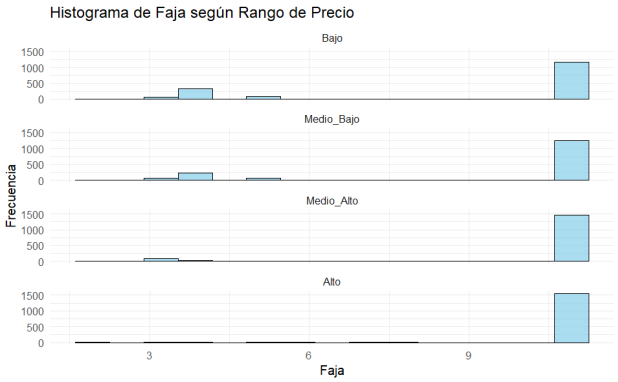
$$H0: Los\ datos\ provienen\ de\ una\ distribución\ Normal$$
$$H1: Los\ datos\ no\ provienen\ de\ una\ distribución\ Normal$$

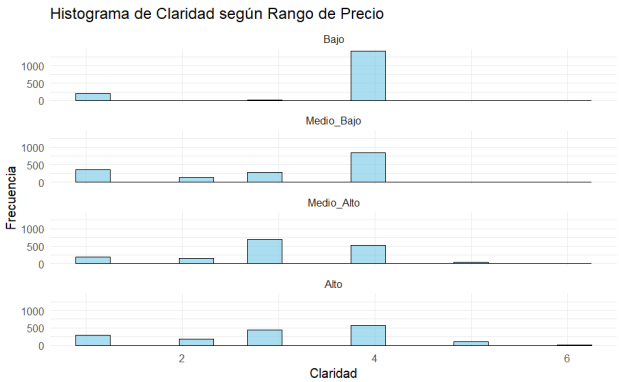
Se procede a continuación a graficar todas las variables en cada una de las categorías de la variable objetivo, siendo esta, precio del diamante.

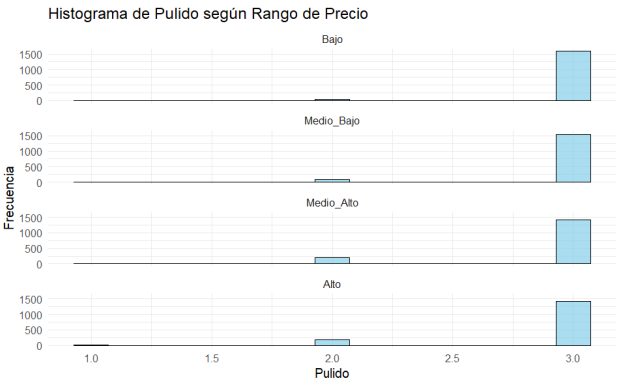

Histograma de Forma según Rango de Precio
Bajo
Medio_Bajo
Medio_Alto
Alto
Frecuencia
Forma

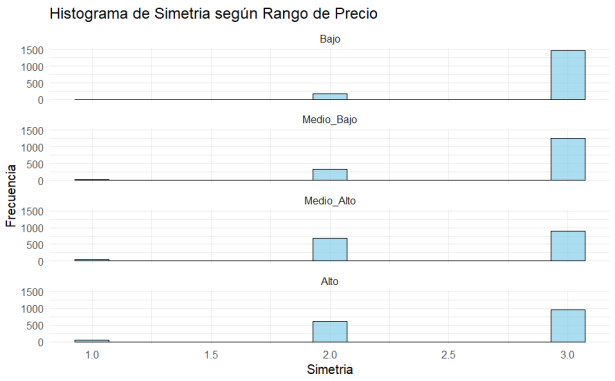

Histograma de Culet según Rango de Precio
Bajo
Medio_Bajo
Medio_Alto
Alto
Frecuencia
Culet

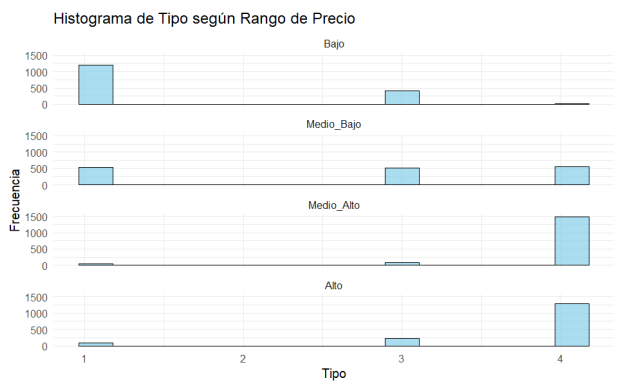

Histograma de Tabla según Rango de Precio
Bajo
Medio_Bajo
Medio_Alto
Alto
Frecuencia
Tabla


Histograma de Faja según Rango de Precio
Bajo
Medio_Bajo
Medio_Alto
Alto
Frecuencia
Faja


Histograma de Claridad según Rango de Precio
Bajo
Medio_Bajo
Medio_Alto
Alto
Frecuencia
Claridad


Histograma de Pulido según Rango de Precio
Bajo
Medio_Bajo
Medio_Alto
Alto
Frecuencia
Pulido


Histograma de Simetria según Rango de Precio
Bajo
Medio_Bajo
Medio_Alto
Alto
Frecuencia
Simetria


Histograma de Tipo según Rango de Precio
Bajo
Medio_Bajo
Medio_Alto
Alto
Frecuencia
Tipo

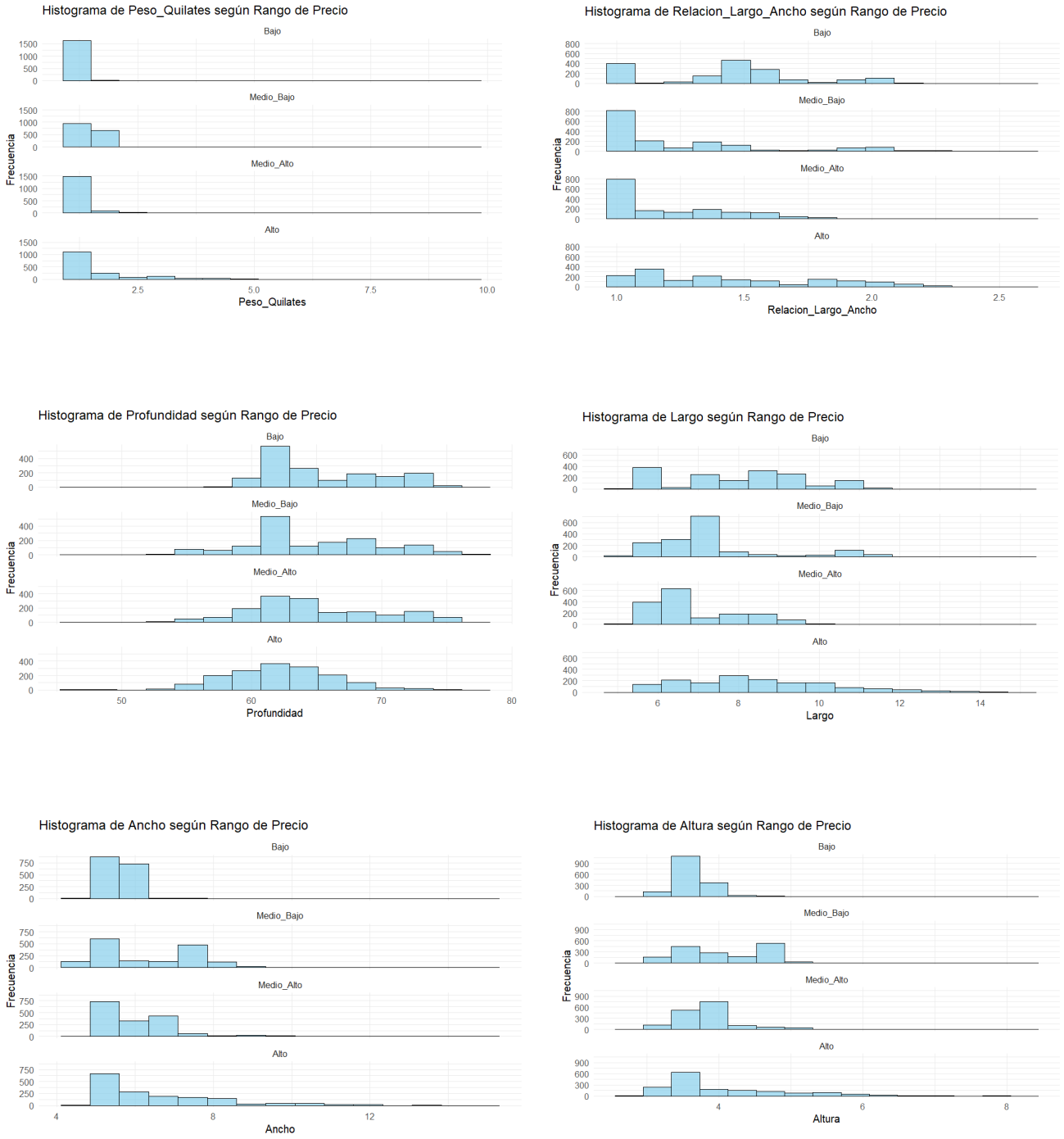


Gráfico 20: Histogramas de las variables independientes.

Al contrario de lo que ocurría en el análisis discriminante con los factores retenidos, vamos a tener que realizar transformaciones en prácticamente todas las variables para normalizarlas. Las variables que vamos a normalizar siguiendo el método del Cuantil Ordenado son: Forma, Tabla, Peso_Quilates, Relación_Largo_Ancho, Profundidad, Largo, Ancho y Altura. El resto de las variables se recodificarán usando el arcoseno hiperbólico. A continuación, con las variables ya normalizadas, se lleva a cabo un test de homogeneidad de la varianza. Las hipótesis son:

$$H0: Las\ varianzas\ no\ son\ significativamente\ diferentes$$
$$H1: Las\ varianzas\ son\ significativamente\ diferentes$$

El test de Levenne nos proporciona un p_valor de prácticamente 0 en todas las categorías rechazando la hipótesis nula. Con este resultado, se procede a hacer un análisis discriminante cuadrático, obteniendo las siguientes ecuaciones:

$$X1 = 2.363 * \text{Culet} + 0.141 * \text{Forma} + 0.178 * \text{Tabla} + 8.978 * \text{Faja} + 1.932 * \text{Claridad} + 1.808 * \text{Pulido} + 1.777 * \text{Simetría} + 1.142 * \text{Tipo} + 0.109 * \text{Peso\_Quilates} + 0.323 * \text{Relación\_Largo\_Ancho} + 0.246 * \text{Profundidad} + 0.111 * \text{Largo} - 0.333 * \text{Ancho} - 0.356 * \text{Altura} + 0.2522644$$

$$X2 = 2.431 * \text{Culet} + 0.010 * \text{Forma} - 0.027 * \text{Tabla} + 9.401 * \text{Faja} + 1.732 * \text{Claridad} + 1.798 * \text{Pulido} + 1.727 * \text{Simetría} + 1.596 * \text{Tipo} + 0.291 * \text{Peso\_Quilates} - 0.342 * \text{Relación\_Largo\_Ancho} + 0.050 * \text{Profundidad} - 0.170 * \text{Largo} + 0.043 * \text{Ancho} + 0.213 * \text{Altura} + 0.249$$

$$X3 = 2.485 * \text{Culet} + 0.300 * \text{Forma} + 0.128 * \text{Tabla} + 10.222 * \text{Faja} + 1.763 * \text{Claridad} + 1.768 * \text{Pulido} + 1.637 * \text{Simetría} + 2.043 * \text{Tipo} - 0.491 * \text{Peso\_Quilates} - 0.327 * \text{Relación\_Largo\_Ancho} + 0.167 * \text{Profundidad} - 0.447 * \text{Largo} - 0.031 * \text{Ancho} + 0.047 * \textit{Altura} + 0.250$$

$$X4 = 2.472 * \text{Culet} - 0.470 * \text{Forma} - 0.201 * \text{Tabla} + 10.729 * \text{Faja} + 1.727 * \text{Claridad} + 1.764 * \text{Pulido} + 1.642 * \text{Simetría} + 1.986 * \text{Tipo} + 0.186 * \text{Peso\_Quilates} + 0.383 * \text{Relación\_Largo\_Ancho} - 0.386 * \text{Profundidad} + 0.478 * \text{Largo} + 0.272 * \text{Ancho} + 0.090 * \textit{Altura} + 0.249$$

A continuación, nos fijamos en la matriz de confusión, donde en verde se señalan los aciertos y en rojo los fallos:

	X1	X2	X3	X4
X1	293	34	0	0
X2	8	198	18	16
X3	11	72	246	99
X4	11	23	60	207

Tabla 12: Tabla de contingencia discriminante.

Para comprender mejor esto, utilizamos el Hit Ratio. Esta medida pondera los casos acertados entre el número total de observaciones.

$$Hit\ Ratio = \frac{Número\ de\ aciertos}{Número\ de\ observaciones}$$

$$Hit\ Ratio = \frac{944}{1296} = 0.728$$

Acertamos en un 72.8% utilizando las reglas discriminantes creadas. La regla creada sin la agrupación por factores resulta ligeramente peor que la anterior creada. Se consigue a continuación el Índice de Significación Práctica para cada grupo.

$$ISP = \frac{o - e}{n - e}$$

Siendo:
O: bien clasificados
E: bien clasificados al azar
N: tamaño muestral

$$ISP_{X1} = \frac{293-(0.25^2*327)}{327-(0.25^2*327)} = 0.889 \qquad ISP_{X2} = \frac{198-(0.25^2*240)}{240-(0.25^2*240)} = 0.813$$

$$ISP_{X3} = \frac{247-(0.25^2*428)}{428-(0.25^2*428)} = 0.553 \qquad ISP_{X4} = \frac{207-(0.25^2*301)}{301-(0.25^2*301)} = 0.667$$

$$ISP = \frac{(293+198+246+207)-(0.25^2*1296)}{1296-(0.25^2*1296)} = 0.710$$

Se observa que todos los valores del Índice de Significación Práctica superan el umbral de 0.25, el cual representa el nivel de rendimiento esperado bajo condiciones de aleatoriedad, es decir, el número de aciertos atribuibles al azar. En particular, el valor de este índice para el primer grupo muestra un leve incremento en comparación con el obtenido mediante el análisis discriminante con retención de factores. En contraste, para los grupos restantes, el Índice de Significación Práctica presenta una disminución. Por su parte, el valor global del índice correspondiente al modelo general también resulta ligeramente inferior al observado previamente, aunque continúa evidenciando un desempeño satisfactorio.

Resulta también interesante en este tipo de análisis calcular la sensibilidad y especificidad de cada clase, ya que esto indica de manera más precisa la eficiencia del modelo en detectar casos reales y la eficiencia en evitar falsos positivos.

$$Sensibilidad = \frac{Verdaderos\ positivos}{Total\ de\ positivos\ reales}$$

$$Especificidad = \frac{Verdaderos\ negativos}{Total\ de\ negativos\ reales}$$

Agrupando los resultados de la tabla de contingencia anterior, se consigue:

	Sensibilidad	Especificidad
X1	0.896	0.970
X2	0.825	0.878
X3	0.574	0.912
X4	0.688	0.886

Tabla 13: Sensibilidad y especificidad discriminante.

En términos generales, el modelo discriminante muestra una sensibilidad y especificidad buena, especialmente en la clasificación de la clase X1, donde alcanza una sensibilidad de 0.896 y una especificidad de 0.970, lo que refleja una elevada capacidad para identificar correctamente los casos pertenecientes a esta categoría y para evitar asignaciones incorrectas desde otras clases. La clase X2 también presenta un desempeño adecuado, con valores de sensibilidad (0.825) y especificidad (0.878) dentro de un rango considerado bueno.

Por el contrario, las clases X3 y X4 muestran un comportamiento menos favorable, con sensibilidades de 0.574 y 0.688, respectivamente, lo que indica una mayor proporción de errores en la identificación de los casos reales de estas clases. A pesar de que ambas mantienen especificidades aceptables (0.912 para X3 y 0.886 para X4), los bajos niveles de sensibilidad podrían limitar la utilidad del modelo en contextos donde resulte especialmente importante una correcta detección de dichas categorías.

El modelo discriminante presenta un rendimiento general satisfactorio, con una buena capacidad para clasificar correctamente a los individuos, especialmente en las clases X1 y X2. En estos grupos, tanto los índices de sensibilidad y especificidad como el índice de significación práctica fueron altos, lo que indica una adecuada utilidad de las variables discriminantes. En cambio, el modelo mostró dificultades para identificar correctamente los casos de las clases X3 y X4, con una sensibilidad reducida y un índice de significación práctica limitado. Aunque la especificidad en estas categorías se mantuvo dentro de rangos aceptables, los resultados sugieren que la capacidad del modelo no es homogénea entre todas las clases.

3.2. Regresión Logística Multinomial con las clases creadas

De manera similar a lo realizado previamente, se lleva a cabo un análisis de regresión logística multinomial utilizando las clases previamente definidas, correspondientes a los niveles de precio bajo, medio-bajo, medio-alto y alto. Se lleva a cabo una partición de los datos, dividiéndolos en un conjunto de entrenamiento y un conjunto de prueba.

El proceso comienza con la creación de un modelo que incluye todas las variables, verificando que el proceso de optimización haya convergido, lo cual indica que el modelo obtenido es adecuado. Posteriormente, se realiza un análisis de tipo II para determinar qué variables son estadísticamente significativas, utilizando la función ANOVA:

```
Analysis of Deviance Table (Type II tests)

Response: Precio
                    LR Chisq Df Pr(>Chisq)
Culet                1426.59  3  < 2.2e-16 ***
Forma                 100.04  3  < 2.2e-16 ***
Tabla                   1.19  3  0.7561878
Faja                   17.81  3  0.0004807 ***
Claridad              366.48  3  < 2.2e-16 ***
Pulido                  2.49  3  0.4773171
Simetria                3.15  3  0.3690906
Tipo                 3091.37  3  < 2.2e-16 ***
Peso_Quilates         446.62  3  < 2.2e-16 ***
Relacion_Largo_Ancho   84.94  3  < 2.2e-16 ***
Profundidad           152.32  3  < 2.2e-16 ***
Largo                  34.77  3  1.361e-07 ***
Ancho                 127.43  3  < 2.2e-16 ***
Altura                141.72  3  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
```

Tabla 14: Anova regresión logística multinomial.

Se puede observar que algunas variables no son estadísticamente significativas; por lo tanto, para mejorar la eficiencia del modelo, se procederá a excluirlas en los siguientes pasos. Al realizar este procedimiento, el proceso no converge, lo cual puede deberse a problemas de multicolinealidad, diferencias significativas en la escala de las variables u otras causas potenciales. Por ello, con el objetivo de seleccionar de manera más adecuada las variables que conformarán el modelo, se pueden aplicar métodos automáticos de selección de variables. Para entender como se llevan a cabo estos métodos automáticos de selección se debe saber:

- La función step (en R) permite seleccionar automáticamente las variables más relevantes para un modelo, evaluando su inclusión o exclusión en función del ajuste global. Este proceso puede realizarse en tres direcciones: hacia adelante, comenzando con un modelo sin variables e incorporándolas progresivamente; hacia atrás, partiendo del modelo completo y eliminando variables no significativas; o en ambas direcciones, combinando ambas estrategias para encontrar el modelo más adecuado.

- La selección de variables mediante step se guía por criterios de información como AIC y BIC, que buscan un equilibrio entre la calidad del ajuste del modelo y su simplicidad. El AIC penaliza la complejidad del modelo de forma más flexible, favoreciendo aquellos con mayor capacidad predictiva. En cambio, el BIC impone una penalización mayor, especialmente con muestras grandes, por lo que tiende a seleccionar modelos más simples y estables.

Tras aplicar las distintas combinaciones de criterios y direcciones en los métodos automáticos de selección de variables, se obtienen varios modelos candidatos. De ellos, dos presentan un total de 30 parámetros, mientras que el resto contienen 36. Es importante verificar si estos modelos son efectivamente distintos, ya que es frecuente que los procedimientos automáticos prioricen estructuras similares.

```
[[1]]
multinom(formula = Precio ~ Tipo + Peso_Quilates + Culet + Relacion_Largo_Ancho +
    Claridad + Profundidad + Forma + Largo + Altura, data = data_train7,
    trace = F)

[[2]]
multinom(formula = Precio ~ Tipo + Peso_Quilates + Culet + Relacion_Largo_Ancho +
    Claridad + Profundidad + Forma + Largo + Altura + Ancho +
    Faja, data = data_train7, trace = F)

[[3]]
multinom(formula = Precio ~ Tipo + Peso_Quilates + Culet + Relacion_Largo_Ancho +
    Claridad + Profundidad + Forma + Largo + Altura, data = data_train7,
    trace = F)

[[4]]
multinom(formula = Precio ~ Tipo + Peso_Quilates + Culet + Relacion_Largo_Ancho +
    Claridad + Profundidad + Forma + Largo + Altura + Ancho +
    Faja, data = data_train7, trace = F)

[[5]]
multinom(formula = Precio ~ Culet + Forma + Faja + Claridad +
    Tipo + Peso_Quilates + Relacion_Largo_Ancho + Profundidad +
    Largo + Ancho + Altura, data = data_train7, trace = F)

[[6]]
multinom(formula = Precio ~ Culet + Forma + Faja + Claridad +
    Tipo + Peso_Quilates + Relacion_Largo_Ancho + Profundidad +
    Largo + Ancho + Altura, data = data_train7, trace = F)

[1] 30 36 30 36 36 36
```

Tabla 15: Modelos creados.

Al realizar esta comprobación, se observa que, en realidad, únicamente se han generado dos modelos distintos. La diferencia fundamental entre ambos radica en el criterio de penalización utilizado: uno se construyó optimizando el AIC, mientras que el otro lo hizo con base en el BIC. A continuación, con

el objetivo de determinar cuál de los modelos obtenidos presenta un mejor rendimiento, se procede a evaluar de forma simultánea los indicadores de Accuracy, AUC e Índice Kappa. Esta comparación conjunta permite identificar con mayor claridad las diferencias entre modelos, facilitando una interpretación visual y cuantitativa de sus respectivas capacidades predictivas.

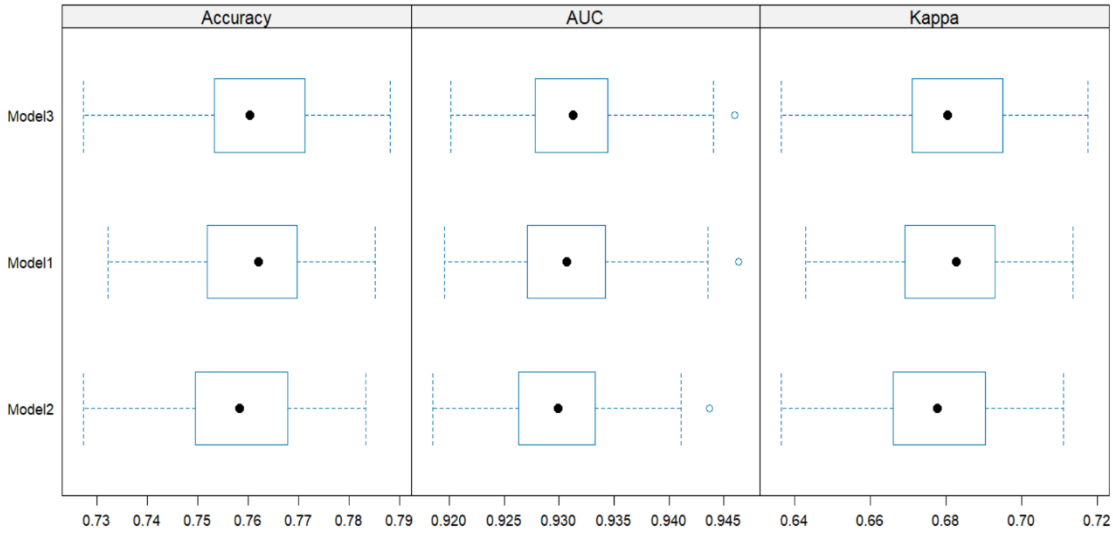


Gráfico 21: Evaluación de los modelos creados.

El Modelo 1 corresponde al modelo que creamos al inicio, el Modelo 2 es el modelo Stepwise BIC y el Modelo 3 es el modelo Stepwise AIC. Estos modelos cuentan con 45, 30 y 36 parámetros respectivamente.

Tras comparar los tres modelos mediante los indicadores de calidad Accuracy, AUC e Índice Kappa, se observa que todos presentan un rendimiento similar. No obstante, el Modelo 3 destaca ligeramente en los tres criterios evaluados. En particular, muestra una mediana de Accuracy levemente superior y una menor variabilidad, lo que sugiere una mayor estabilidad en las predicciones. Asimismo, en términos del AUC, mantiene un rendimiento alto y consistente, con escasa presencia de valores atípicos. Visualizando el último indicador, el Índice Kappa alcanza su valor más elevado en este modelo, lo que indica una mejor concordancia entre las predicciones del modelo y las clases reales.

Aunque el Modelo 3 muestra un rendimiento ligeramente superior en términos de Accuracy, AUC e Índice Kappa, su mayor número de parámetros podría generar algunos inconvenientes. Los modelos con más parámetros, como el Modelo 3, son más propensos al sobreajuste, lo que significa que podrían ajustarse demasiado a los datos de entrenamiento y no generalizar bien a nuevos datos.

Por otro lado, el Modelo 2, al tener menos parámetros, puede ofrecer una mejor generalización y ser menos sensible al ruido en los datos. Además, los modelos más simples suelen ser más eficientes en términos computacionales, lo que reduce los costos en tiempo de entrenamiento e inferencia.

Finalmente, el Modelo 2 podría ser una opción más estable, generalizable y eficiente, especialmente si la diferencia de rendimiento no es significativa.

Para comprender mejor la información que aporta este modelo, se saca por pantalla el ODDS-ratio de los efectos del modelo.

	Tipo	Peso_Quilates	Culet	Relacion_Largo_Ancho	Claridad	Profundidad	Forma
X2	1655.265	47710798	5.231444e-07	12474.109	0.3173182	0.7488902	0.8508913
X3	2382100.760	1910848649	3.833542e-13	1321.251	0.2793642	0.6879014	0.9910085
X4	5337836.608	11091645309	2.126919e-13	104564.098	0.2590356	0.6548774	0.7784210

	Largo	Altura
X2	0.1419913	1.005518
X3	0.2955532	2.846043
X4	0.2329950	4.619067

Tabla 16: ODDS-ratio regresión logística multinomial.

- Tipo: Los coeficientes de esta variable también son positivos, lo cual indica que una mejora en la categoría o tipo del diamante incrementa la probabilidad de obtener un precio más elevado. En concreto, el coeficiente es 2.81 para los diamantes de precio medio-bajo, 4.42 para los de precio medio-alto y 2.20 para los de precio alto, en comparación con los de precio bajo. Aunque el impacto es menor que en otras variables como peso o altura, sigue siendo un factor relevante en la determinación del precio, especialmente en los rangos medio-bajo y medio-alto.

- Peso en Quilates: El odds ratio aumenta drásticamente con el nivel de precio. Para los diamantes de precio medio-bajo (X2) es de aproximadamente 4.77 millones, y se incrementa hasta más de 1.1 mil millones para los de precio alto (X4). Esto indica que el peso del diamante tiene una influencia sumamente significativa en el aumento del precio: a mayor peso, mucho mayor probabilidad de que el diamante pertenezca a una categoría de precio superior.

- Culet: Aunque los valores son muy pequeños, todos los coeficientes del culet son positivos. Esto sugiere que una mejora en esta característica también incrementa la probabilidad de un mayor precio. El efecto es leve, pero consistente: 5.23×10^{-7} para precio medio-bajo, 3.83×10^{-13} para medio-alto y 2.13×10^{-13} para precio alto.

- Relación Largo-Ancho: Al igual que el peso, esta variable presenta coeficientes positivos y de gran magnitud, lo que indica un fuerte impacto sobre el precio. Los valores son 12,474.1 para diamantes de precio medio-bajo, 1,321.3 para medio-alto y 104,564.1 para alto. Esto evidencia que una mayor proporción largo-ancho incrementa significativamente la probabilidad de que el diamante sea más caro.

- Claridad: Los coeficientes son positivos en todos los casos, lo que sugiere que una mayor claridad del diamante se asocia con una mayor probabilidad de alcanzar un precio más alto. Los valores específicos son 0.317 para precio medio-bajo, 0.280 para medio-alto y 0.259 para alto, respecto a la categoría de precio bajo.

- Profundidad: Esta variable también muestra un efecto positivo sobre el precio, con coeficientes de 0.749 para diamantes de precio medio-bajo, 0.687 para medio-alto y 0.655 para alto. Esto implica que una mayor profundidad del diamante incrementa la probabilidad de un precio superior.

- Forma: Los valores de los coeficientes son 0.851 para precio medio-bajo, 0.991 para medio-alto y 0.778 para alto. Todos son positivos, lo que indica que una mejor forma contribuye al aumento en la probabilidad de un mayor precio.

- Largo: En esta variable, también se observa un patrón positivo: 0.142 para diamantes de precio medio-bajo, 0.296 para medio-alto y 0.233 para alto. Esto sugiere que un mayor largo del diamante incrementa la probabilidad de un precio más elevado, aunque el efecto es moderado.

- Altura: Finalmente, la altura del diamante presenta coeficientes positivos y crecientes: 1.006 para precio medio-bajo, 2.846 para medio-alto y 4.619 para alto. Esto refleja una relación clara entre mayor altura y probabilidad de que el diamante pertenezca a una categoría de precio superior.

Finalmente, se analiza este modelo de manera independiente con el objetivo de comprender mejor sus fortalezas. El análisis comienza con una tabla de contingencia:

	X1	X2	X3	X4
X1	1248	97	3	3
X2	43	877	68	43
X3	8	242	882	301
4	7	79	346	943

	X1	X2	X3	X4
X1	305	23	1	0
X2	12	237	13	9
X3	5	46	226	69
X4	4	17	84	244

Tabla 17: Tablas de contingencia regresión logística multinomial. Entrenamiento y prueba.

La tabla ubicada a la izquierda presenta el desempeño del modelo en el conjunto de entrenamiento, detallando tanto las predicciones correctas como las incorrectas. En contraste, la tabla a la derecha expone los resultados de las predicciones realizadas sobre el conjunto de prueba. Se calcula la tasa de acierto de ambas bases:

$$Tasa\ de\ Acierto\ Entrenamiento = \frac{3950}{5190} = 0.761 \qquad Tasa\ de\ Acierto\ Prueba = \frac{1012}{1295} = 0.781$$

Las tasas de acierto de ambas bases son bastante buenas. A continuación, se calcula el Índice de Kappa de Cohen, que evalúa si las predicciones son mejores que lo que cabría esperar por azar. Para la base de prueba, conseguimos un valor de 0.709 que implica que el modelo tiene una concordancia buena, teniendo un desempeño mejor que el azar.

Posterior a esto, se saca por pantalla la sensibilidad y la especificidad del modelo. Mientras que la sensibilidad refleja la proporción de casos positivos reales que han sido correctamente identificados como tales por el modelo (es decir, pertenecer a una clase y ser clasificados en esa misma clase), la especificidad representa la proporción de casos negativos reales que han sido correctamente reconocidos como negativos (es decir, no pertenecer a una clase y ser clasificados como tales por el modelo).

	Sensibilidad	**Especificidad**
X1	0.936	0.975
X2	0.734	0.965
X3	0.698	0.876
X4	0.758	0.892

Tabla 18: Sensibilidad y especificidad regresión logística multinomial.

A diferencia de lo observado en la regresión logística multinomial realizada con los factores retenidos, en este caso se obtuvieron valores satisfactorios tanto de sensibilidad como de especificidad para todas las clases definidas.

Para finalizar este apartado, se presenta la curva ROC correspondiente al modelo seleccionado. El área bajo la curva (AUC) alcanzó un valor de 0.934, lo que indica un excelente desempeño global del modelo. La clase que mostró una mejor capacidad predictiva fue X1 (precio bajo), lo cual puede estar influido por el hecho de que fue tomada como clase de referencia al construir el modelo. Por otro lado, la clase X3 (precio medio-alto) fue la que presentó un rendimiento comparativamente inferior; sin embargo, con un valor de predicción de 0.896, este sigue considerándose elevado. En conjunto, este modelo demostró un desempeño superior al obtenido mediante la regresión logística multinomial basada en los factores retenidos.

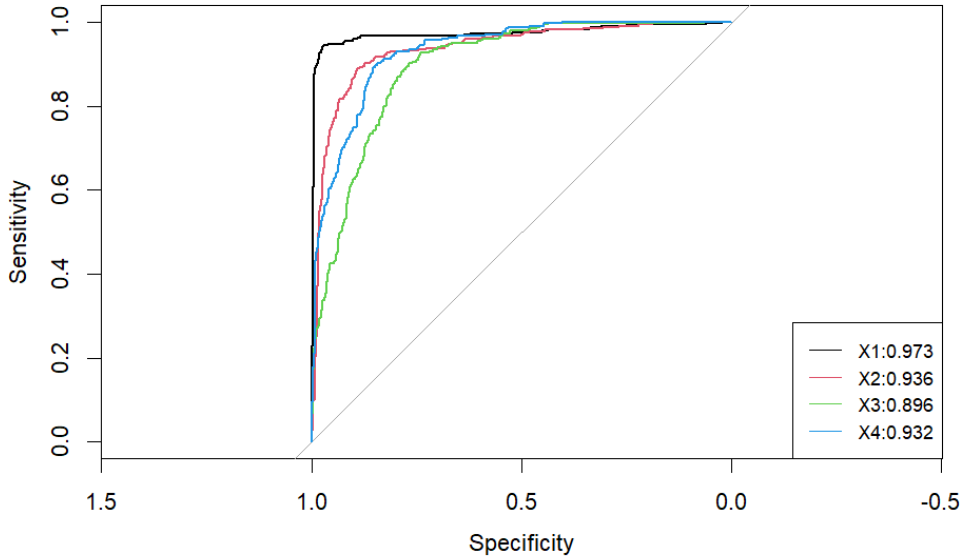


Gráfico 22: Curva ROC regresión logística multinomial

3.3. Árbol de regresión con los valores originales de la variable precio

Previo al desarrollo del modelo, se realizó una partición de la base de datos, asignando el 80% de los registros al conjunto de entrenamiento y el 20% restante al conjunto de prueba.

Tal como se ha expuesto previamente, el árbol de regresión es un modelo cuyo objetivo es predecir una variable dependiente de naturaleza continua. Una de las principales ventajas del árbol de regresión es su capacidad para captar relaciones no lineales entre las variables, así como su interpretación intuitiva, ya que el modelo se puede visualizar como un esquema jerárquico de decisiones. No obstante, es importante tener en cuenta que, si no se aplican técnicas de poda o regularización, el modelo puede sobreajustarse a los datos de entrenamiento, disminuyendo así su capacidad de generalización.

Por ello, se comienza construyendo un árbol de gran tamaño, sin restricciones, con el fin de capturar toda la complejidad presente en los datos.

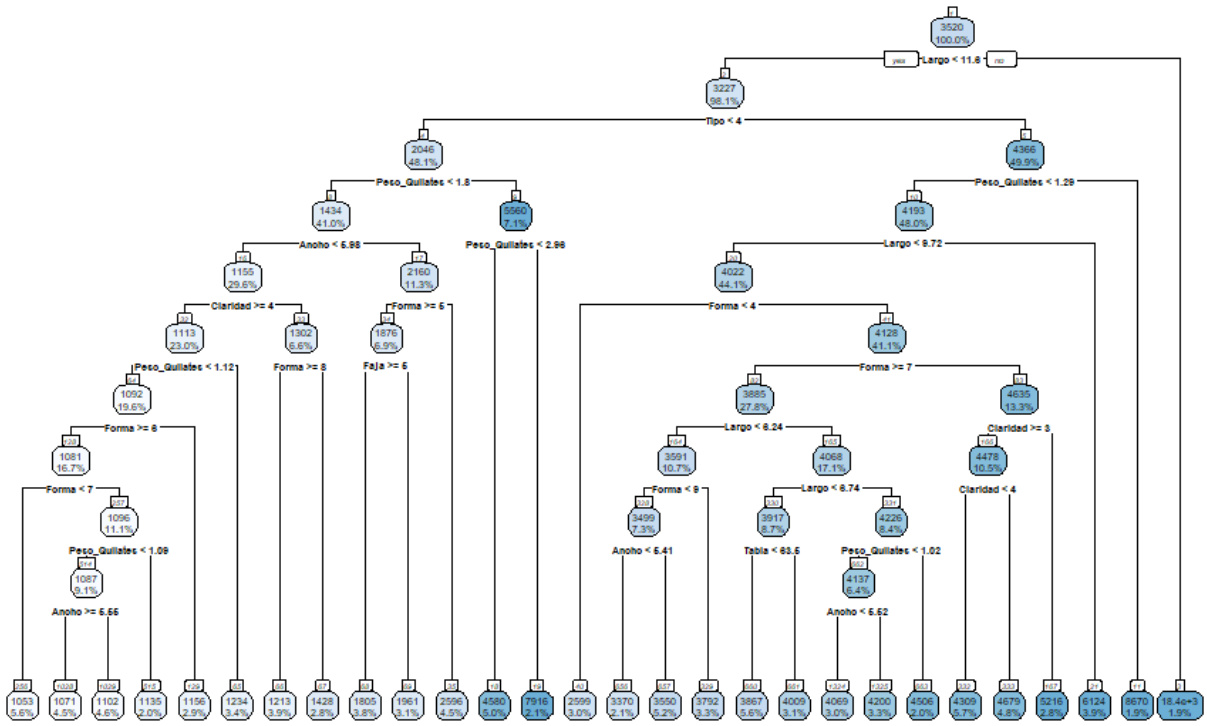


Gráfico 23: Árbol de regresión

Este árbol cuenta con 10 niveles de profundidad y un total de 28 hojas. El modelo, tal como está, resulta ambiguo y difícil de interpretar, por lo que no se considera adecuado como modelo final. Para construir uno más apropiado, será necesario evaluar su fiabilidad y aplicar criterios que simplifiquen su estructura, mejorando su capacidad de interpretación.

Con el fin de evaluar la calidad del modelo, se calcula el R^2, el cual es 0.727 en el conjunto de entrenamiento y 0.694 en el conjunto de prueba. El R^2 indica la capacidad del modelo para explicar la variabilidad de la variable dependiente, y en este caso, los resultados obtenidos son satisfactorios.

En cuanto a la importancia de las variables en el modelo, se observa que Largo se destaca como la variable con mayor poder predictivo, lo que indica que tiene una influencia significativa en la clasificación de los individuos en sus respectivas categorías. Le siguen Peso_Quilates y Tipo, que también aportan información relevante al proceso de discriminación, aunque en menor medida. En contraste, el resto de las variables presentan un peso predictivo considerablemente menor, lo que sugiere que su contribución al modelo es limitada en comparación con las variables principales.

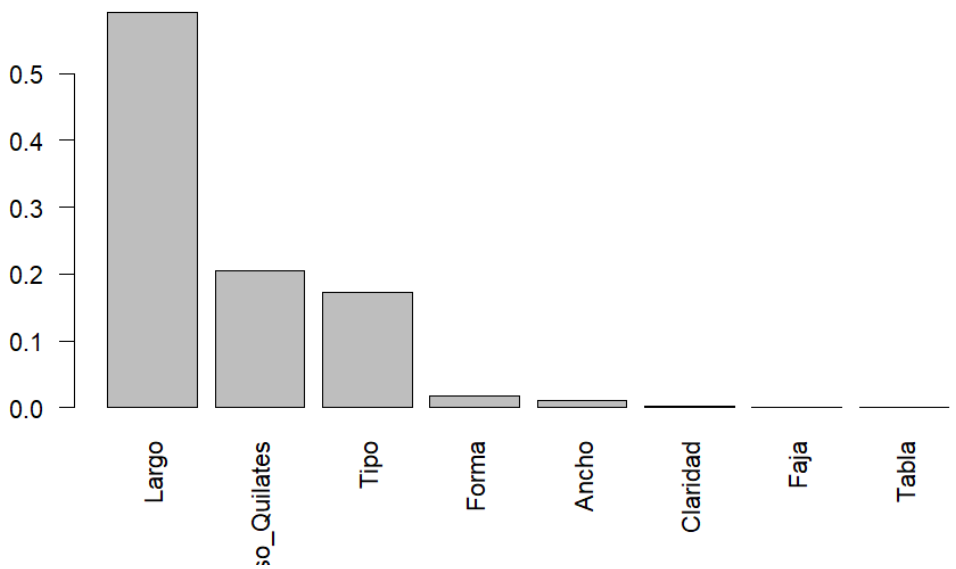


Gráfico 24: Importancia de las variables

Con este modelo, se procede a realizar la poda. En el siguiente gráfico, se identifica el punto óptimo para podar el árbol, donde se minimiza el error cometido sin perder información relevante del modelo. En este caso, se elige un valor de $cp = 0.011$.

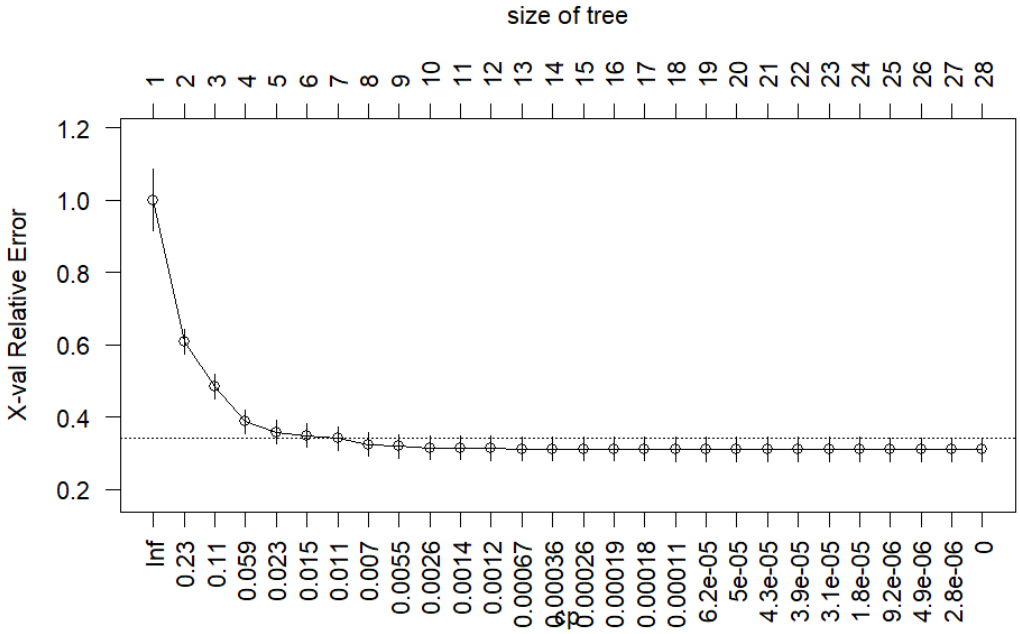


Gráfico 25: Corte óptimo árbol de regresión.

Como resultado se consigue el siguiente árbol, que cuenta con una profundidad de 4 niveles y un total de 7 hojas. Con esto, se ha logrado reducir enormemente el tamaño del árbol y mejorar su interpretabilidad. Al reducir el tamaño del árbol mediante la poda, eliminamos ramas que no contribuyen significativamente a la predicción, lo que ayuda a evitar el sobreajuste y mejora la capacidad del modelo para generalizar sin perder valor predictivo.

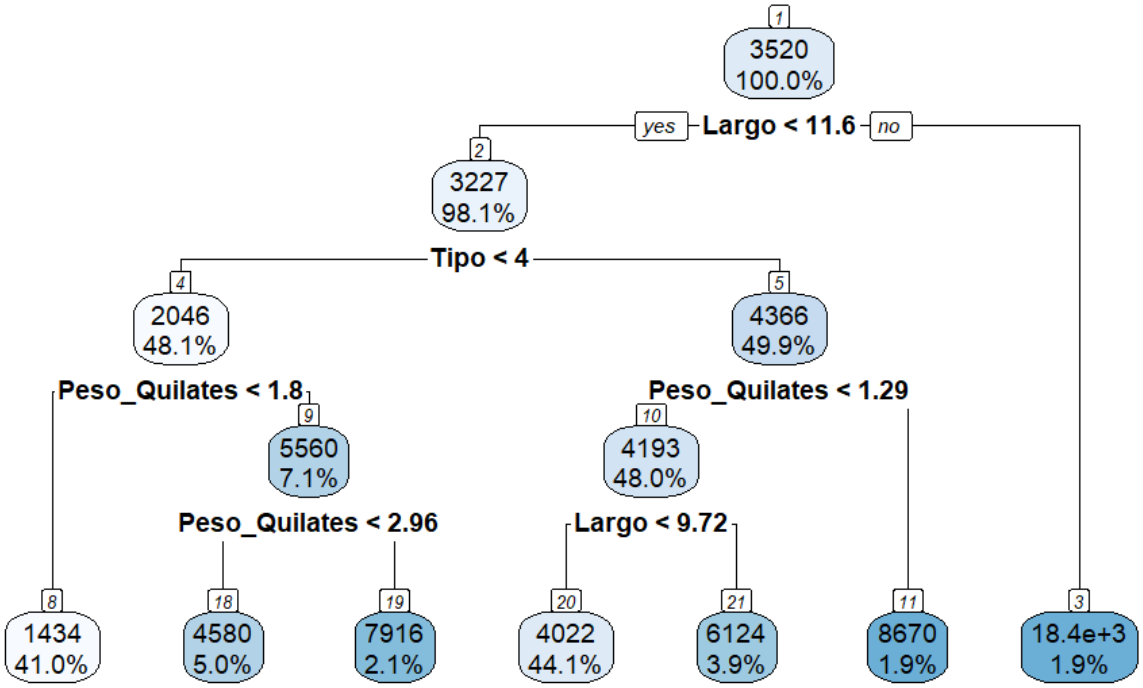


Gráfico 26: Árbol de regresión.

Una posible manera de interpretarlo sería, por ejemplo: los diamantes que cuentan con un Largo menor de 11.6, un Tipo igual o mayor de 4, un Peso_Quilates menor de 1.29 y un Largo menor de 9.72 tendrán un precio aproximado de 4022$, constituyendo un 44.1% de la muestra.

Finalmente, es crucial evaluar el valor del R^2, ya que este indicador permite valorar la efectividad del modelo. Se obtuvo un valor de 0.703 en el conjunto de entrenamiento y de 0.670 en el conjunto de prueba. A pesar de la considerable reducción en el tamaño del árbol, la disminución en la capacidad de explicar la variabilidad de la variable dependiente ha sido mínima.

En este caso, también es preferible llevar a cabo el análisis sin agrupar las variables en factores, ya que se ha logrado mejorar el R^2 y se obtiene una clasificación directa de los diamantes, sin necesidad de recurrir a transformaciones previas.

3.3. Árbol de clasificación con las clases creadas

Para evaluar los resultados del árbol de clasificación, se divide el conjunto de datos, asignando el 80% para entrenar el modelo (data_train) y el 20% restante para validar su desempeño (data_test).

Se genera un árbol de clasificación con varios niveles, con el objetivo de realizar una poda posterior, aunque este modelo inicial puede resultar complejo de interpretar. El árbol resultante es el siguiente:

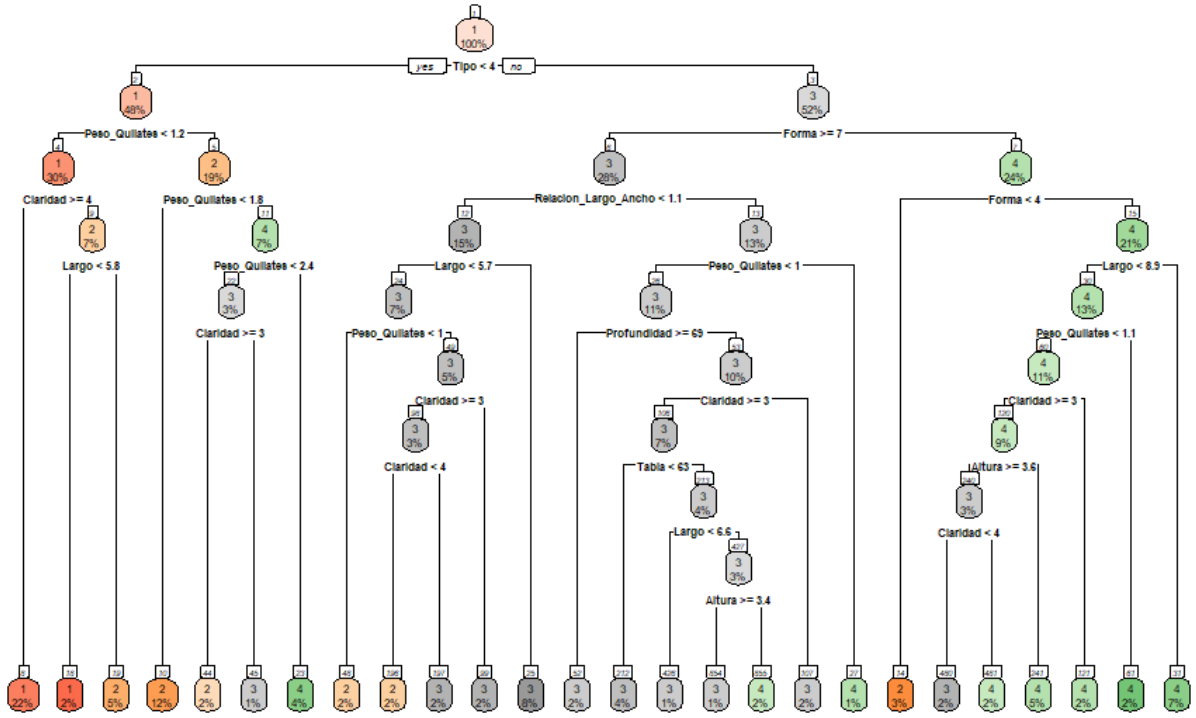


Gráfico 27: Árbol de clasificación.

Este árbol, cuenta con una profundidad de 9 niveles y 26 hojas. Resulta difícilmente interpretable de esta manera por lo que más adelante se procederá con la poda.

Las variables más importantes en este modelo, Peso_Quilates, Tipo y Forma. Estas características se consideran fundamentales para predecir la clase de los diamantes, probablemente debido a su relación directa con las categorías de precio, que es la variable a predecir. El resto de las variables son de menor importancia predictiva

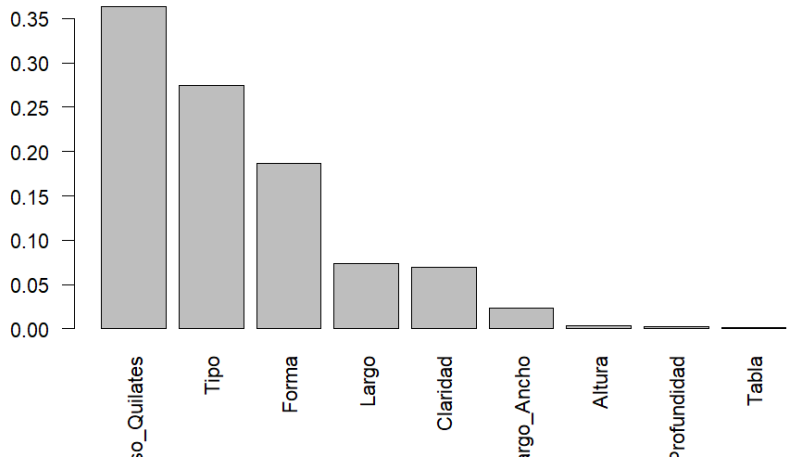


Gráfico 28: Importancia de las variables

Para determinar el tamaño óptimo del árbol, se emplea la técnica de poda con el fin de reducir el sobreajuste, asegurando que no se pierda información relevante. Se elige un valor de $cp = 0.0024$

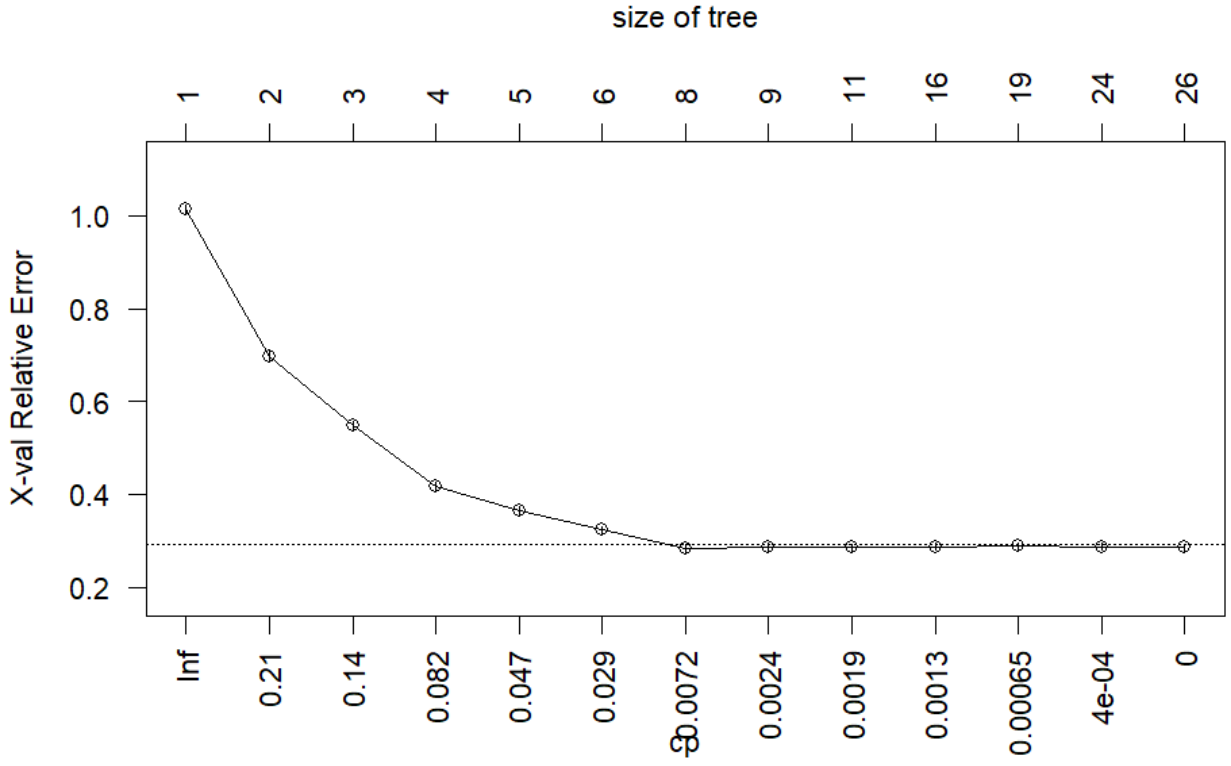


Gráfico 29: Corte óptimo árbol de clasificación.

El árbol resultante presenta una estructura de 4 niveles y 9 hojas. Dado que se trata de un árbol de clasificación, no es posible calcular su R^2; sin embargo, es factible obtener la matriz de confusión, la tasa de acierto, el índice de Kappa, así como las métricas de sensibilidad y especificidad.

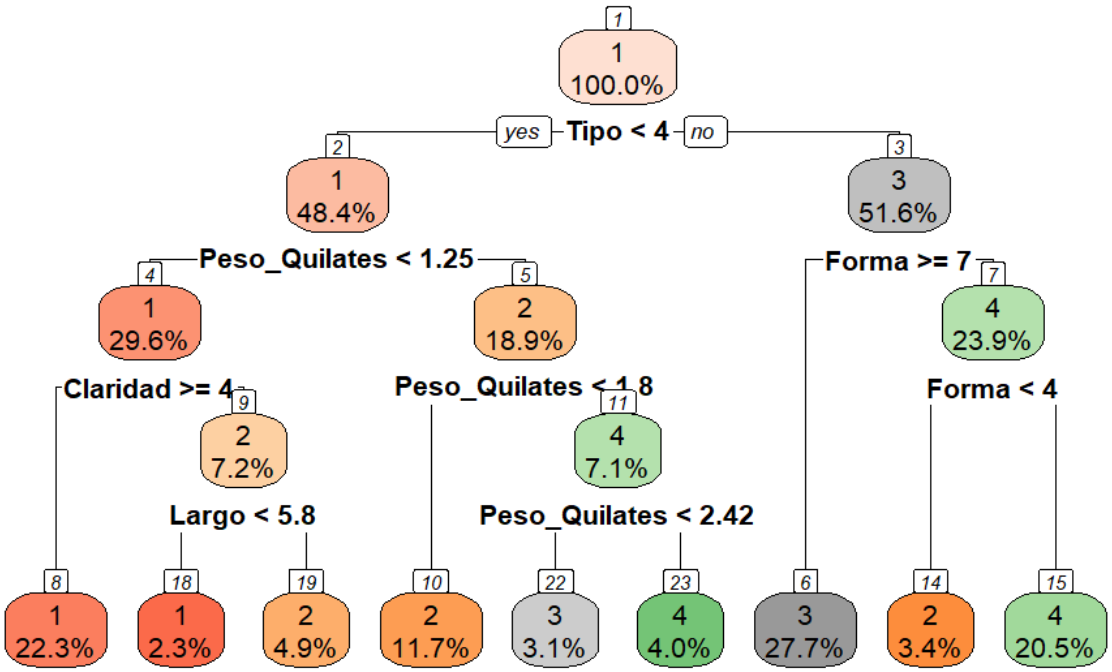


Gráfico 30: Árbol de clasificación.

Este árbol presenta una mayor variedad de colores, cada uno de los cuales indica una clase específica. El color rojo corresponde a la clase 1, asociada al precio bajo; el naranja a la clase 2, que representa precio medio-bajo; el gris a la clase 3, vinculada al precio medio-alto; y, finalmente, el verde a la clase 4, indicando precio alto.

Una interpretación posible sería la siguiente: los diamantes con tipo menor a 0.4, un Peso_Quilates inferior a 1.25 y una Claridad mayor o igual a 0.4 representan el 22.3% de la muestra, correspondiendo a diamantes de precio bajo.

Finalmente, es necesario evaluar las tasas de acierto y los indicadores de calidad del modelo.

	X1	X2	X3	X4
X1	1228	46	3	2
X2	64	928	31	15
X3	7	285	992	314
X4	10	32	273	959

	X1	X2	X3	X4
X1	298	8	1	0
X2	17	229	7	5
X3	3	75	232	87
X4	5	15	84	230

Tabla 19: Tablas de contingencia árbol de clasificación. Entrenamiento y prueba.

La tasa de acierto obtenida en la base de entrenamiento es de 0.791, lo que se considera un valor alto, dado que los aciertos al azar en un escenario con cuatro clases serían de 0.25. En cuanto a la base de prueba tiene una tasa de acierto de 0.763, que es ligeramente inferior, pero sigue siendo buena.

Evaluamos a continuación El Índice Kappa en ambas bases. Para la base de entrenamiento es un valor de 0.722 y para la base de prueba es de 0.684, lo cual se considera un valor bueno, para ambas bases.

Dado que ambos valores no difieren significativamente de los obtenidos en la base de entrenamiento, podemos concluir que no hay evidencia de sobreajuste.

Para la sensibilidad y especificidad obtenemos en ambas bases:

	Sensibilidad	**Especificidad**
Base de entrenamiento	0,938	0,719
Base de prueba	0,923	0,7

Tabla 20: Sensibilidad y especificidad árbol de clasificación.

En ambos casos se consiguen buenos valores para la sensibilidad y especificidad. Se captan bien a los que pertenecen a cada grupo y a los que no pertenecen a los grupos.

4. Conclusión del estudio

4.1. Conclusiones del estudio a partir del análisis factorial

El estudio ha demostrado que el precio de los diamantes está principalmente determinado por una combinación de factores, entre los cuales el tamaño y la calidad destacan como los más influyentes. En general, los diamantes más grandes y de mejor calidad tienen una mayor probabilidad de alcanzar las categorías de precio más alto. No obstante, variables como la forma, la proporción y la geometría también desempeñan un papel importante, aunque su influencia no siempre es lineal ni directa.

Se evidencia que para que un diamante sea valorado con un precio elevado, no basta con poseer una sola característica destacada; es necesaria una interacción favorable entre varias propiedades. Por ejemplo, un buen tamaño y geometría adecuada suelen ser determinantes, pero en ciertos casos, incluso diamantes con calidad moderada o baja pueden alcanzar un precio alto si presentan una forma atractiva o proporciones específicas que mejoran su valoración estética.

Además, algunas características que se podrían considerar positivas, como una geometría o proporción muy alta, pueden estar asociadas con una menor probabilidad de un precio elevado, lo que sugiere la existencia de rangos óptimos para estas variables, más allá de los cuales no aportan valor adicional e incluso pueden disminuir la valoración comercial.

Dado que las variables se relacionan entre sí, la siguiente tabla muestra los valores esperados del precio según las distintas combinaciones de dichas variables. Los valores resaltados en verde corresponden a condiciones que se cumplen de forma constante, mientras que los señalados en rojo representan casos que ocurren solo bajo ciertas combinaciones específicas.

Variable	Precio Alto	Precio Medio-Alto	Precio Medio-Bajo	Precio Bajo
Geometría	≥ 2.46	< 2.46 (en combinación con otras variables)	< 0.154 (combinación con forma y proporción)	≥ -0.372 (con calidad o proporción altas)
Calidad	≥ 0.12	< 0.138	≤ 0.138 (con forma alta y geometría baja)	≤ 0.138 (con forma baja o proporción alta)
Tamaño	≥ 0.521	≥ 0.237 y ≤ 1.37	< 0.237	< 0.237 o < 1.37
Forma	≥ -0.707	< -0.707 (combinado con tamaño y calidad)	≥ -0.134 (con geometría baja)	< -0.134 (combinado con proporción y geometría)
Proporción	-	≥ 0.403 (en combinación con forma y geometría)	< 0.403	≥ 0.403 (con forma baja y geometría alta)

Tabla 21: Resultados a partir del Análisis Factorial

4.2. Conclusiones del estudio sin el análisis factorial

A través de los distintos modelos utilizados en este estudio se ha podido identificar un conjunto de variables que, de forma consistente, influyen en el precio de los diamantes. Aunque cada modelo aplica un enfoque distinto, los resultados muestran una coincidencia clara en varios factores clave.

Uno de los aspectos que más destaca es la importancia del peso en quilates. Todos los modelos coinciden en que, a mayor peso, mayor es la probabilidad de que el diamante pertenezca a una categoría de precio alto. En la regresión logística, este efecto se ve reforzado por coeficientes y odds ratios muy elevados, y tanto en los árboles como en el análisis discriminante, el peso de los quilates aparece como una de las variables que mejor separa las distintas clases de precio.

También se observa una fuerte influencia del tipo del diamante. Esta variable, que agrupa aspectos generales de calidad, se relaciona con precios más altos especialmente en los niveles medio-alto y alto. Aunque su impacto no es tan extremo como el del peso, sí aporta una diferenciación importante en el modelo discriminante y en los árboles.

Por otro lado, hay un grupo de variables relacionadas con el acabado y la calidad visual del diamante (como la claridad, el pulido, la simetría o el culet) que muestran un impacto positivo y bastante constante en los tres modelos. Estas características aportan valor añadido y parecen ser especialmente relevantes cuando el peso no es excesivamente alto, reforzando la idea de que la percepción de calidad influye directamente en el precio final.

Las proporciones físicas del diamante, especialmente la relación largo-ancho, también aparecen como variables relevantes, aunque con una influencia algo más moderada. Según los modelos, una mejor proporción favorece el posicionamiento en precios altos, aunque este efecto depende de cómo se combinen con otras características.

En general, los modelos no solo coinciden en qué variables son importantes, sino que también muestran que el precio se determina por una combinación de factores, más que por una sola característica. Es decir, un diamante pesado no siempre será caro si no tiene buena simetría, o si su forma no es apreciada.

Variable	Precio Bajo	Precio Medio-Bajo	Precio Medio-Alto	Precio Alto
Peso (quilates)	< 1.25	1.25 – 2.42	1.25 – 1.8	≥ 1.29
Tipo	< 4	≥ 4	< 4	≥ 4
Claridad	Baja (< 4)	< 4	< 4	≥ 4
Largo	< 5.8 o < 9.72	< 11.6	5.8 – 9.72	≥ 11.6
Altura	Baja	Media	Media (↑)	Alta (↑)
Relación Largo/Ancho	Baja	Variable	Media (↑)	Alta (↑)
Forma	≥ 4 y < 7	≥ 7	< 4 y < 7	< 4
Culet	Positivo (↓)	Positivo (↓)	Positivo (↑)	Positivo (↑)
Otros factores	Proporción y forma débiles	Claridad baja	Ancho y pulido moderado	Profundidad media-alta

Tabla 22: Resultados sin el Análisis Factorial

4.3. Conclusiones combinando ambos estudios

Para concluir el estudio, se analizan todos los resultados obtenidos sacando conclusiones claras:

En la categoría de precio alto, los diamantes se distinguen por su excelente calidad tanto técnica como visual. Los puntajes elevados en los análisis factoriales indican que estos diamantes reúnen atributos deseables de manera consistente. En particular, se trata de piedras grandes (peso ≥ 1.29 quilates), con formas apreciadas, alta claridad (pocas imperfecciones), y proporciones óptimas (profundidad ≥ 11.6%). Además, presentan una buena relación entre largo y ancho, lo que favorece su simetría y brillo. Estos diamantes reflejan un alto nivel de corte, pulido y geometría, justificando así su valor elevado.

Los diamantes de precio medio-alto también muestran características positivas, aunque con algunas limitaciones. Su tamaño es competitivo (entre 1.25 y 1.8 quilates), pero suelen tener formas y niveles de claridad algo menores. La proporción y la profundidad son adecuadas, y su forma general es armónica. Aunque no alcanzan los estándares más altos, combinan varios atributos de calidad que los hacen atractivos dentro de una gama superior media.

En la categoría precio medio-bajo, los diamantes presentan calidad variable. Aunque algunos tienen buena forma, suelen mostrar baja claridad y proporciones menos ideales. Tienen una geometría más débil, y su simetría visual puede no ser del todo equilibrada. Esto se traduce en un valor más accesible.

Finalmente, los diamantes de precio bajo muestran deficiencias claras en casi todas las dimensiones. Son pequeños (menos de 1.25 quilates), con formas poco valoradas, claridad baja y proporciones desbalanceadas. La geometría suele ser pobre y la relación entre largo y ancho, desfavorable. Estos diamantes no cumplen con los estándares óptimos de corte ni de simetría, lo que explica su bajo precio en el mercado.

5. Bibliografía

- Valencia Delfa, J., & Vicente Hernanz, Á. (2015). *Análisis multivariante I*. Cersa.
- Alonso Revenga, J. M., & Calviño Martínez, N. (2025). *Introducción a la ciencia de datos con R: Análisis supervisado*. García Maroto Editores.
- Alonso Revenga, J. M., & Calviño Martínez, N. (2025). *Introducción a la ciencia de datos con R: Preparación de los datos y análisis no supervisado*. García Maroto Editores.