



**FACULTAD  
DE INGENIERIA**  
Universidad de Buenos Aires

**75.50**

**Introducción a Sistemas  
Inteligentes  
TP - Final**

**Profesores:** Ochoa María Alejandra

**Alumnos:**

Nombre y Apellido	Padrón	Email
Maximiliano Scoccimarro	93784	maximiliano.scoccimarro@gmail.com
Mariano Ezequiel Andres	96838	mariando.andres@gmail.com

**Cuatrimestre:** 2° 2018



# Índice

<b>Índice</b>	<b>1</b>
<b>1- Comprensión del negocio</b>	<b>3</b>
Objetivos del negocio	3
Escenario actual	3
Objetivos del negocio	3
Criterios de éxito del negocio	3
Situación	3
Inventario de recursos	3
Requisitos, supuestos y requerimientos	4
Riesgos y contingencias	4
Terminología	4
Costos y beneficios	4
Objetivos de la Minería de Datos	5
Objetivos de la minería de datos	5
Criterios de éxito de la minería de datos	5
Plan de proyecto	5
Plan de proyecto	5
Valoración de las herramientas	5
<b>2- Comprensión de los datos</b>	<b>6</b>
Recolección de datos Iniciales	6
Descripción de los datos	6
Descripción global del set de datos	6
Descripción de atributos	6
Exploración de los datos	11
Histogramas de atributos numéricos	11
Distribución del precio con respecto a atributos numéricos	20
Distribución atributos categóricos	28
Verificación de la calidad de los datos	33
<b>3- Preparación de los datos</b>	<b>35</b>
Selección de los datos	35



Inclusión/Exclusión de datos	35
Limpieza de los datos	35
Reporte de limpieza de datos	35
Estructura de los datos	37
Derivación de atributos	37
Discretización de atributos numéricos	37
Transformación de valores de atributos	38
Integración de los datos	38
Formateo de los datos	38
Data Set preparado	39
<b>4- Modelado</b>	<b>43</b>
Técnica de modelado	43
Algoritmo de Inducción (C4.5)	43
Funcionamiento	43
Red bayesiana (Naive Bayes)	43
Funcionamiento	44
Supuestos de modelado	44
Diseño de pruebas	44
Modelo	45
Algoritmo de inducción	45
Red bayesiana	50
Comparación	55
Reglas generadas	58
Evaluación del modelo	60
<b>5- Evaluación</b>	<b>62</b>
Valoración de los resultados	62
Revisión del proceso	62
Próximos Pasos	62
<b>Bibliografía</b>	<b>63</b>



# 1- Comprensión del negocio

## **Objetivos del negocio**

### **Escenario actual**

El descubrimiento de los diamantes se estima que fue hace 6000 años. Desde esa época hasta ahora siempre han sido atesorados por diversas razones, como por su belleza, extrema rareza, íconos religiosos o únicas características físicas como es su dureza. Debido a estas razones, los diamantes son hoy en día comercializados globalmente como gemas. Existen muchos negocios dedicados a la compra y venta de diamantes, para ello, necesitan tasarlos previamente. Dado que estas gemas tienen distintas características como peso, color, tipo de corte, medidas, etc. puede ser una tarea difícil de realizar.

### **Objetivos del negocio**

El objetivo del negocio es el de elaborar reglas de negocio que puedan ser utilizadas para construir un tasador de diamantes, de modo que ayude a los comercios o cualquier individuo interesado a valorizarlos.

### **Criterios de éxito del negocio**

Las reglas obtenidas deben ser simples y entendibles. Esto es, el 90% de los individuos que las utilicen no deben tener inconvenientes aplicándolas. Además, deben poder inferir un rango de precio de cualquier diamante observado.

## **Situación**

### **Inventario de recursos**

Los recursos disponibles para llevar a cabo el proyecto incluyen:

- Dos expertos en análisis y explotación de datos
- Una computadora de altas capacidades técnicas de hardware



- Software necesario para realizar gráficos y modelados.
- Un set de datos con miles de registros que contienen información de diamantes y sus respectivos precios.

## ***Requisitos, supuestos y requerimientos***

Un requerimiento sumamente importante es el tamaño del set de datos. Para realizar un análisis estadísticamente válido se requiere una cantidad mínima de 10.000 registros.

Se supone que los datos son reales y han sido recopilados correctamente. Además, no se tiene en cuenta la inflación en este estudio y se supone que todos los registros fueron recopilados en una ventana de tiempo pequeña donde los precios de los diamantes no variaron.

## ***Riesgos y contingencias***

Como riesgo es posible que las reglas obtenidas no reflejen la realidad, por lo tanto, la eficacia de las mismas no sea la esperada en un contexto de tasación de diamantes.

## ***Terminología***

A continuación se listan los términos y traducciones al español de algunos atributos que se encuentran en inglés (dado que el set de datos está en inglés).

Carat: quilate

Cut: corte

Price: precio

Clarity: claridad

Depth: profundidad

Table: tabla

Length: largo

Width: ancho

Height: altura

## ***Costos y beneficios***

Los costos para realizar este estudio incluyen la contratación de los expertos para realizar el análisis y el equipamiento de hardware y software. Gracias a que el set de datos ya se encontraba disponible no fue necesario gastar esfuerzo en su recopilación, lo cual reduce enormemente el costo del proyecto. Los beneficios son altos, ayudar a los individuos a tasar los diamantes les generará una alta reducción de los costos de



operaciones. Por esta razón, el costo-beneficio de realizar el proyecto es altamente positivo.

## **Objetivos de la Minería de Datos**

### **Objetivos de la minería de datos**

Realizar un análisis estadístico del set de datos y elaborar reglas y un modelo predictivo que dada las características de un diamante clasifique al mismo en el rango de precios correcto.

### **Criterios de éxito de la minería de datos**

La precisión de la clasificación sobre el set de datos de prueba debe ser del 70% o más. Además, la cantidad de reglas generadas debe ser entre 10 y 15. Todas las reglas deben tener una confianza del 50% o más.

## **Plan de proyecto**

### **Plan de proyecto**

El proyecto consiste en la ejecución de las siguientes tareas:

1. Descargar y almacenar el set de datos
2. Realizar un análisis estadístico de las distintas variables del set de datos, transformar los datos y depurar los registros y columnas innecesarias.
3. Construir un modelo de clasificación utilizando algoritmos de inducción y redes bayesianas.
4. Analizar los resultados generados por el modelo, evaluarlo y obtener las reglas.
5. Evaluar los resultados de la ejecución el proyecto.

### **Valoración de las herramientas**

Hemos elegido utilizar las herramientas Weka y Elvira para los procesos de explotación de datos del presente trabajo. Las elegimos porque tenemos conocimiento práctico con ambas y nos resultaron útiles en trabajos anteriores de explotación de datos, y porque soportan los algoritmos que vamos a aplicar sobre los datos. Vamos a usar Weka para aplicar algoritmos de inducción y Elvira para redes bayesianas.



## 2- Comprensión de los datos

### *Recolección de datos Iniciales*

El dataset se descargó de la página de kaggle (<https://www.kaggle.com/shivam2503/diamonds>).

Es un set de datos de un único archivo csv de casi 54000 registros. Cada registro contiene características del diamante y su respectivo precio.

### *Descripción de los datos*

#### *Descripción global del set de datos*

Formato: CSV (comma separated value)

Tamaño en disco: 3118 KB

Cantidad de registros: 53940

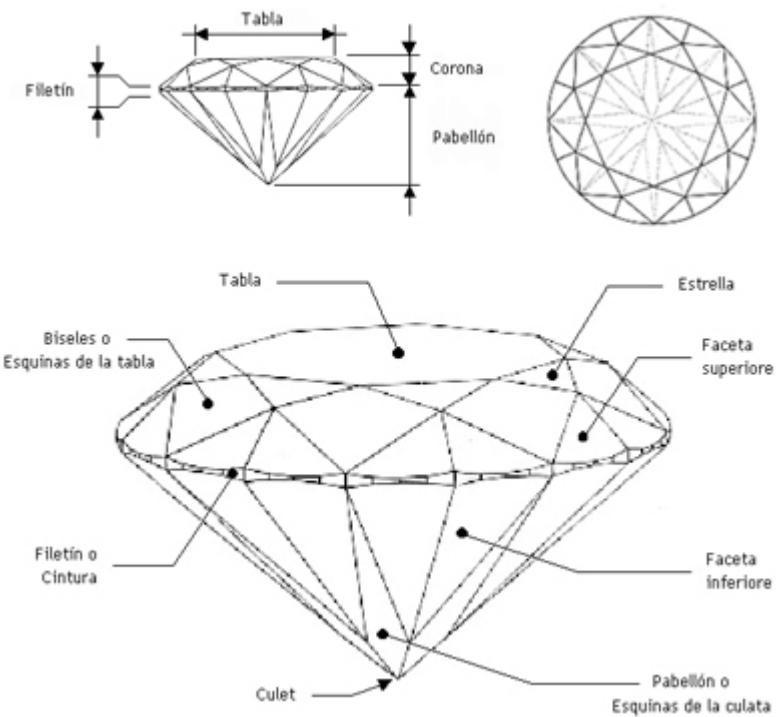
Cantidad de columnas: 10

#### *Descripción de atributos*

Nombre	Descripción	Tipo	Posibles valores/ Rango de valores
Identificador	Representa el identificador o índice de cada registro	Numérico	1-53940
Carat	Peso del diamante medido en quilates	Numérico	0.20-2.8
Cut	Calidad del corte del diamante	Categórico	Ideal, Premium, Very Good, Good, Fair
Color	Color del diamante	Categórico	D, E, F, G, H, I, J



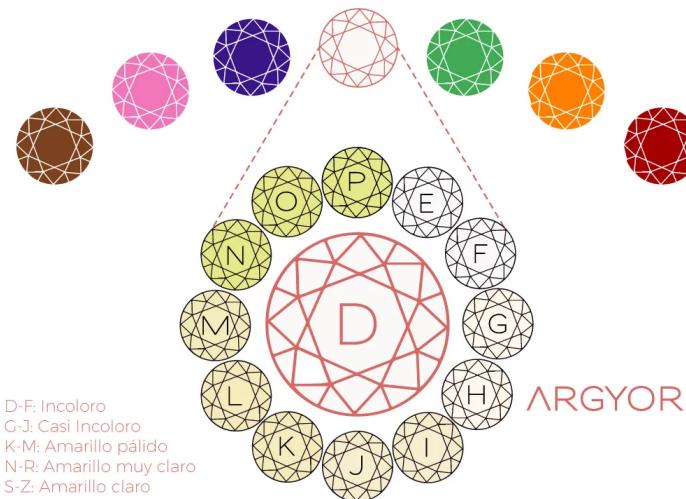
Clarity	Medida de ausencia de imperfecciones en el diamante	Categórico	IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1
Depth %	La altura del diamante, medida desde el culet hasta la tabla, dividida por su diámetro de filetín promedio	Numérico	0-100
Table %	El ancho de la tabla del diamante dividido por el ancho total	Numérico	0-100
Length	Largo del diamante en mm	Numérico	3.73-9.05
Width	Ancho del diamante en mm	Numérico	3.68-8.98
Height	Altura del diamante en mm	Numérico	1.53-5.60
Precio	Valor monetario en dólares del diamante	Numérico	326-18823



### Descripción detallada de las 4C

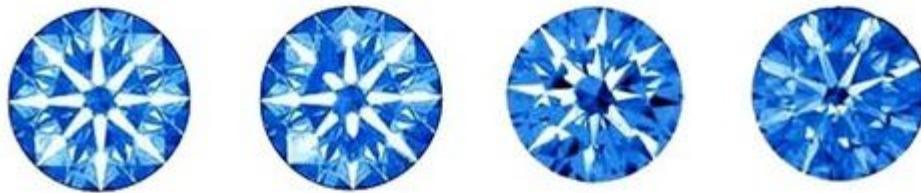
**Carat:** representa el peso de un diamante y se expresa con el número de quilates y dos decimales. Al momento de pesar un diamante, la norma internacional dice que sólo cuando la tercera cifra decimal es un 9, el peso debe indicarse pasando a la centésima superior. Por ejemplo, si el peso de un diamante da 1.631, 1.632, 1.633, 1.634, 1.635, 1.636, 1.637 o 1.638 se indica 1.63 y si es 1.639 se indica 1.64. Además, la regla de equivalencia con los gramos es 1 quilate = 0.2 gramos, o sea 1 gramo = 5 quilates.

**Color:** el color es una de las características más importantes del diamante. Cuanto más transparente es, mejor. De esta manera, las calificaciones de mejor a peor son D, E, F (incoloro), G, H, I, J (casi incoloro), K, L, M (amarillo pálido), N, O, P, Q, R (amarillo muy claro), S, T, U, V, W, X, Y, Z (amarillo claro).



**Cut:** la calidad del corte influye en el brillo, el fuego (cantidad de colores que 'viven' dentro del diamante) y centelleo (destellos de luz) de un diamante. En orden de mejor a peor calidad de corte tenemos: Ideal, Premium, Very Good, Good y Fair.

Cut



Excellent + Very Good.

These diamonds will have a high degree of brilliance, fire and scintillation.

Good

This grade will generally be a bit darker or lacking scintillation.

Fair

Diamonds in this category lack brightness, fire and scintillation.

Poor

Diamonds in this category show very little brightness, fire and scintillation.

**Clarity:** la claridad se refiere a la cantidad de imperfecciones o inclusiones que tiene el diamante, como por ejemplo, fracturas, un diamante dentro de otro, líquido, etc. Las categorías son las siguientes:

FL: significa Flawless o Perfecto. No presenta imperfecciones



IF: significa Internally Flawless o Internamente Perfecto. No presenta imperfecciones internas visibles con una lupa X10 de aumento, pero puede presentar alguna imperfección externa.

VVS1: significa Very Very Small Inclusions o Muy Muy Pequeñas Imperfecciones de grado 1. Presenta una única imperfección pequeña interna visible por un experto con una lupa X10 de aumento.

VVS2: significa Very Very Small Inclusions o Muy Muy Pequeñas Imperfecciones de grado 2. Presenta varias imperfección pequeñas internas visibles por un experto con una lupa X10 de aumento.

VS1: significa Very Small Inclusions o Muy Pequeñas Imperfecciones de grado 1. Presenta una única imperfección muy pequeña interna visible con una lupa X10 de aumento.

VS2: significa Very Very Small Inclusions o Muy Pequeñas Imperfecciones de grado 2. Presenta varias imperfección muy pequeñas internas visibles con una lupa X10 de aumento.

S1: significa Small Inclusions o Pequeñas Imperfecciones de grado 1. Presenta una única imperfección pequeña interna visible con una lupa X10 de aumento.

S2: significa Very Small Inclusions o Pequeñas Imperfecciones de grado 2. Presenta varias imperfección pequeñas internas visibles con una lupa X10 de aumento.

I1: Included o Imperfecto de grado 1: presenta una imperfección visible a simple vista.

I2: Included o Imperfecto de grado 2: presenta varias imperfecciones visibles a simple vista que también disminuyen el brillo.

I3: Included o Imperfecto de grado 3: presenta varias imperfecciones visibles a simple vista que también disminuyen el brillo y comprometen la estructura del diamante generando un riesgo de agrietamiento y/o ruptura.



## *Exploración de los datos*

### *Histogramas de atributos numéricicos*

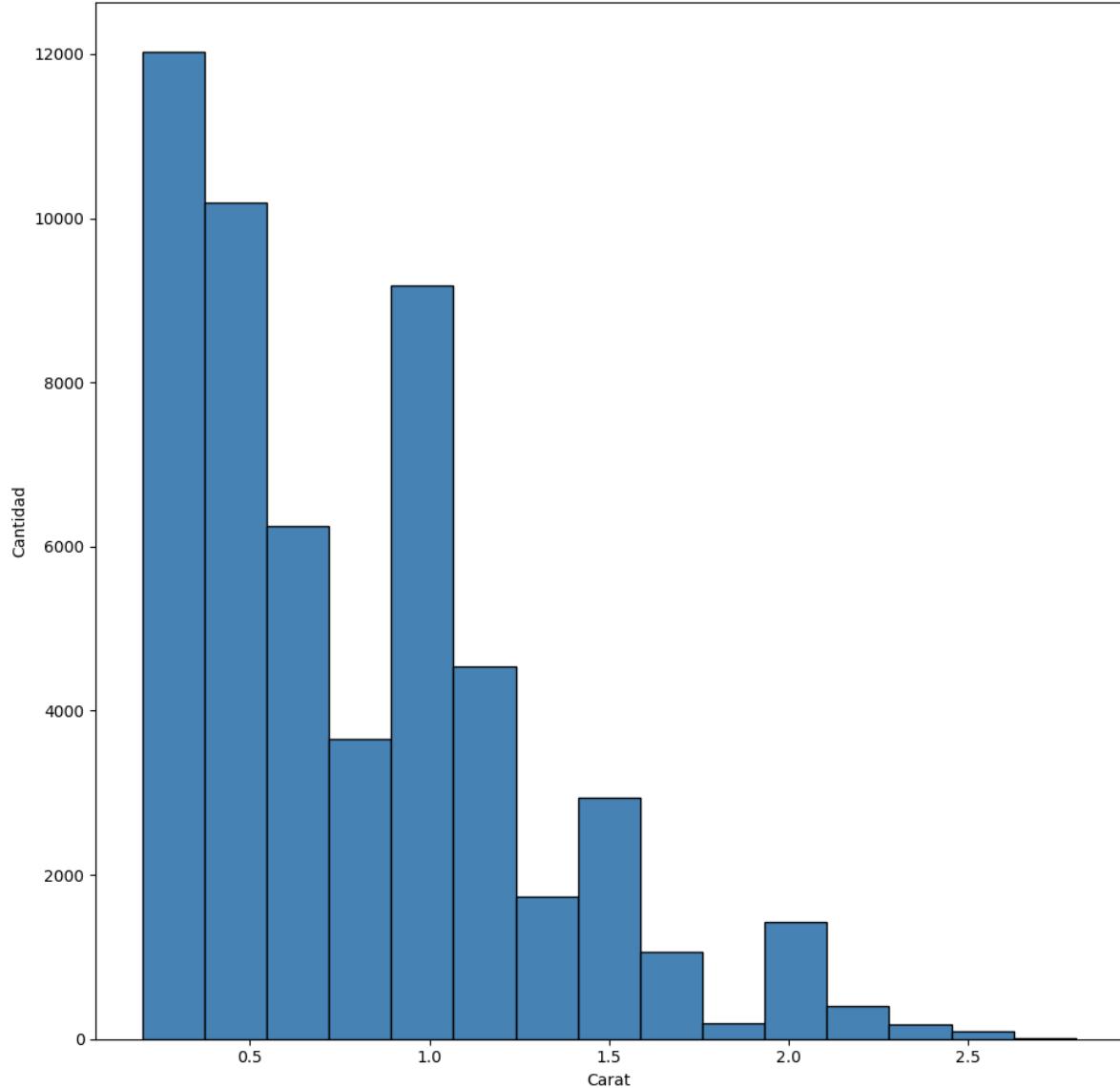
Antes de la exploración recordemos que podemos dividir el rango de valores de una variable numérica en cuartiles. Los cuartiles son los tres valores que dividen un conjunto de datos ordenados en cuatro partes porcentualmente iguales. Es decir, se arman 4 intervalos cada uno conteniendo 25% de los datos.

Primero veamos las distribuciones de los principales atributos numéricos en histogramas:



## Carat

Histograma Carat



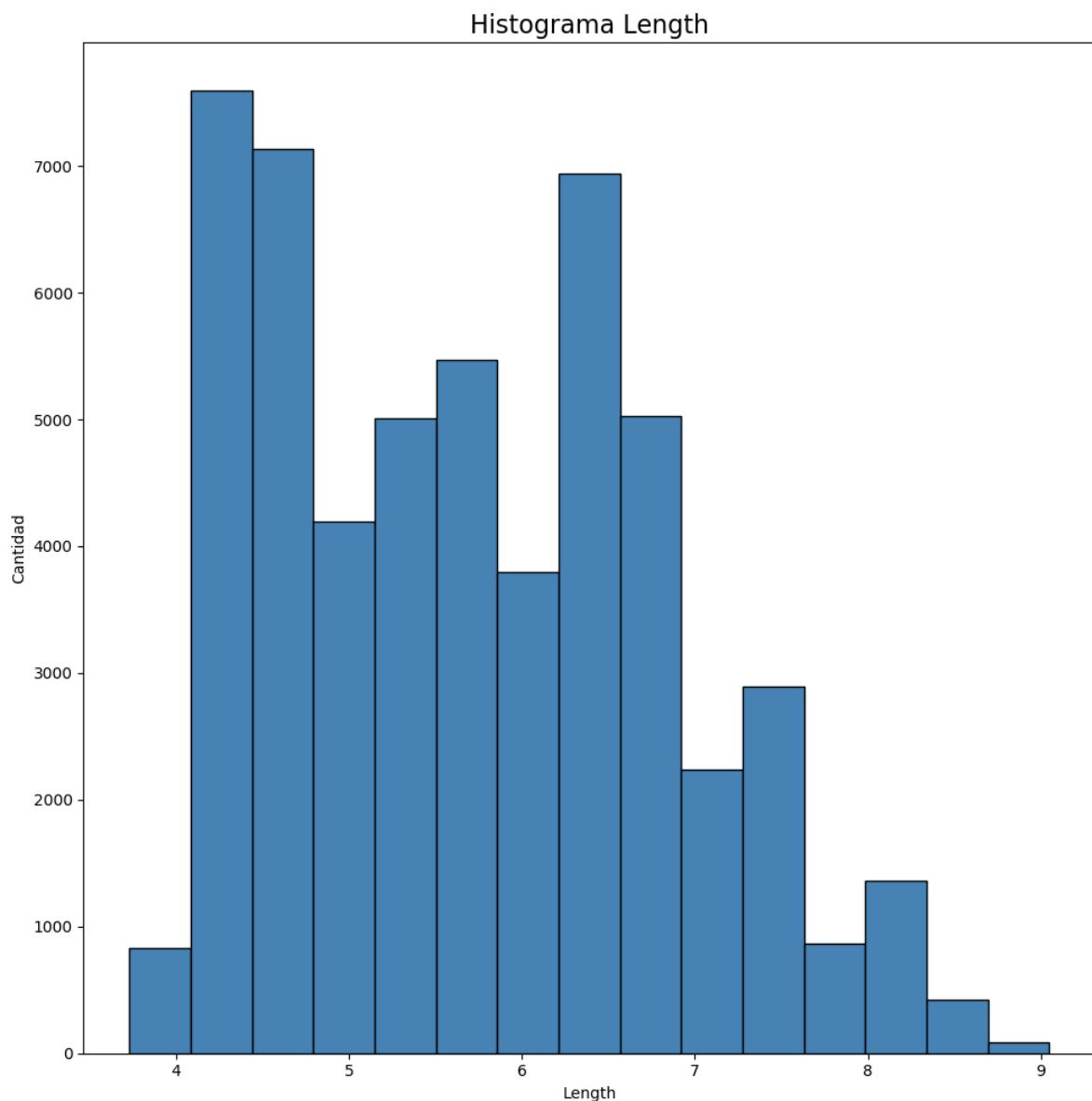
Media	0.796
Desvío estandar	0.468
Mínimo	0.2
Máximo	2.8
25%	(0.2, 0.4)
50%	(0.4, 0.7)



75%	(0.7, 1.04)
100%	(1.04, 2.8)

Podemos observar que la mayoría de los diamantes tienen un carat menor a 1.0, sin embargo algunos tienen un carat muy alto arriba de 2.0. Esto quiere decir que hay una gran varianza entre los diamantes.

## Length



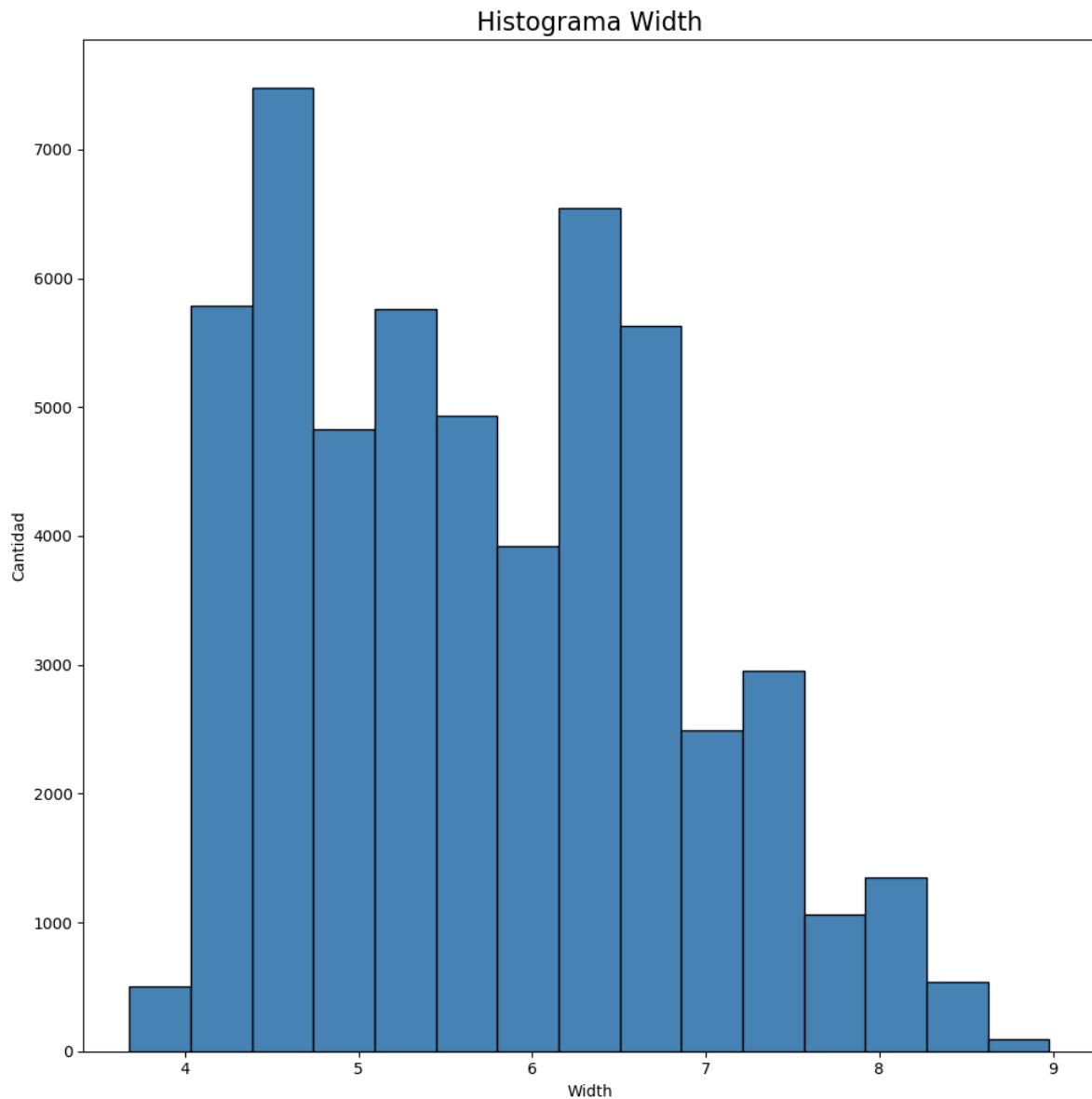


Media	5.728
Desvío estándar	1.114
Mínimo	3.73
Máximo	9.05
25%	(3.73, 4.71)
50%	(4.71, 5.7)
75%	(5.7, 6.54)
100%	(6.54, 9.05)

El length de los diamantes es muy variado. La mayoría siendo menor a 7.5. Hay unos pocos diamantes con un largo bastante alto.



## Width



Media	5.73
Desvío estándar	1.106
Mínimo	3.68
Máximo	8.98
25%	(3.68, 4.72)



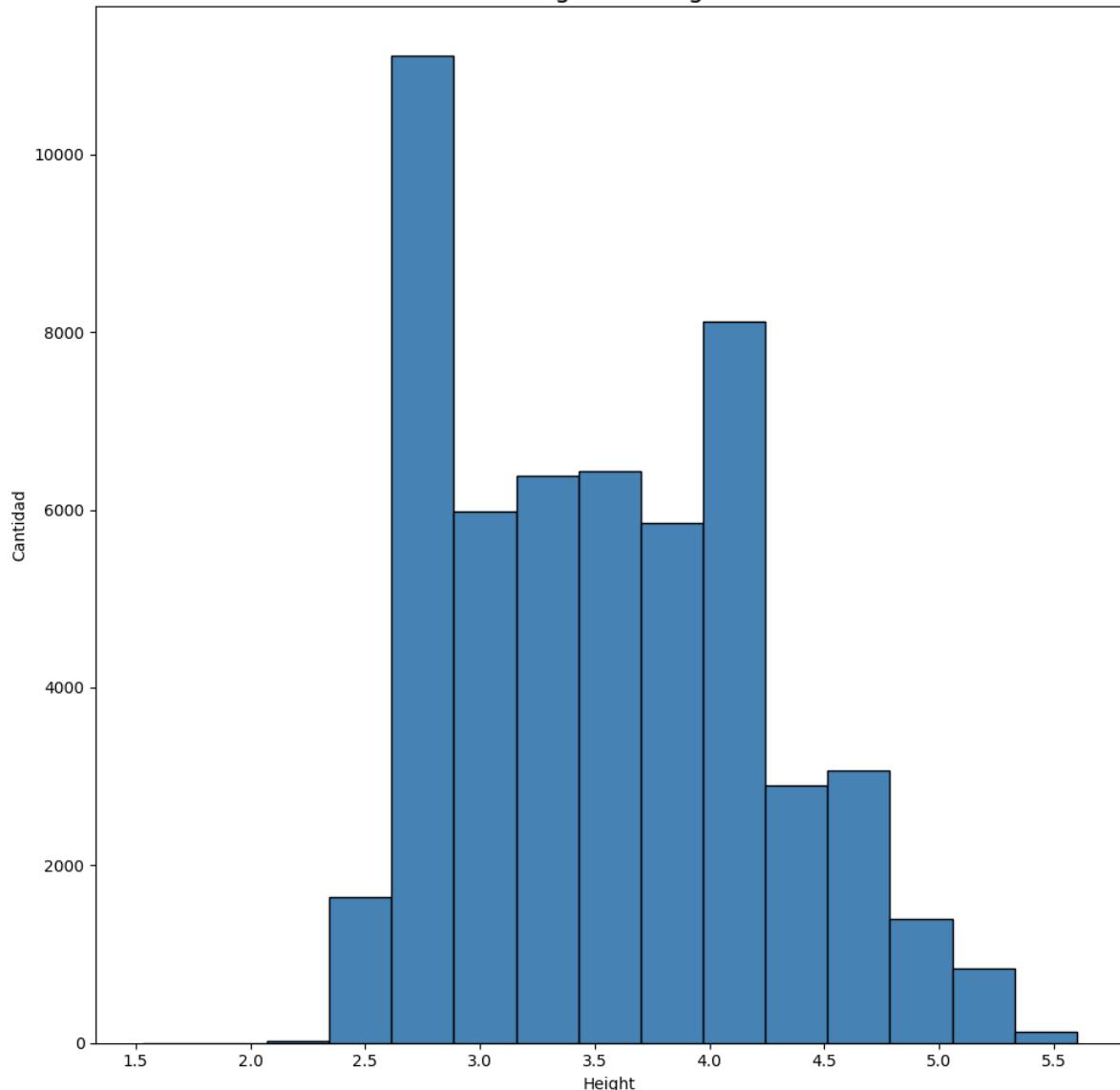
50%	(4.72, 5.71)
75%	(5.71, 6.53)
100%	(6.53, 8.98))

La distribución del width de los diamantes es muy similar a la del length. Esto se debe a que son atributos correlacionados, lo cual es lógico, ya que un diamante no puede tener un largo extremadamente alto y un ancho bajo. Si así lo fuera su forma sería muy extraña y dejaría de ser perfecta.



## Height

Histograma Height



Media	3.537
Desvío estándar	0.688
Mínimo	1.53
Máximo	5.6
25%	(1.53, 2.91)

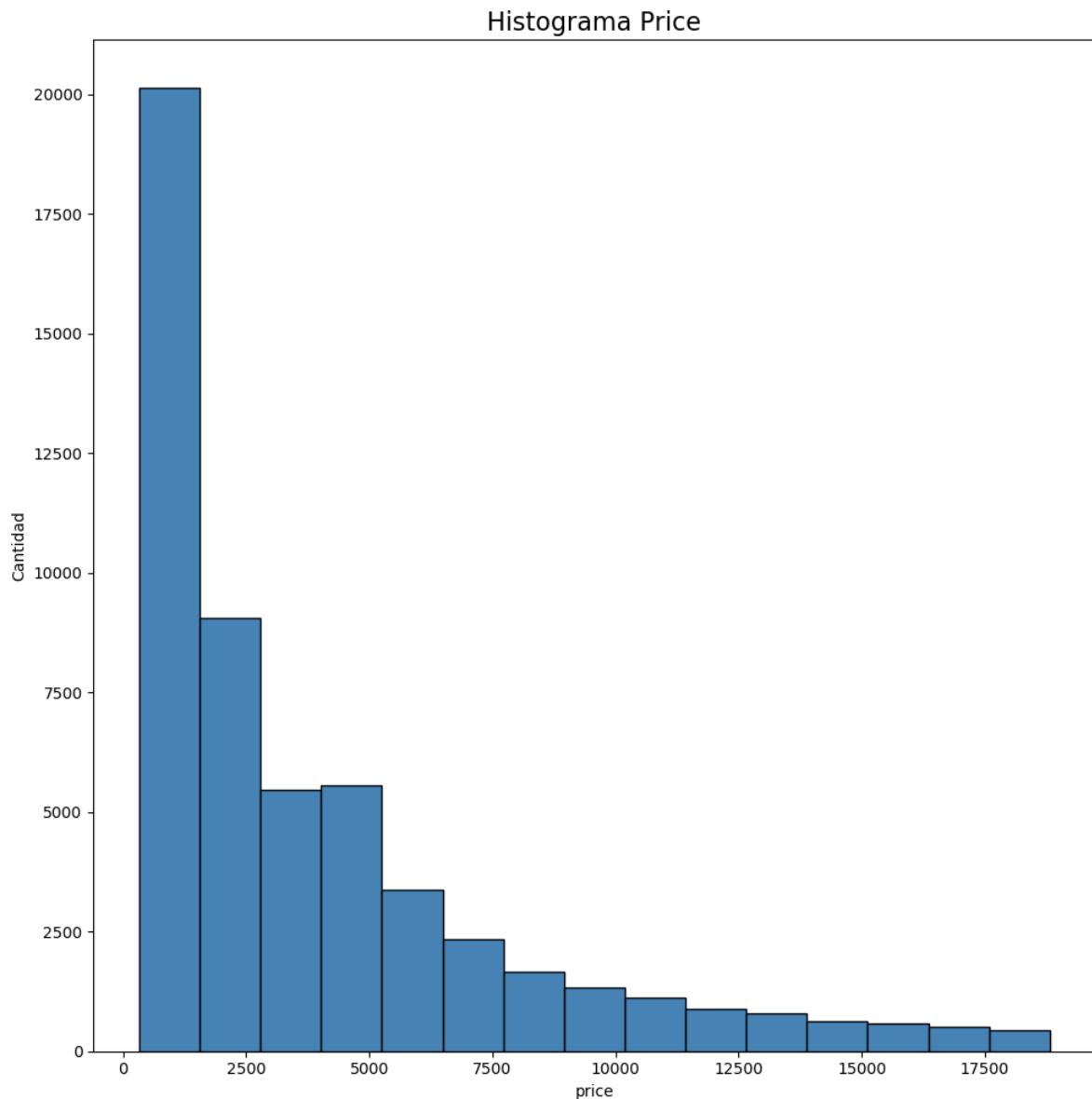


50%	(2.91, 3.52)
75%	(3.52, 4.03)
100%	(4.03, 5.6)

Observemos que hay casi una distribución uniforme entre los diamantes con height 3.0 y 4.0. Más allá de eso la altura de los mismos varía ampliamente desde su valor más chico (1.53) al más grande (5.6).



## Price



Media	3921.512
Desvío estándar	3975.384
Mínimo	326
Máximo	18823
25%	(326, 949)



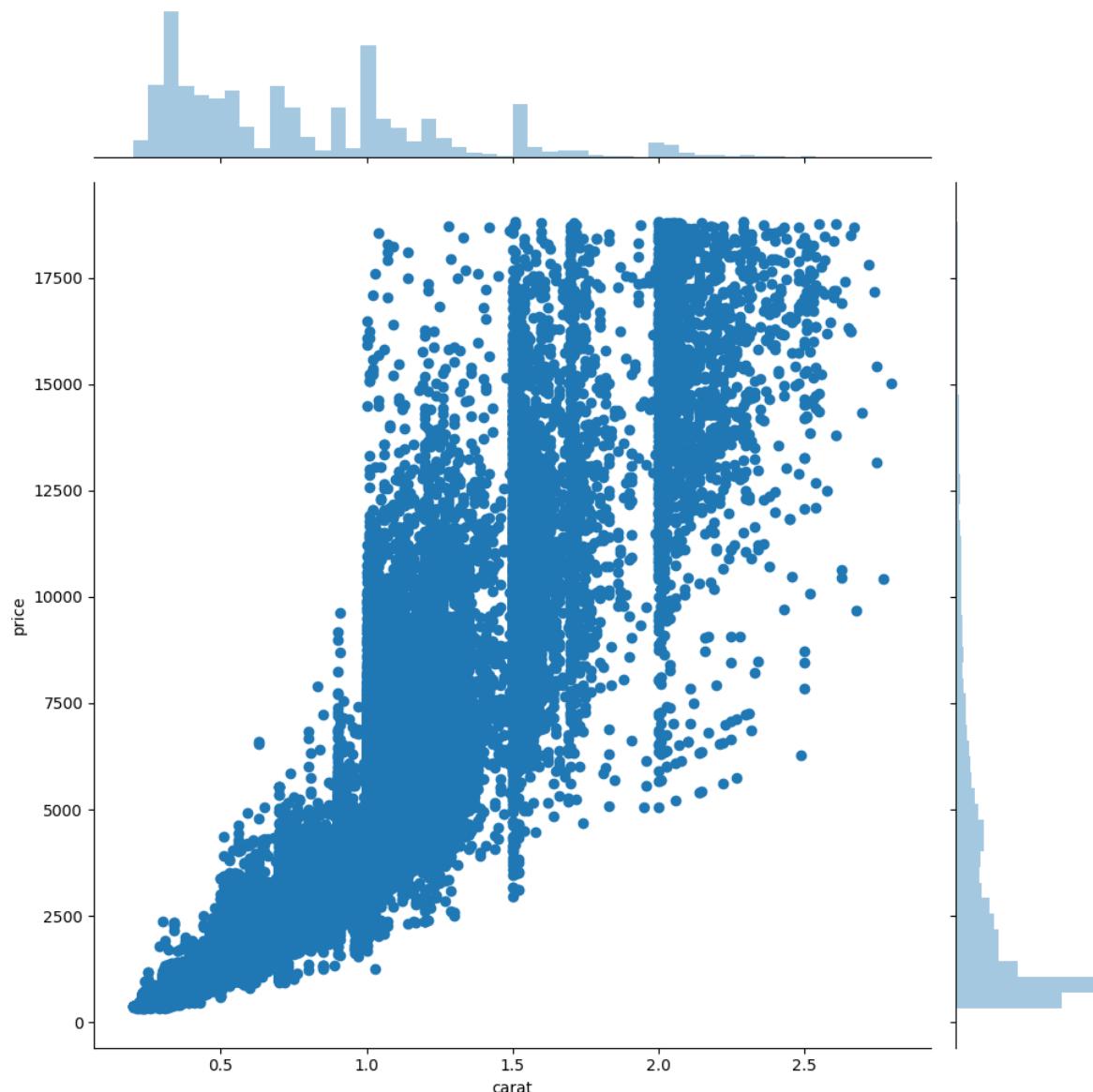
50%	(949, 2399)
75%	(2399, 5311)
100%	(5311, 18823)

Observemos que la amplitud de valores que toma el price es muy grande, siendo 326 el mínimo y 18823 el máximo, por esta razón es que el desvío estándar es más grande que la media. Además, la mayor parte de los registros son diamantes “baratos” menores a 6000 dólares.

### ***Distribución del precio con respecto a atributos numéricos***

A continuación graficamos las variables numéricas en el eje X y al precio en el eje Y y analizamos las relaciones entre estas variables.

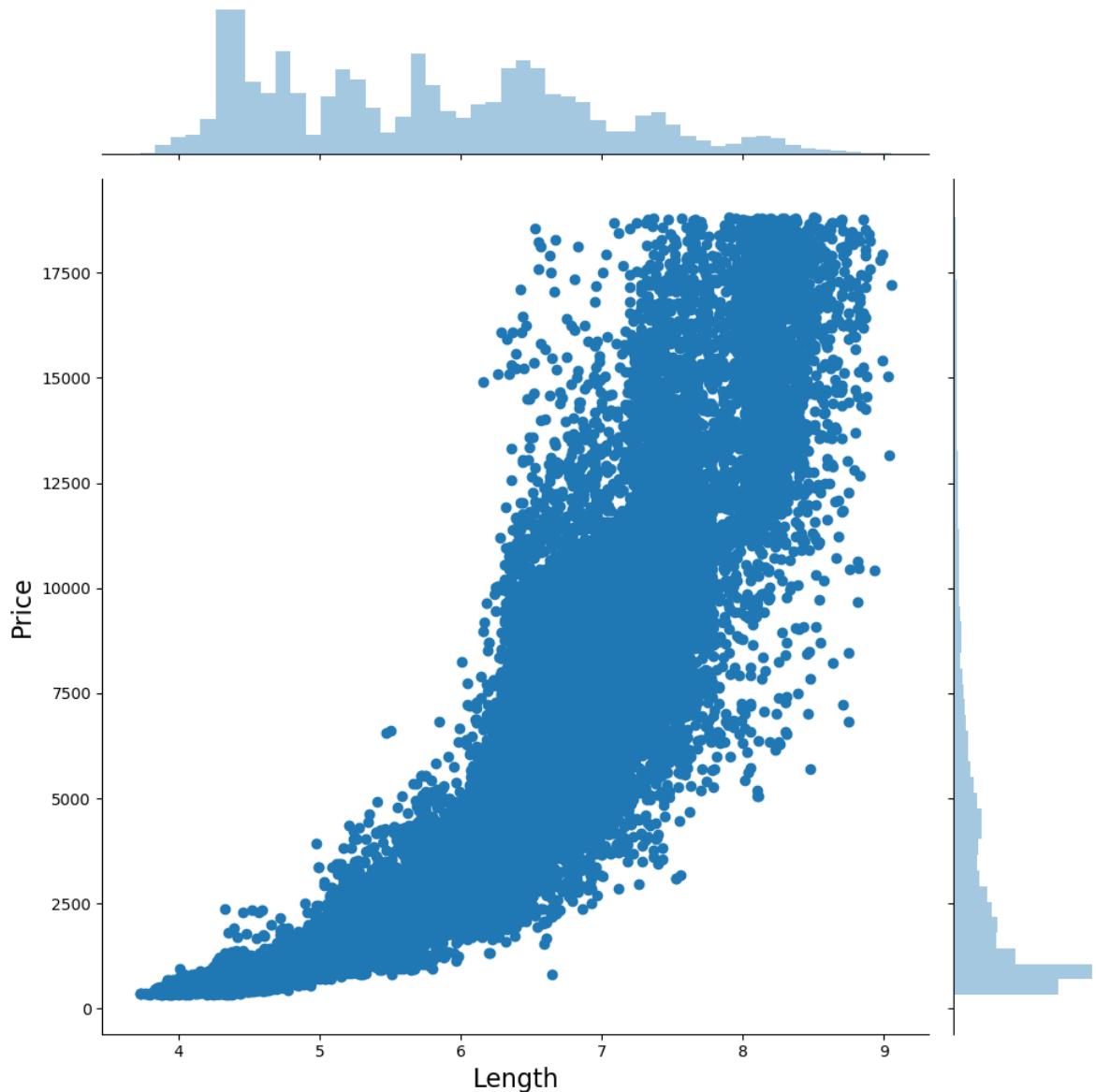
## Carat vs Price

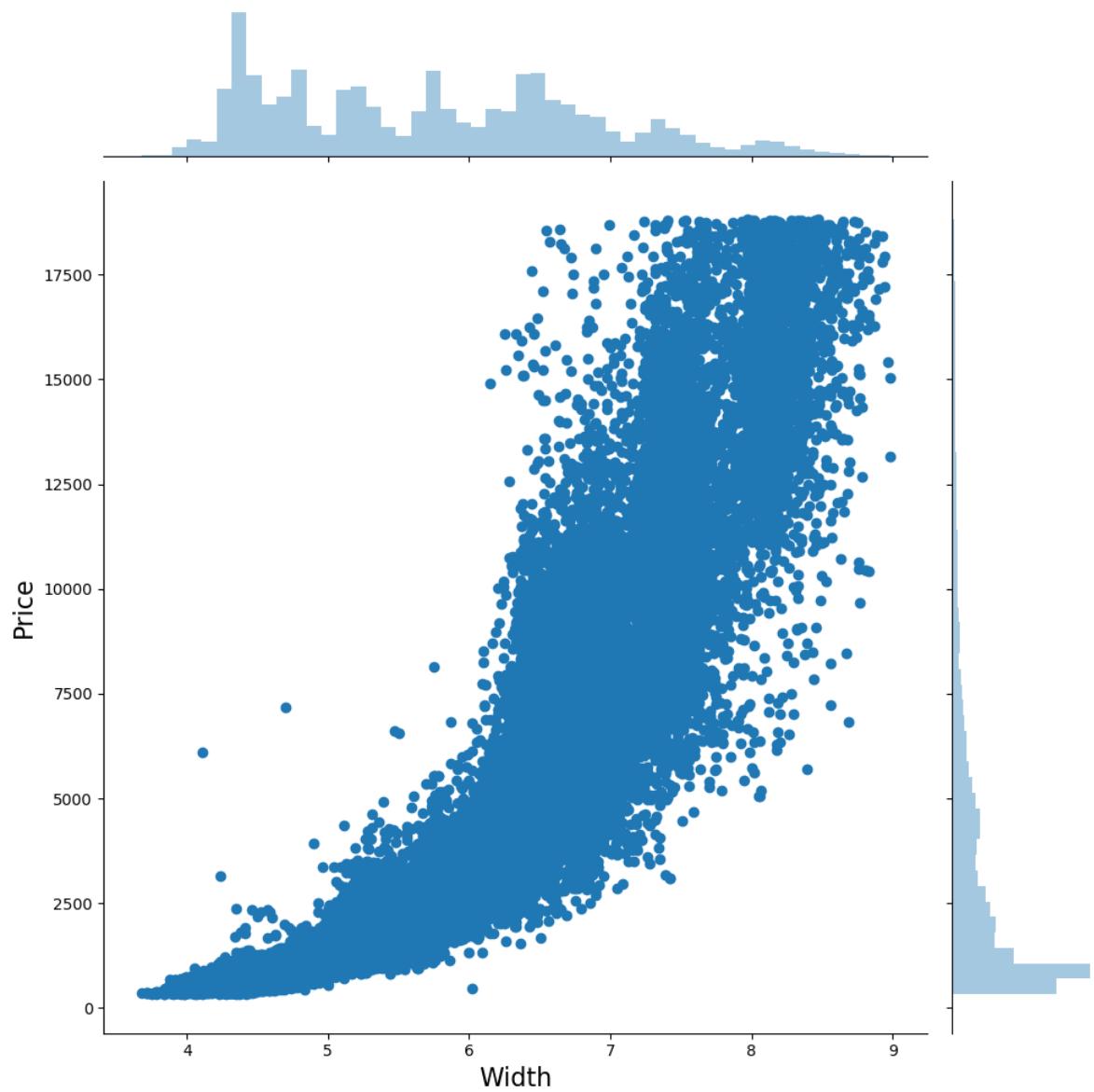


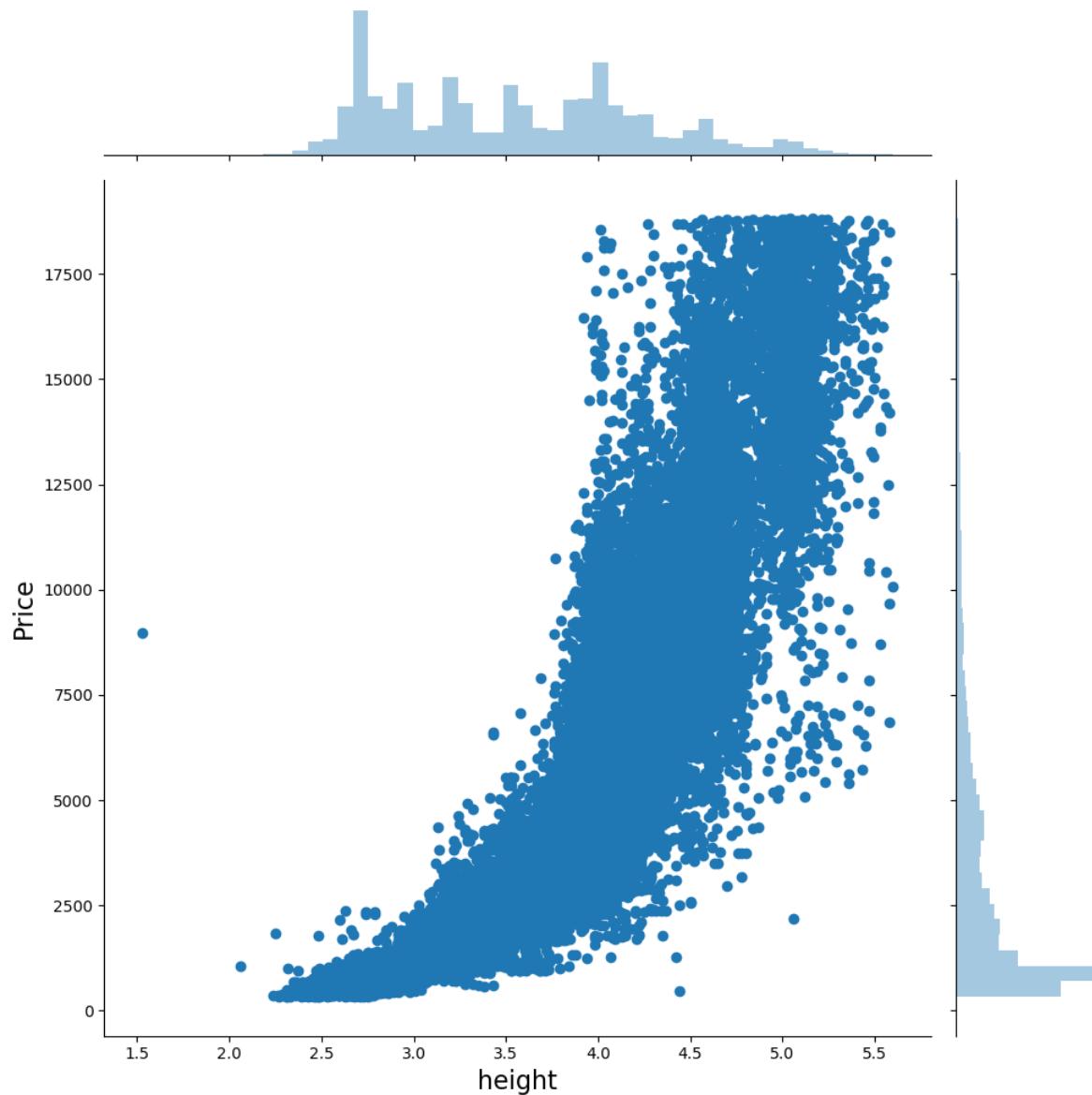
Notemos que a medida que el carat aumenta el price también. De hecho, parece ser un crecimiento exponencial. Esto tiene sentido, ya que, los diamantes de alto carat son muy raros, de modo que un diamante de por ejemplo 3 carats vale mucho más que dos diamantes de 1.5 carats cada uno. De esta manera, el carat es un muy buen candidato como atributo predictor, ya que, el precio muestra una correlación positiva.



## Length , Width, Height vs Price



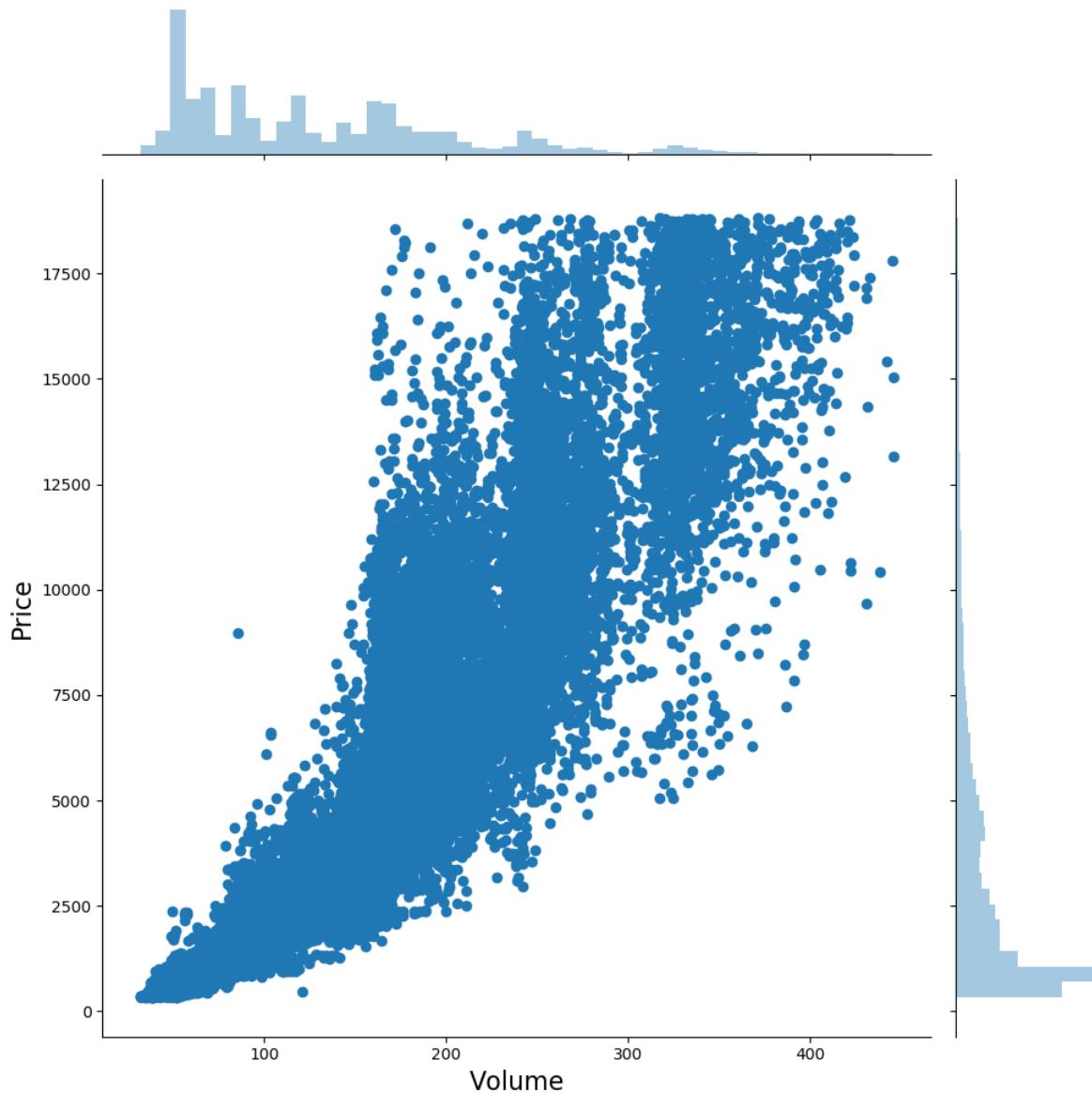




Podemos observar que los tres gráficos son muy similares. Esto se debe a que los atributos length, width y height están correlacionados e impactan de la misma manera al price. Notemos como el price crece a medida que las dimensiones del diamante (length, width, height) aumentan. Al igual que con el carat, el crecimiento es exponencial. También son buenos candidatos como atributos predictores, ya que, el precio muestra una correlación positiva.

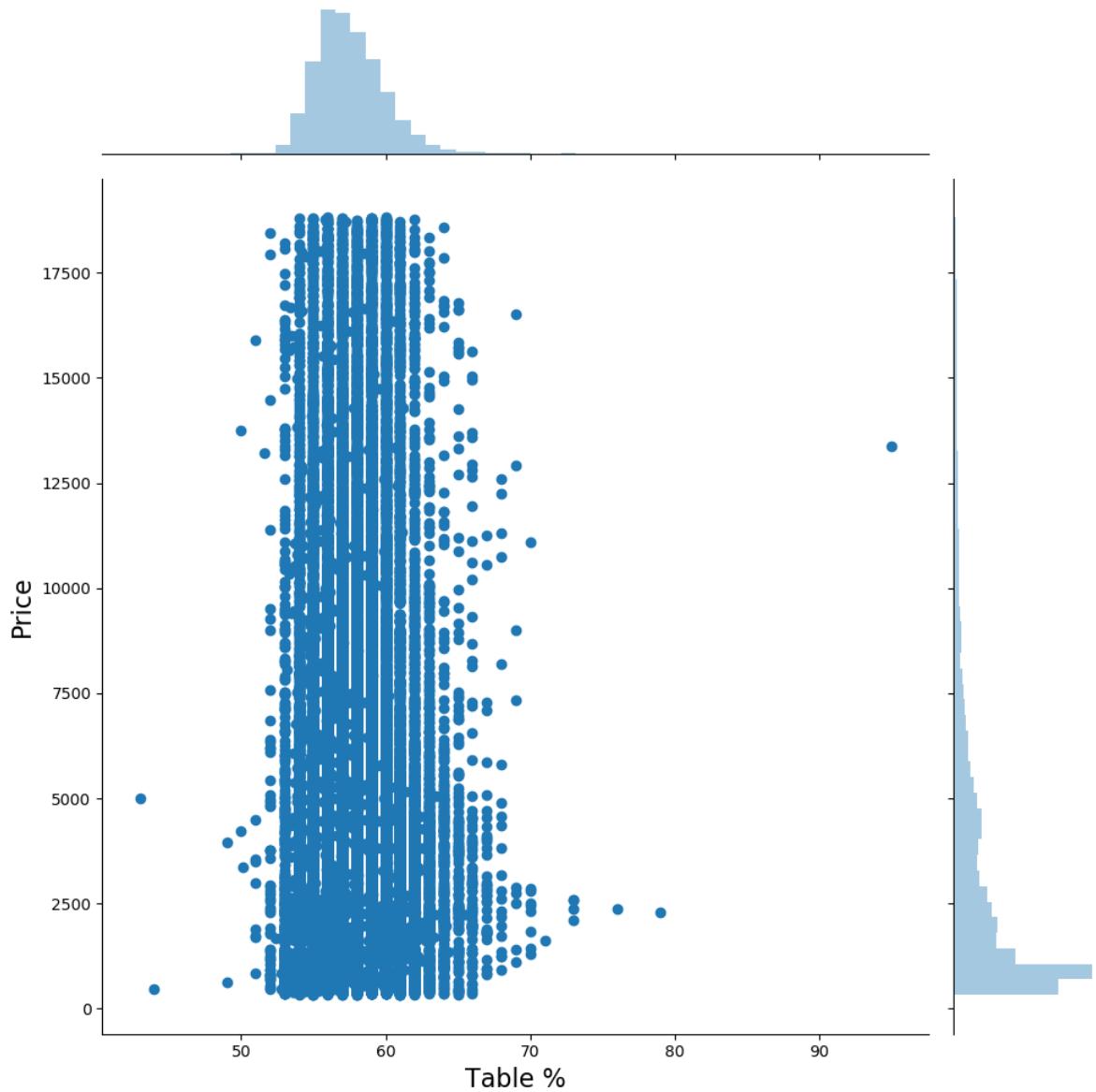
## Volume vs Price

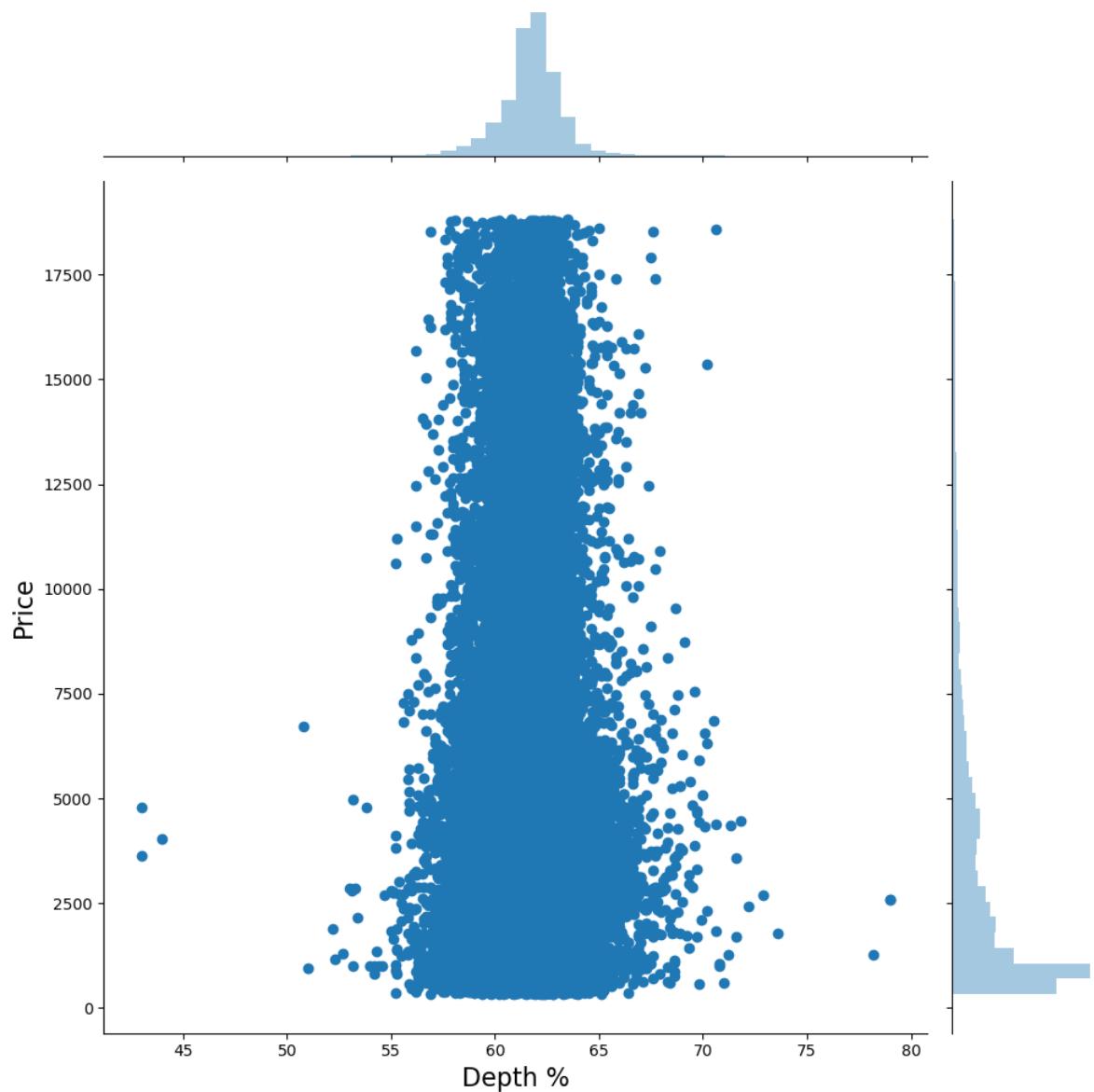
De los atributos Length, Width y Height podemos derivar el volumen del diamante. A continuación vemos cómo varía el precio según el volumen.



Podemos observar un crecimiento exponencial del precio. El crecimiento es similar que para los atributos Length, Width y Height por lo que podemos decir que el volumen resume bastante la información que nos brindan los atributos anteriores. Se podría utilizar el volumen como variable para reducir la complejidad del modelo.

## Table %, Depth % vs Price



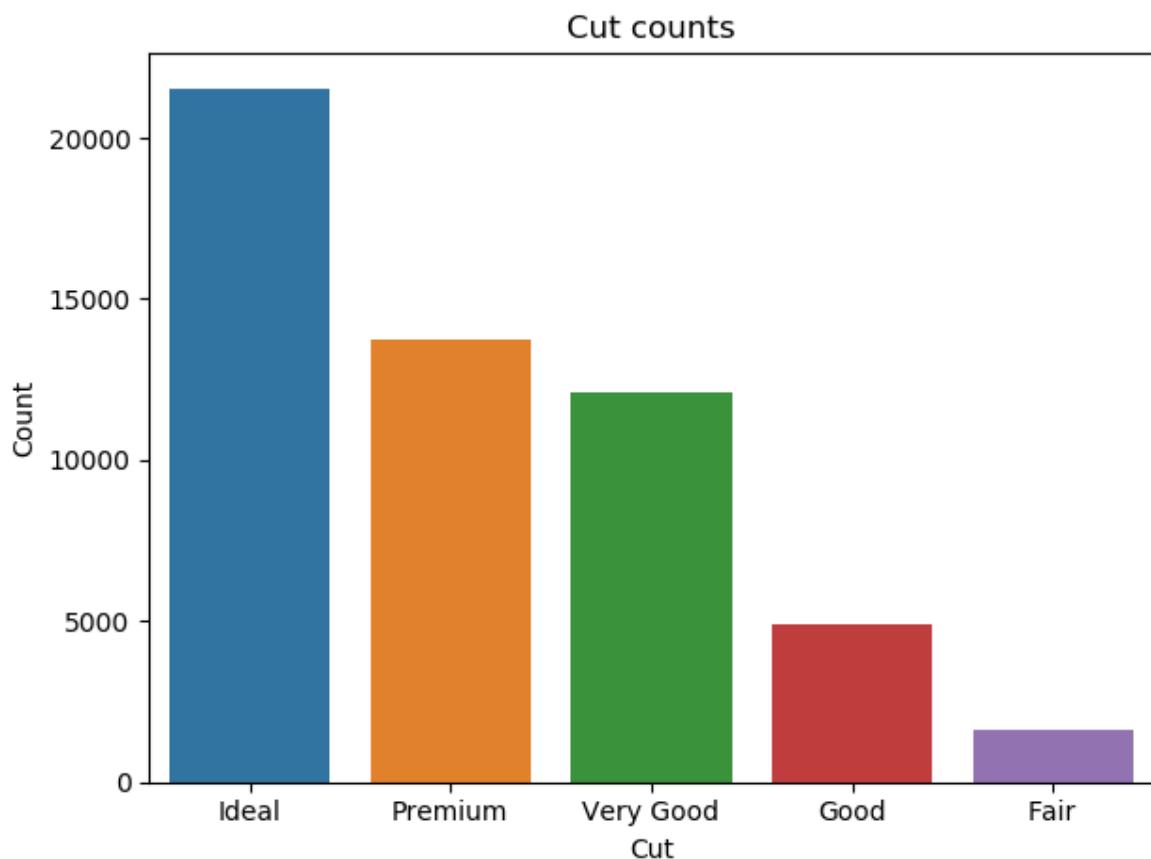


Podemos observar que tanto el table % como el depth % no son atributos buenos para predecir el precio del diamante. Esto se debe a que para un mismo table % o depth % el precio varía enormemente, desde un precio muy bajo a un precio altísimo y distribuido uniformemente. Por estas razones, es probable que estos atributos generen ruido si son utilizados como predictores. Lo mejor sería eliminarlos.



## Distribución atributos categóricos

### Cut

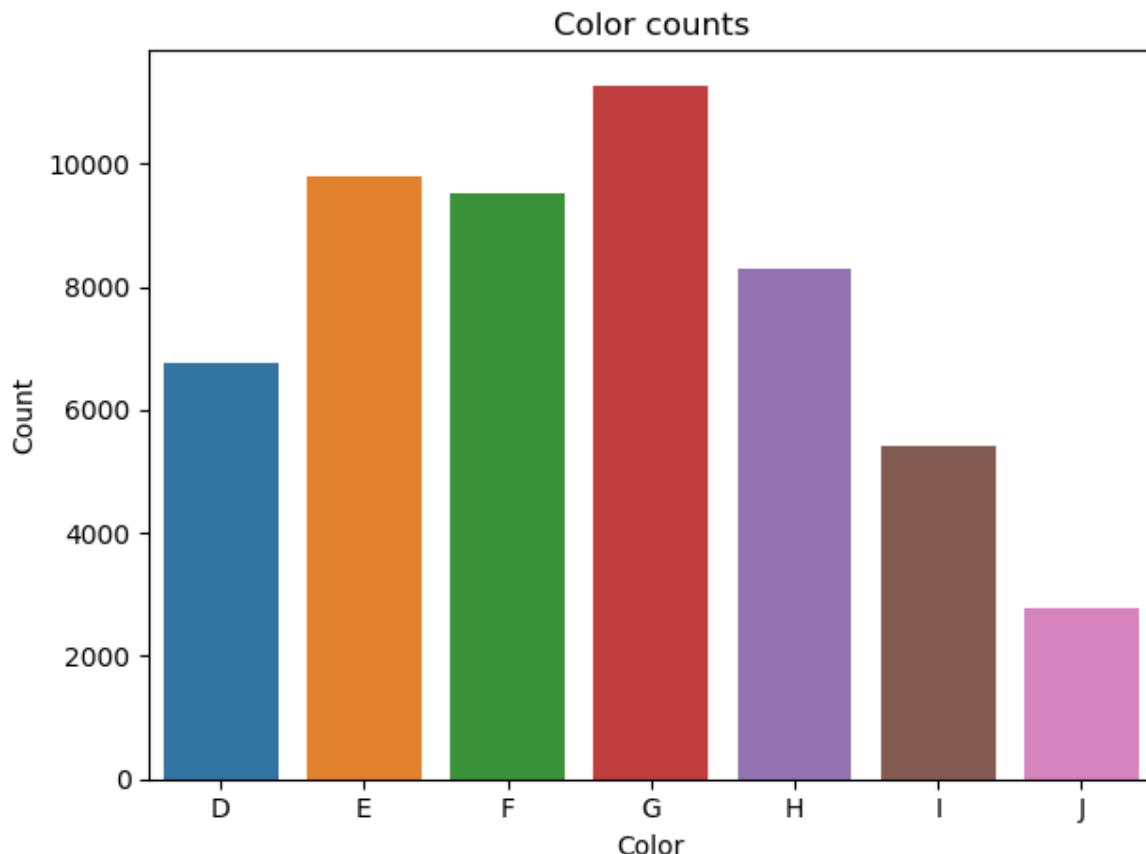


	Count	% of total
Ideal	21540	40%
Premium	13762	25%
Very Good	12075	22%
Good	4896	10%
Fair	1594	3%

Notemos que abundan más los diamantes con mejor cut, es decir, Ideal, Premium y Very Good, mientras que hay bastantes menos registros con cut Good o Fair.



## Color

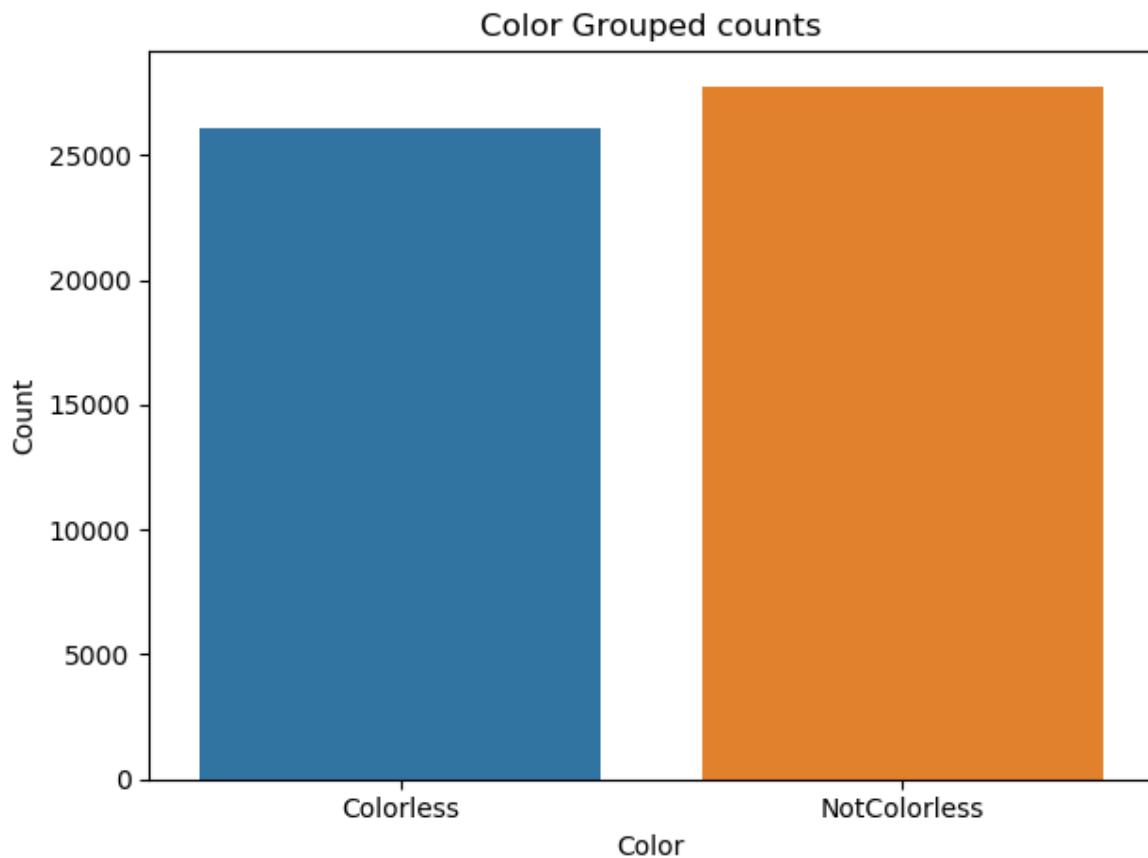


	Count	% of total
D	6772	13%
E	9792	18%
F	9536	18%
G	11280	21%
H	8286	15%
I	5405	10%
J	2796	5%

Podemos observar que hay una mayoría de registros con categoría G de color. A su vez, hay muy pocos diamantes con categoría J.



Dado que los valores D,E,F corresponden a diamantes colorless (es decir, transparentes) y los valores G,H,I,J a diamantes que no son transparentes, podemos agrupar en dos categorías este atributo. Entonces, el atributo color será Colorless o NotColorless. Abajo mostramos cómo se distribuye.

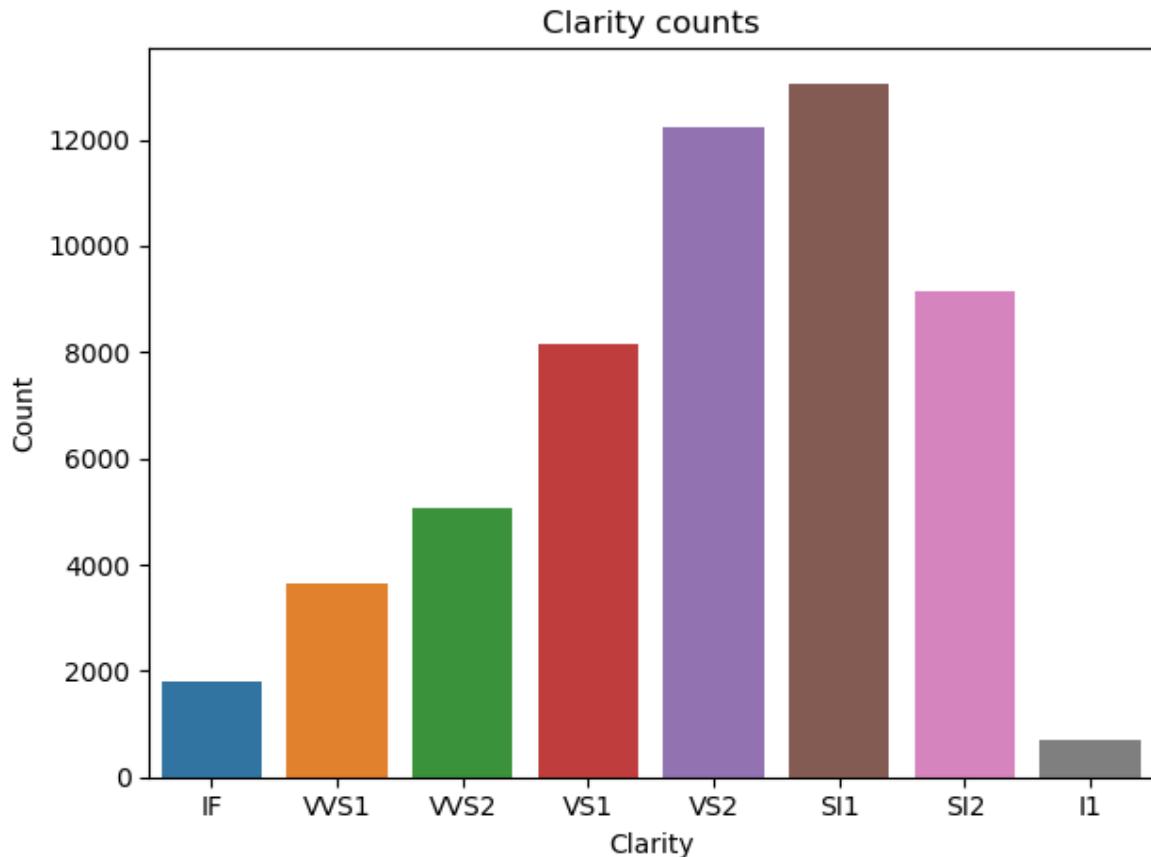


	Count	% of total
Colorless	27767	48%
NotColorless	26100	52%

Notemos que hay ligeramente mayor cantidad de diamantes NotColorless, es decir, que no son transparentes.



## Clarity

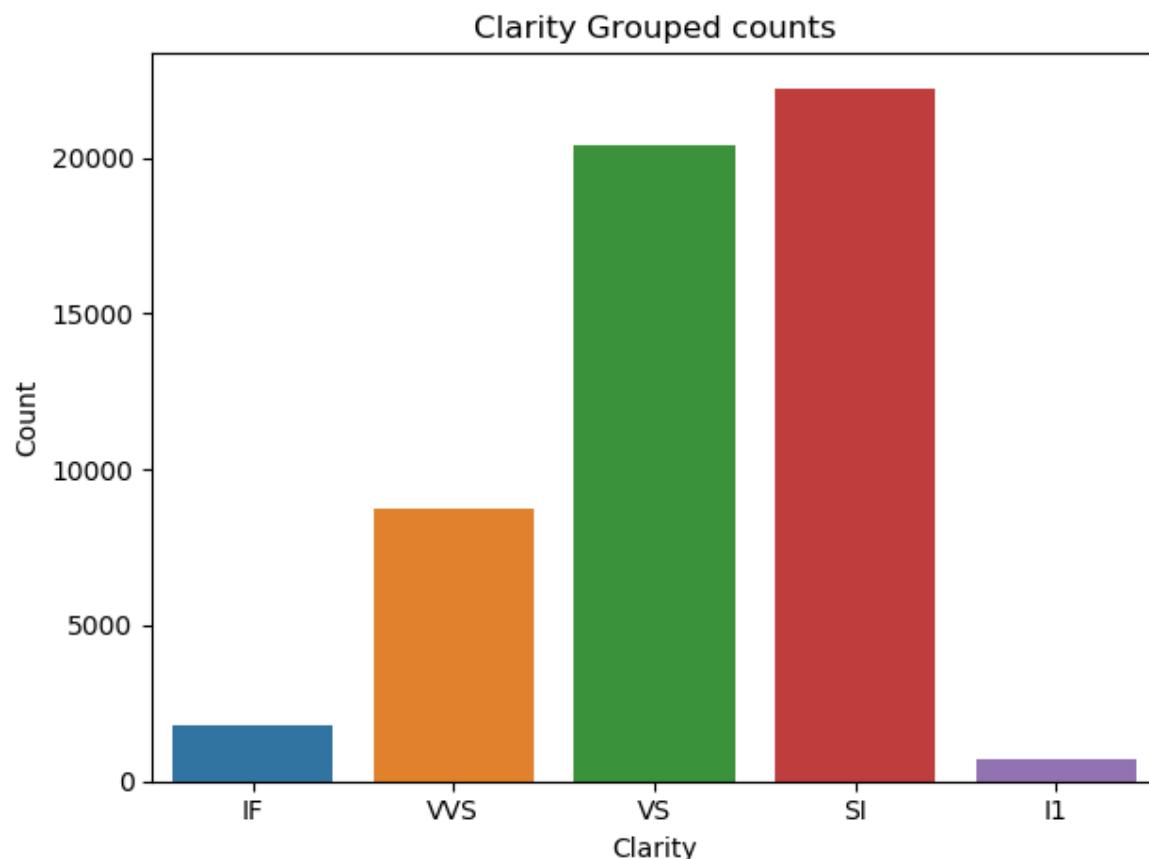


	Count	% of total
IF	1790	3%
VVS1	3654	7%
VVS2	5066	9%
VS1	8167	15%
VS2	12253	23%
SI1	13062	24%
SI2	9165	17%
I1	710	1%

Podemos ver como hay una gran cantidad de diamantes con una clarity de calidad media (VS1, VS2, SI1, SI2). Por otro lado, hay en menor medida registros con clarity más alta (IF, VVS1, VVS2) y muy pocos registros con clarity baja como es I1.

Dado que los valores VVS1-VVS2, VS1-VS2, SI1-SI2 son conceptualmente muy similares, agrupamos los valores en las siguientes categorías VVS, VS Y SI. Los valores IF y I1 quedan igual.

A continuación mostramos la distribución del agrupamiento:



	Count	% of total
IF	1790	3%
VVS	8720	16%
VS	20420	39%
SI	22227	41%



I1	710	1%
----	-----	----

Como podemos observar, la mayoría de los diamantes tienen un clarity de tipo VS o SI. Además, hay muy pocos diamantes en los extremos (IF y I1).

## **Verificación de la calidad de los datos**

A continuación se listan los problemas de calidad de los datos encontrados y las posibles soluciones para cada uno.

### **Atributos innecesarios**

Existen atributos como el identificador de cada registro que no aportan información sobre la clase a predecir.

Las posibles soluciones son:

- Eliminar el atributo identificador.

### **Atributos ruidosos**

Existen atributos como table % y depth % que parecen no tener relevancia en el precio del diamante. Esto se ve evidenciado en el análisis de sus distribuciones. Mantener estos atributos puede generar una complejidad innecesaria y baja en la performance del modelo.

Las posibles soluciones son:

- Eliminar los atributos ruidosos.

### **Registros duplicados**

Un registro se considera duplicado si hay algún otro registro cuyos atributos son idénticos. Este dataset contiene 146 registros duplicados. Los registros duplicados en este set de datos son redundantes y pueden generar ruido.

Las posibles soluciones son:

- Eliminar los registros duplicados.

### **Registros con valores inválidos**

Algunos registros contienen atributos con valores inválidos. Esto sucede con los atributos Length, Width y Height que representan el largo, ancho y alto del diamante respectivamente. En estos registros el valor de Length, Width y Height es igual a 0. Esto es claramente un error, dado que estos atributos representan dimensiones del diamante. Son 20 los registros que presentan este problema.

Las posibles soluciones son:



- Eliminar los registros donde el atributo Length, Width o Height sea 0.
- Reemplazar el valor del atributo Length, Width o Height donde sea 0 por la media de Length, Width o Height respectivamente.
- Reemplazar el valor del atributo Length, Width o Height donde sea 0 por la mediana de Length, Width o Height respectivamente.

### Registros con valores atípicos

Hay registros que tienen valores atípicos comparados con el resto. Para las variables numéricas con distribución normal se consideran outlier valores que se encuentran fuera del intervalo media +/- 3 desvíos standard. Estos registros pueden dificultar la discretización de las variables numéricas y reducir la eficacia del modelo. Se encontraron outliers en los atributos Length, Width y Height totalizando 51 registros.

Las posibles soluciones son:

- Eliminar los registros que contengan valores atípicos.
- Reemplazar los valores atípicos por la media de la variable.
- Reemplazar los valores atípicos por la mediana de la variable.



## 3- Preparación de los datos

### Selección de los datos

#### Inclusión/Exclusión de datos

Se ha decidido incluir todos los registros del dataset en cuestión, ya que la cantidad de registros no presenta un problema para su correcto análisis (sujetos a los criterios de depuración de datos y atributos detallados en la sección “limpieza de datos”, que detalla las decisiones tomadas y sus respectivas justificaciones).

### Limpieza de los datos

#### Reporte de limpieza de datos

Problema	Soluciones posibles	Decisión tomada	Justificación
Atributos innecesarios	<ul style="list-style-type: none"><li>• Eliminar el atributo ID</li></ul>	Eliminar el atributo ID	<b>El atributo ID no tiene ningún valor para el análisis en cuestión.</b>
Atributos ruidosos	<ul style="list-style-type: none"><li>• Eliminar los atributos ruidosos</li></ul>	Eliminar los atributos ruidosos (table % y depth %)	<b>En el análisis de comprensión de los datos se ha visto que dichos atributos parecen no tener relevancia en la determinación del precio del diamante, por lo tanto se ha decidido eliminarlos para no agregar complejidad innecesaria al modelo.</b>
Registros	<ul style="list-style-type: none"><li>• Eliminar los</li></ul>	Eliminar los	<b>Los registros</b>



duplicados	registros duplicados	registros duplicados	<b>duplicados no agregan ningún valor al análisis que se está realizando, por lo tanto se ha decidido eliminarlos.</b>
Registros con valores inválidos	<ul style="list-style-type: none"> <li>• Eliminar los registros donde el atributo Length, Width o Height sea 0.</li> <li>• Reemplazar el valor del atributo Length, Width o Height donde sea 0 por la media de Length, Width y Height respectivamente.</li> <li>• Reemplazar el valor del atributo Length, Width o Height donde sea 0 por la mediana de Length, Width y Height respectivamente.</li> </ul>	Eliminar los registros donde el atributo Length, Width o Height sea 0.	<b>Menos del 0.03% del total de registros presentan valores inválidos, al ser dicha cantidad despreciable se ha decidido eliminarlos</b>
Registros con valores atípicos	<ul style="list-style-type: none"> <li>• Eliminar los registros que contengan valores atípicos.</li> <li>• Reemplazar los valores atípicos por la media de la variable.</li> </ul>	Eliminar los registros que contengan valores atípicos.	<b>Menos del 0.09% del total de registros presentan valores atípicos, al ser dicha cantidad despreciable se ha decidido eliminarlos</b>



	<ul style="list-style-type: none"> <li>• Reemplazar los valores atípicos por la mediana de la variable.</li> </ul>		
--	--	--	--

## Estructura de los datos

### Derivación de atributos

Atributo derivado	Derivado de	Método aplicado
Volume	<ul style="list-style-type: none"> <li>• Length</li> <li>• Width</li> <li>• Height</li> </ul>	$\text{Volume} = \text{Length} * \text{Width} * \text{Height}$

### Discretización de atributos numéricos

Para utilizar las herramientas y disminuir la complejidad del modelo y su análisis se ha decidido discretizar las variables numéricas. La técnica de discretización elegida fue la de partir en 4 intervalos cada variable numérica. Para ello, se utilizaron los quartiles. A cada intervalo se lo categorizó con un nombre.

Atributo	Discretización
Carat	<ul style="list-style-type: none"> <li>• <b>Low</b> (&lt; 0.4)</li> <li>• <b>Medium</b> (0.4-0.7)</li> <li>• <b>High</b> (0.7-1.04)</li> <li>• <b>Very High</b> (&gt; 1.04)</li> </ul>
Volume	<ul style="list-style-type: none"> <li>• <b>Low</b> (&lt; 31.71)</li> <li>• <b>Medium</b> (31.71-65.34)</li> <li>• <b>High</b> (65.34-114.93)</li> <li>• <b>Very High</b> (&gt; 114.93)</li> </ul>
Price	<ul style="list-style-type: none"> <li>• <b>Low</b> (&lt; 954)</li> <li>• <b>Medium</b> (954-2409)</li> </ul>



	<ul style="list-style-type: none"><li>• <b>High</b> (2409-5324)</li><li>• <b>Very High</b> (&gt; 5324)</li></ul>
--	--

## *Transformación de valores de atributos*

Atributo	Valores actuales	Valores nuevos
Color	<ul style="list-style-type: none"><li>• D</li><li>• E</li><li>• F</li><li>• G</li><li>• H</li><li>• I</li><li>• J</li></ul>	<ul style="list-style-type: none"><li>• <b>Colorless</b> (D, E, F)</li><li>• <b>Not Colorless</b> (G, H, I, J)</li></ul>
Clarity	<ul style="list-style-type: none"><li>• IF</li><li>• VVS1</li><li>• VVS2</li><li>• VS1</li><li>• VS2</li><li>• SI1</li><li>• SI2</li><li>• I1</li></ul>	<ul style="list-style-type: none"><li>• <b>IF</b></li><li>• <b>VVS</b> (VVS1, VVS2)</li><li>• <b>VS</b> (VS1, VS2)</li><li>• <b>SI</b> (SI1, SI2)</li><li>• <b>I1</b></li></ul>

## *Integración de los datos*

No fue necesario realizar un proceso de integración de los datos.

## *Formateo de los datos*



Herramienta	Cambios sobre los datos
Weka	Cambiar la posición del atributo clase para que sea el último atributo de la fila

## Data Set preparado

Link al dataset depurado

<https://drive.google.com/open?id=1V40-FOjECJU35XIHL4fUtJpHXWxUJaEQ>

Código utilizado para la depuración:

```
import pandas as pd
import csv
from sklearn.utils import shuffle

def createVolume(dataFrame):
    exportDataframe(dataFrame, "temp.csv")
    file = open("temp.csv")
    reader = csv.reader(file, delimiter=',')
    columns = next(reader)
    xIndex = 0
    yIndex = 0
    zIndex = 0
    for x in range(len(columns)):
        if columns[x] == 'x':
            xIndex = x
        if columns[x] == 'y':
            yIndex = x
        if columns[x] == 'z':
            zIndex = x
    output = open("temp2.csv", "w")
    columns.append("volume")
    output.write(",".join(columns)+"\n")
    for row in reader:
        volume = str(round(float(row[xIndex]) * float(row[yIndex]) *
float(row[zIndex]), 2))
        row.append(volume)
```



```
        output.write(",".join(row)+"\n")
output.close()
file.close()
dataFrame = pd.read_csv('temp2.csv')
return dataFrame.drop('x', axis=1).drop('y', axis=1).drop('z', axis=1)

def printUnique(dataFrame):
    for col in dataFrame:
        print(dataFrame[col].unique())

def printColumns(dataFrame):
    print(dataFrame.columns)

def groupColumnsDataframe(dataFrame):
    colorReplace = {"D": "Colorless", "E": "Colorless", "F": "Colorless",
"G": "Not Colorless", "H": "Not Colorless", "I": "Not Colorless", "J": "Not
Colorless"}
    dataFrame["color"] = dataFrame["color"].replace(colorReplace)

    clarityReplace = {"IF": "IF", "VVS1": "VVS", "VVS2": "VVS", "VS1": "VS",
"VS2": "VS", "SI1": "SI", "SI2": "SI",
        "I1": "I1"}
    dataFrame["clarity"] = dataFrame["clarity"].replace(clarityReplace)

    return dataFrame

def dropUnnecessaryColumns(dataFrame):
    return dataFrame.drop('table', axis=1).drop('depth',
axis=1).drop('Unnamed: 0', axis=1)

def exportDataframe(dataFrame, name):
    dataFrame.to_csv(name, sep=",", encoding="utf-8", index=False)

def countXYZWithZeroValue(dataFrame):
    return len(dataFrame[(dataFrame['x'] == 0) | (dataFrame['y'] == 0) |
(dataFrame['z'] == 0)])]

def eliminateXYZWithZeroValue(dataFrame):
    return dataFrame[(dataFrame[['x', 'y', 'z']] != 0).all(axis=1)]]

def eliminateOutliers(dataFrame, includeDepthAndTable=False):
    columns = ["x", "y", "z"]
    if includeDepthAndTable:
        columns.append("depth")
```



```
    columns.append("table")
rowIndeces = []
for column in columns:
    mean = dataFrame[column].mean()
    std = dataFrame[column].std()
    lower = mean - 3 * std
    upper = mean + 3 * std
    indeces = dataFrame[(dataFrame[column] < lower) | (dataFrame[column]
> upper)].index
    for index in indeces:
        if index in rowIndeces:
            continue
        rowIndeces.append(index)
dataFrame = dataFrame.drop(rowIndeces)
return dataFrame

def eliminateDuplicates(dataFrame):
    return dataFrame.drop_duplicates()

def main():
    dataFrame = pd.read_csv("diamonds.csv")

    dataFrame = dropUnnecessaryColumns(dataFrame)
    dataFrame = eliminateDuplicates(dataFrame)
    dataFrame = eliminateXYZWithZeroValue(dataFrame)
    dataFrame = eliminateOutliers(dataFrame)
    dataFrame = createVolume(dataFrame)
    dataFrame = groupColumnsDataframe(dataFrame)

    # Discretize numerical variables
    dataFrame["carat"], bins = pd.qcut(dataFrame[ "carat"], 4, labels=[ "Low",
"Medium", "High", "Very High"], retbins=True)
    dataFrame[ "volume"], bins = pd.qcut(dataFrame[ "volume"], 4,
labels=[ "Low", "Medium", "High", "Very High"], retbins=True)
    dataFrame[ "price"], bins = pd.qcut(dataFrame[ "price"], 4, labels=[ "Low",
"Medium", "High", "Very High"], retbins=True)
    # Shuffle dataframe
    dataFrame = shuffle(dataFrame)

    # Change order columns
    dataFrame = dataFrame.reindex(columns=[ 'carat', 'cut', 'color',
'clarity', 'volume', 'price'])

    fileName = "diamonds_processed.csv"
```



```
exportDataframe(dataFrame, fileName)  
  
main()
```

El mismo fue realizado en Python 3 utilizando librerías de preprocesamiento como numpy, pandas y csv.



## 4- Modelado

### **Técnica de modelado**

Las técnicas de modelado que se van a utilizar son:

#### **Algoritmo de Inducción (C4.5)**

Los algoritmos de inducción se utilizan en la minería de datos para modelar las clasificaciones en los datos mediante árboles de decisión. Pertenecen a los métodos inductivos del Aprendizaje Automático que aprenden a partir de ejemplos preclasificados.

#### **Funcionamiento**

El algoritmo genera un árbol de decisión a partir de datos ya clasificados, en el que en cada nodo del árbol representa un atributo y cada arista un valor posible de cada atributo. El árbol de decisión generado se evalúa de arriba hacia abajo (**Top-Down Induction Tree**), donde cada nodo (atributo) implica la selección de un valor para dicho atributo, moviéndose hacia abajo en el árbol en cada selección, hasta llegar al último nivel donde se encontrará el valor de clase, clasificando así al registro.

El orden de los nodos no es arbitrario, para construir el árbol, el algoritmo calcula la **ganancia (disminución en entropía)** de cada atributo, seleccionando al de mayor ganancia para realizar la división en cada paso. El uso de este valor (ganancia) como determinante para decidir el atributo por el cual dividir se justifica con el concepto de entropía (que mide la incertidumbre que se tiene sobre los datos), que según la teoría de la información, la información se maximiza cuando la entropía se minimiza.

#### **Red bayesiana (Naive Bayes)**

Los algoritmos de Bayes se utilizan para resolver problemas de clasificación supervisada. Una **red bayesiana** es un **grafo acíclico** (lineal) dirigido, donde cada nodo representa una variable y cada arco una dependencia probabilística donde se especifica la probabilidad de cada variable dados sus padres. La variable a la que apunta el arco es independiente (causa – efecto) de la que está en el origen.



## Funcionamiento

El algoritmo genera la red bayesiana donde se pueden apreciar las dependencias probabilísticas de los atributos del modelo para inferir qué clase es más probable dado determinado valor de los atributos.

## Supuestos de modelado

Técnica de modelado	Supuestos y Precondiciones
Algoritmo de inducción	<ul style="list-style-type: none"><li>Los datos se encuentran clasificados</li><li>La clase a predecir debe ser discreta</li></ul>
Red Bayesiana	<ul style="list-style-type: none"><li>Los datos se encuentran clasificados</li><li>La clase a predecir debe ser discreta</li><li>Todos los atributos deben ser discretos</li></ul>

## Diseño de pruebas

Para verificar la calidad del modelo se separará el set de datos en dos sets, uno de entrenamiento y otro de prueba. Esto se hace dividiendo al azar el set de datos.

Set de entrenamiento: contiene el 70% de los datos. Se utiliza para entrenar el modelo.  
Set de prueba: contiene el 30% de los datos. Se utiliza para validar la precisión del modelo.

La métrica utilizada como forma de validación de la calidad del modelo es la precisión. Esta se calcula como casos correctamente clasificados / casos totales. Dado que la clase “precio” toma 4 valores posibles (Low, Medium, High y Very High) se pueden calcular una métrica de precisión para cada uno de los valores posibles.

Cabe aclarar que, dado que la clase está distribuida uniformemente, si clasificamos al azar o por valor de clase mayoritaria obtenemos un 25% de precisión. Este valor lo



podemos utilizar como base para verificar que nuestro modelo sea mejor que un modelo que clasifique al azar.

## Modelo

### Algoritmo de inducción

Para la construcción del árbol de decisión (J48) se utilizó la herramienta Weka.

La configuración del modelo final utilizada es:

Input

- Carat
- Cut
- Color
- Clarity
- Volume

Target

- Price

batchSize	100
binarySplits	False
collapseTree	False
confidenceFactor	0.25
debug	False
doNotCheckCapabilities	False
oNotMakeSplitPointActualValue	False
minNumObj	250
numDecimalPlaces	2
numFolds	3
reducedErrorPruning	False
saveInstanceData	False
seed	1
subtreeRaising	True
unpruned	False



Luego de la ejecución del algoritmo obtenemos el siguiente resultado:

```
==== Run information ====

Scheme:      weka.classifiers.trees.J48 -O -C 0.25 -M 250
Relation:    diamonds_processed
Instances:   53574
Attributes:  6
              carat
              cut
              color
              clarity
              volume
              price
Test mode:   split 70.0% train, remainder test

==== Classifier model (full training set) ====

J48 pruned tree
-----
volume = Medium
|   carat = Medium: Medium (12073.0/2587.0)
|   carat = High: Medium (463.0/210.0)
|   carat = Very High: Very High (3.0/1.0)
|   carat = Low
|     |   clarity = VVS: Medium (258.0/51.0)
|     |   clarity = SI: Low (248.0/12.0)
|     |   clarity = VS: Low (307.0/138.0)
|     |   clarity = IF: Medium (46.0/4.0)
|     |   clarity = I1: Low (0.0)
volume = High
|   clarity = VVS
|     |   color = Not Colorless: High (637.0/234.0)
|     |   color = Colorless: Very High (497.0/196.0)
|   clarity = SI
|     |   carat = Medium: Medium (270.0/109.0)
|     |   carat = High: High (6964.0/1207.0)
|     |   carat = Very High: High (112.0/12.0)
```



```
|   |   carat = Low: Low (1.0)
|   clarity = VS: High (4526.0/1747.0)
|   clarity = IF: Very High (153.0/66.0)
|   clarity = I1: Medium (230.0/110.0)
volume = Very High
|   clarity = VVS: Very High (1042.0/5.0)
|   clarity = SI: Very High (7172.0/1969.0)
|   clarity = VS: Very High (4624.0/295.0)
|   clarity = IF: Very High (181.0)
|   clarity = I1: High (370.0/163.0)
volume = Low: Low (13397.0/1858.0)
```

Number of Leaves : 23

Size of the tree : 30

Time taken to build model: 0.06 seconds

==== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

==== Summary ===

Correctly Classified Instances	12699	79.0132 %
Incorrectly Classified Instances	3373	20.9868 %
Kappa statistic	0.7202	
Mean absolute error	0.1612	
Root mean squared error	0.2839	
Relative absolute error	42.9827 %	
Root relative squared error	65.5527 %	
Total Number of Instances	16072	

==== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure
MCC	ROC Area	PRC Area	Class		

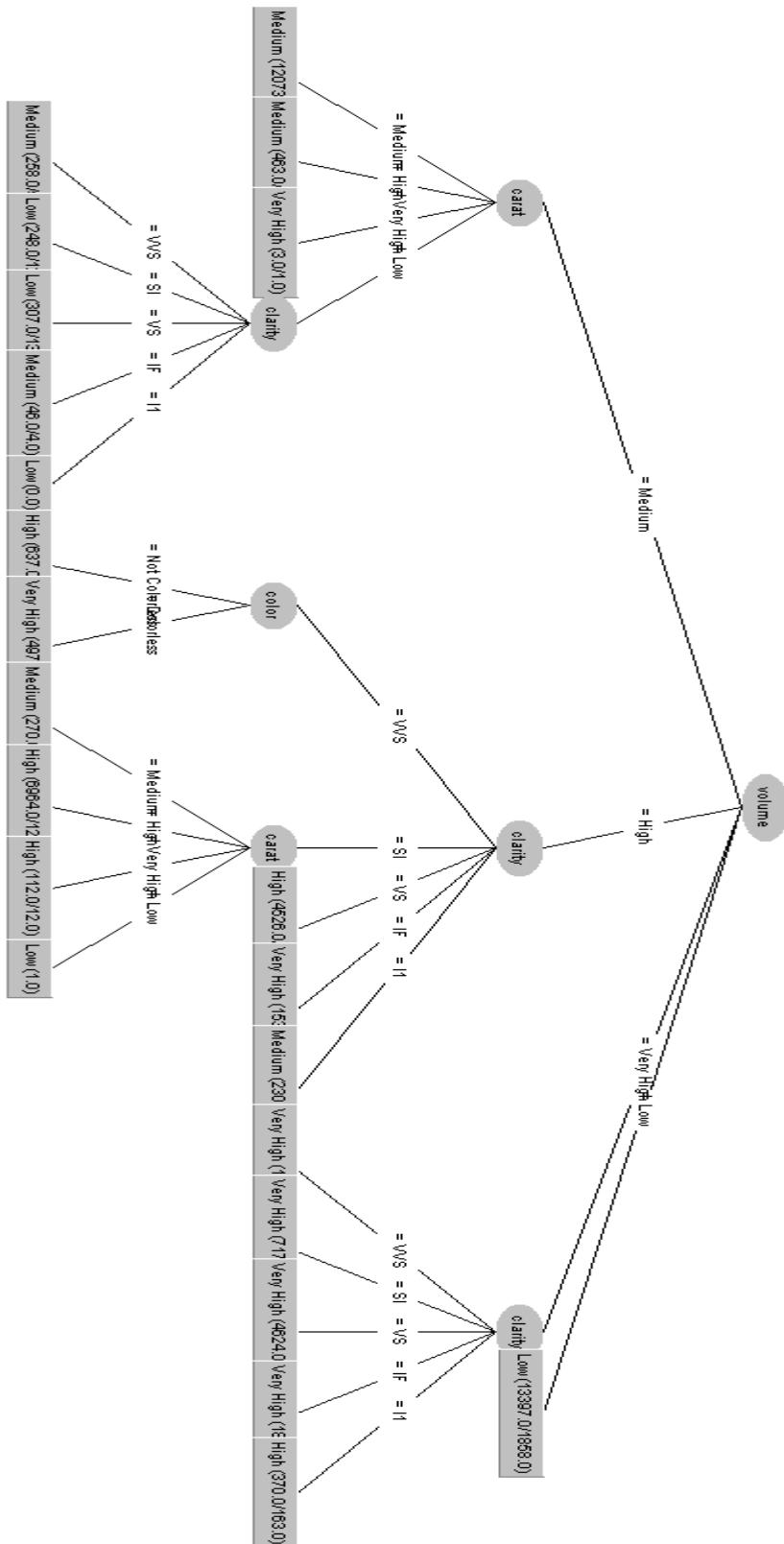


		0,730	0,072	0,772	0,730	0,750
0,671	0,896	0,699		Medium		
		0,698	0,086	0,732	0,698	0,715
0,622	0,903	0,710		High		
		0,837	0,064	0,815	0,837	0,826
0,766	0,962	0,851		Very High		
		0,897	0,058	0,835	0,897	0,865
0,819	0,958	0,820		Low		
Weighted Avg.		0,790	0,070	0,788	0,790	0,789
0,719	0,930	0,770				

==== Confusion Matrix ===

a	b	c	d	<-- classified as
2923	374	1	705	a = Medium
454	2831	768	0	b = High
1	659	3382	0	c = Very High
410	1	0	3563	d = Low

## Visualización del árbol de decisión generado



El árbol generado tiene un total de 30 nodos, 23 hojas y un máximo de 4 niveles.



## Red bayesiana

Para la construcción de la red bayesiana se utiliza la herramienta Elvira. Esta permite crear una red bayesiana Naive Bayes.

Los parámetros de entrada son:

Input

- Carat
- Cut
- Color
- Clarity
- Volume

Target

- Price

Open data cases file X

Cases file :  Browse

Preprocess options Machine learning Post learning

**CLASSIFIER STRUCTURE**

Naïve-Bayes ▼

Laplace correction

Substructure ▼

Todas ▼

Filter  Wrapper

95 %  99 %

Greedy  Umda

K parameter

A classic naïve Bayes model is built with all the predictive variables, that is, it assumes conditional independence among the predictive variables given the class.

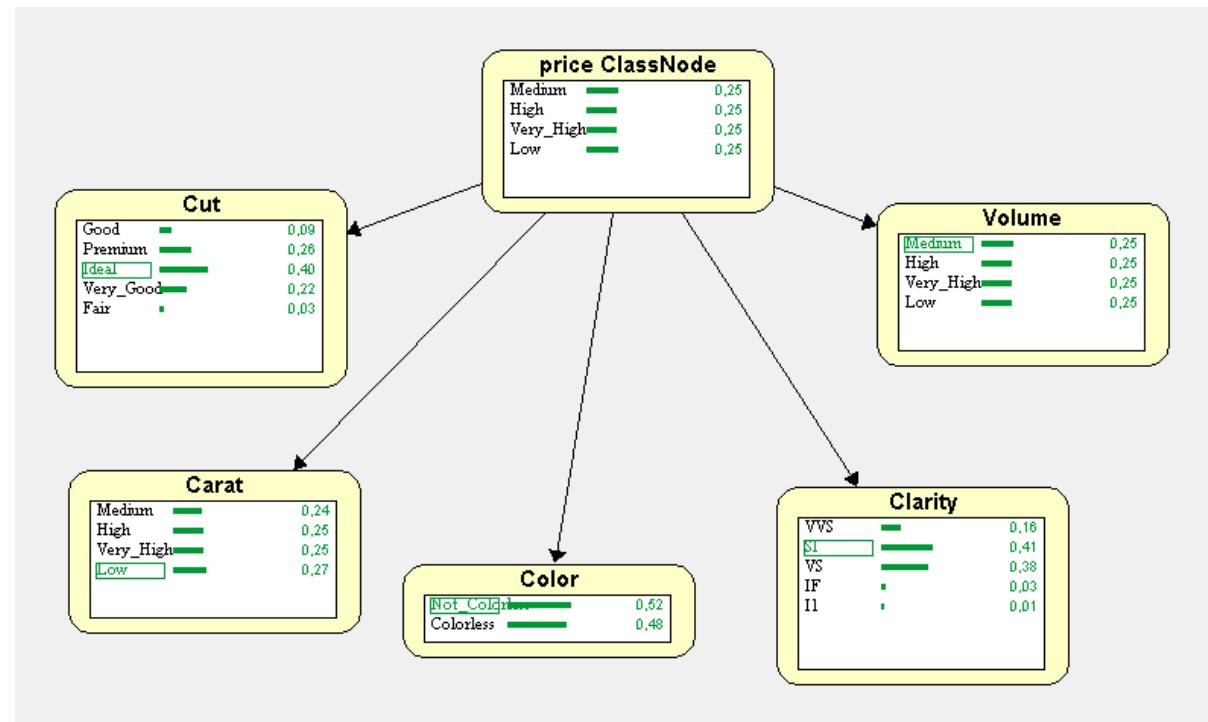
Process

Supervised clasification Unsupervised clasification Factorization

Cancel

Podemos visualizar la red bayesiana creada. En la misma se muestra que el precio depende de todas las demás variables y que éstas son independientes entre sí (de ahí el nombre Naive Bayes). Además, se muestran las probabilidades a priori de cada uno de los valores de los atributos y la clase.

## Probabilidades A Priori



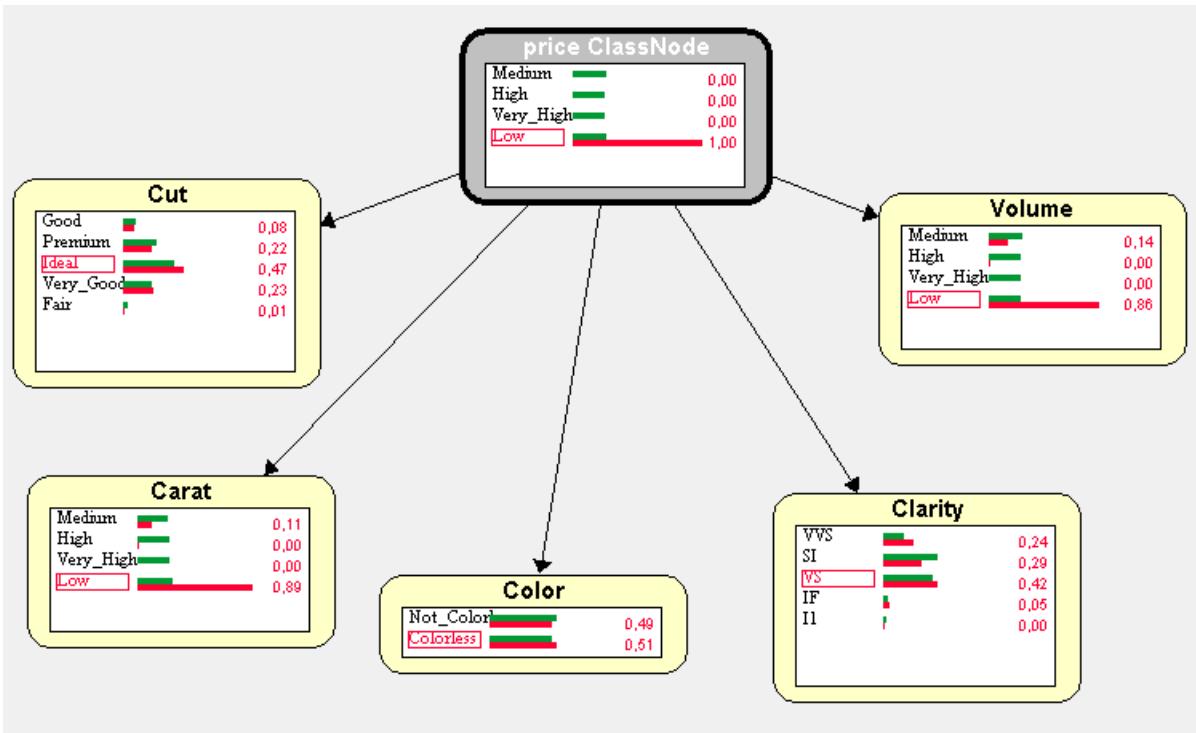
Atributo	Valor	A Priori
Precio	Low	0.25
	Medium	0.25
	High	0.25
	Very High	0.25
Carat	Low	0.27
	Medium	0.24
	High	0.25
	Very High	0.25
Volume	Low	0.25
	Medium	0.25
	High	0.25
	Very High	0.25



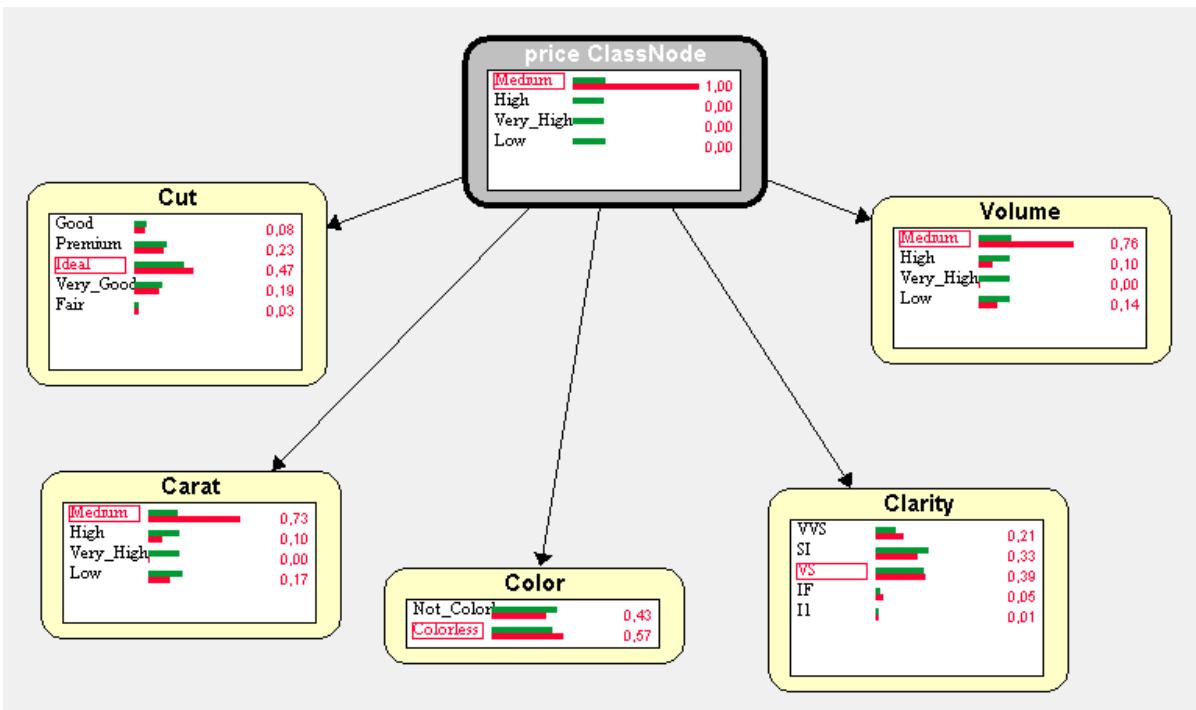
<b>Clarity</b>	IF	0.03
	VVS	0.16
	VS	0.38
	SI	0.41
	I1	0.01
<b>Color</b>	Colorless	0.48
	Not Colorless	0.52
<b>Cut</b>	Ideal	0.40
	Premium	0.26
	Very Good	0.22
	Good	0.09
	Fair	0.03

Las probabilidades a priori muestran que los valores de price, carat, volume y color están distribuidos bastante uniforme. En clarity predominan VS y SI y en cut los valores Ideal, Premium y Very Good.

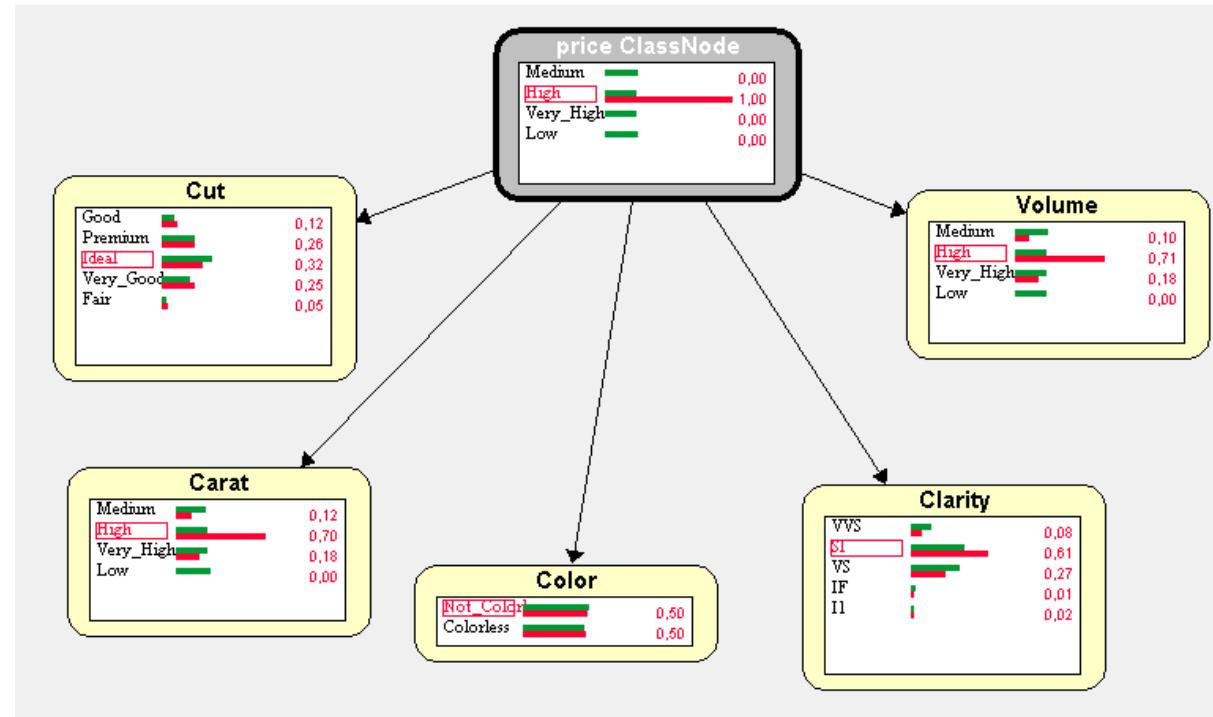
### Probabilidades dado Low Price



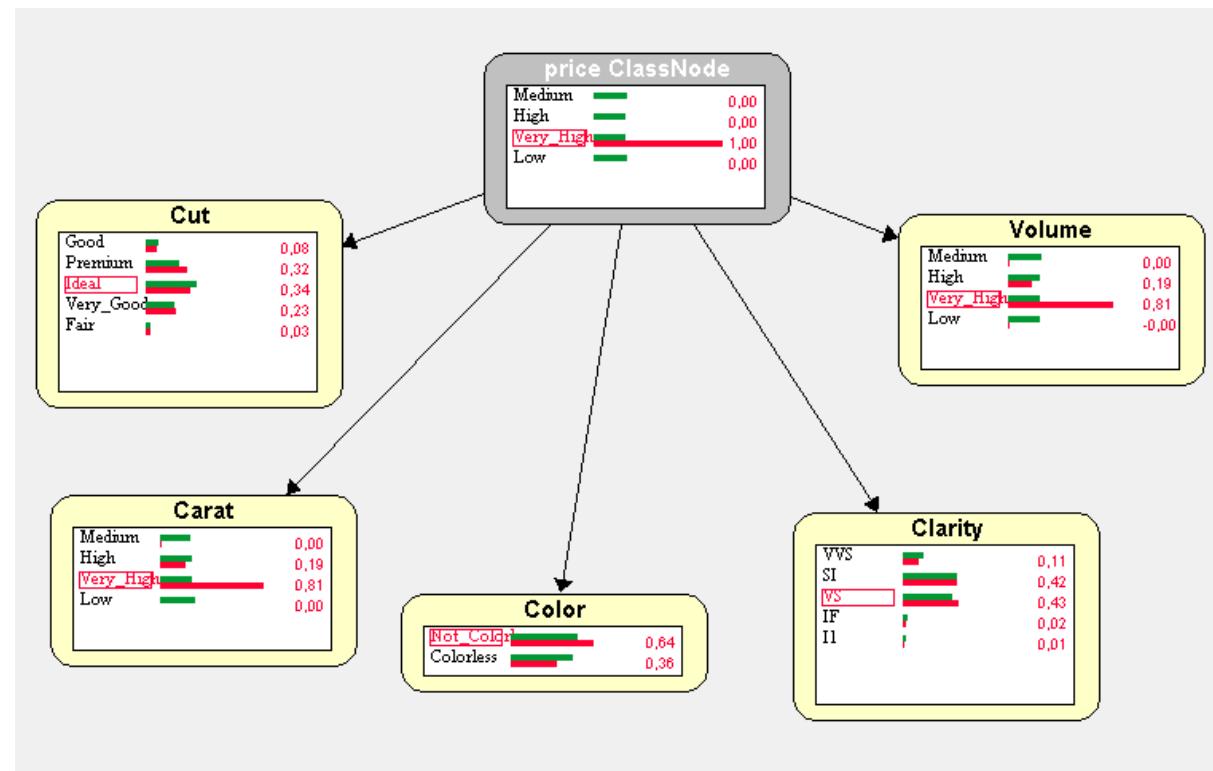
### Probabilidades dado Medium Price



### Probabilidades dado High Price



### Probabilidades dado Very High Price





Notemos que los atributos con más alta probabilidad a posteriori son Volume y Carat, esto se debe a que son los más importantes y brindan mayor información de predicción. A continuación mostramos los valores de máxima probabilidad a posteriori para los atributos Volume y Carat.

Price	Volume a posteriori	Carat a posteriori
Low	Low (0.86)	Low (0.89)
Medium	Medium (0.76)	Medium (0.73)
High	High (0.71)	High (0.70)
Very High	Very High (0.81)	Very High (0.81)

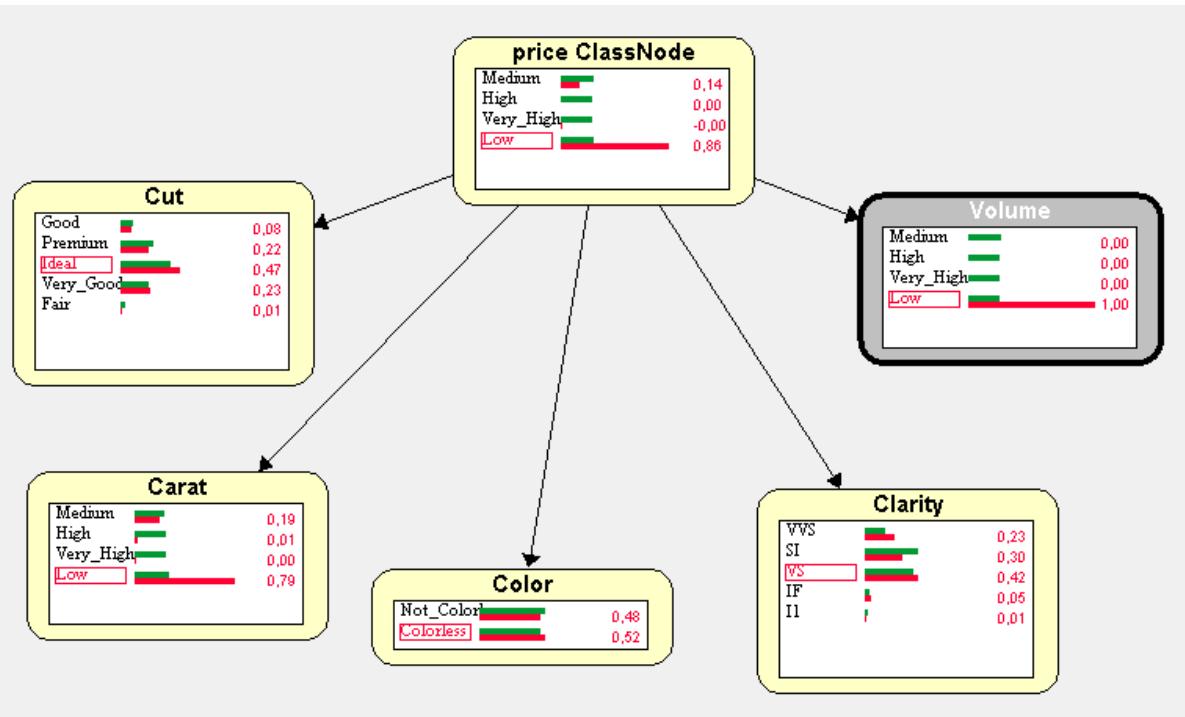
## Comparación

Podemos tomar algunas de las reglas generadas por el algoritmo de inducción y calcular la probabilidad que arroja la red bayesiana de pertenecer a una determinada clase. De esta manera, es posible validar que las reglas generadas por el árbol de decisión sean de calidad.

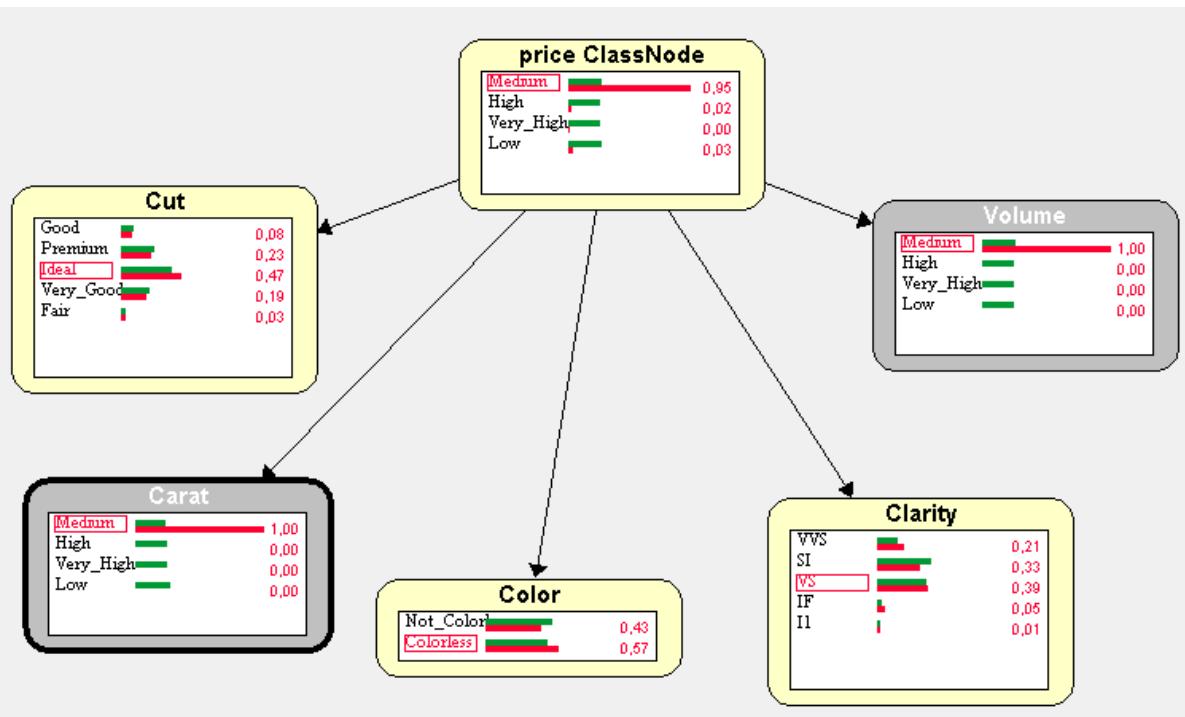
### Reglas

Regla	Volume	Carat	Clarity	Color	Cut	Price (predicción)
R1	Low	-	-	-	-	Low
R2	Medium	Medium	-	-	-	Medium
R3	High	High	SI	-	-	High
R4	Very High	-	VS	-	-	Very High

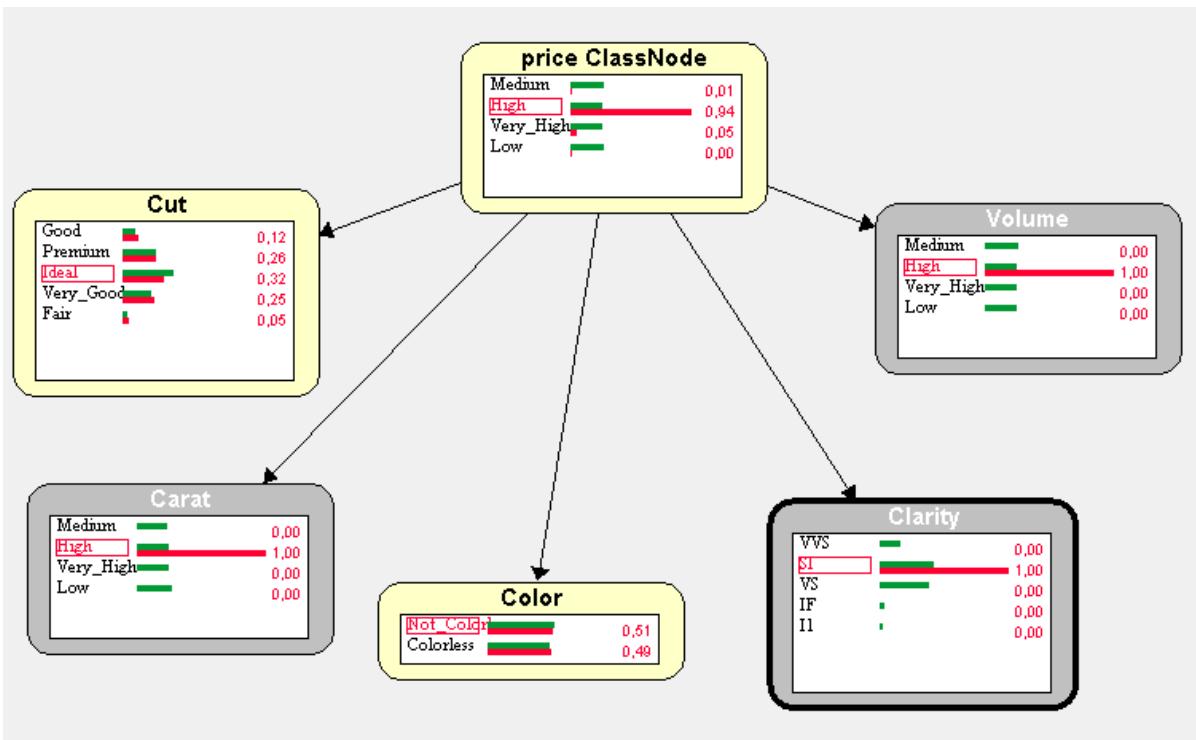
Regla 1



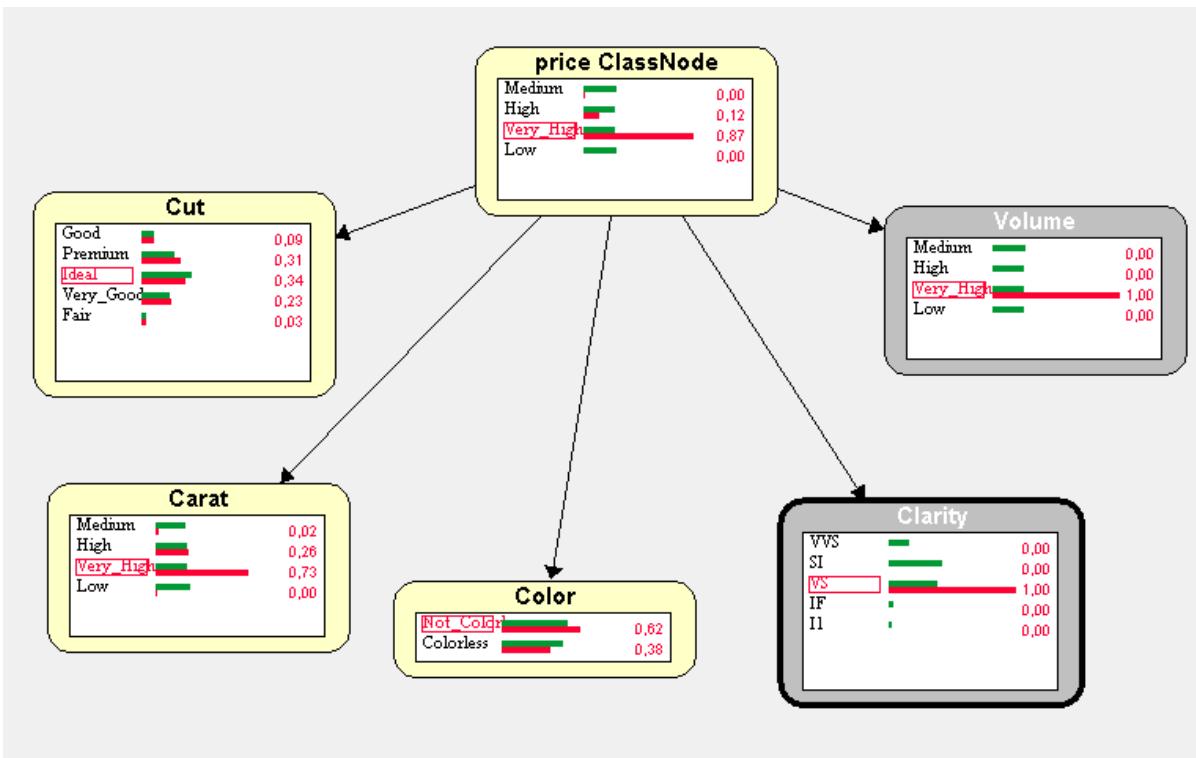
Regla 2



Regla 3



Regla 4





### Probabilidades calculadas por la red

Regla	Probabilidad Price			
	Low	Medium	High	Very High
R1 (Low)	0.86	0.14	0.00	0.00
R2 (Medium)	0.03	0.95	0.02	0.00
R3 (High)	0.00	0.01	0.94	0.05
R4 (Very High)	0.00	0.00	0.12	0.88

Podemos ver que la red bayesiana valida las reglas detalladas anteriormente con alta probabilidad.

## Reglas generadas

A partir de los análisis hechos anteriormente podemos obtener reglas de modo que sean utilizadas para tasar diamantes. Las reglas generadas son:

Price = Low			
Regla	Volume	Carat	Clarity
R1	Low	-	-
R2	Medium	Low	VS, SI, I1
Regla	Confianza %		
R1	86 %		
R2	73 %		

Price = Medium			
Regla	Volume	Carat	Clarity
R3	Medium	Medium, High	-



R4	Medium	Low	IF, VVS
R5	High	Medium	SI
R6	High	-	I1
<b>Regla</b>	<b>Confianza %</b>		
R3	78 %		
R4	82 %		
R5	60 %		
R6	52 %		

Price = High				
Regla	Volume	Carat	Clarity	Color
R7	High	-	VS	-
R8	High	-	VVS	Not Colorless
R9	High	High, Very High	SI	-
R10	Very High	-	I1	-
<b>Regla</b>	<b>Confianza %</b>			
R7	61 %			
R8	63 %			
R9	83 %			
R10	56 %			

Price = Very High				
Regla	Volume	Carat	Clarity	Color
R11	High	-	IF	-



R12	High	-	VVS	Colorless
R13	Very High	-	IF, VVS, VS, SI	-
<b>Regla</b>	<b>Confianza %</b>			
R11	57 %			
R12	61 %			
R13	83 %			

Estas 13 reglas permiten generar un modelo clasificador que dadas las características de un diamante devuelva un rango de precios (bajo, medio, alto y muy alto) y así utilizarlo como tasador.

## Evaluación del modelo

Las reglas generadas tienen que superar el 50% de confianza.

Regla	Confianza %
R1	86 %
R2	73 %
R3	78 %
R4	82 %
R5	60 %
R6	52 %
R7	61 %
R8	63 %
R9	83 %
R10	56 %
R11	57 %



R12	61%
R13	83%

Notemos que las 13 reglas generadas superan el 50 % de confianza solicitada en los criterios de éxito.

La evaluación de la precisión del modelo consiste en ejecutar las pruebas como se ha detallado en el apartado diseño de pruebas.

#### Resultado de la evaluación

La precisión al clasificar el set de prueba es de **78.85%**, clasificando correctamente de los 16072 registros totales a 12699. La precisión lograda es mucho mayor a la precisión base del 25% que arroja un clasificador al azar.

La precisión obtenida para cada uno de los valores de la clase es:

Price	Precisión %	Error %
Low	83.50 %	16.50 %
Medium	77.20 %	22.80 %
High	73.20 %	26.80 %
Very High	81.50 %	18.50 %
Promedio	<b>78.85 %</b>	21.15 %



## 5- Evaluación

### **Valoración de los resultados**

Como resultado del modelo, hemos obtenido [13 reglas](#) que (con una [confianza de más de 50%](#)) nos permiten clasificar con un [gran nivel de precisión](#) el rango de precio de los diamantes en base a sus atributos. Además, el modelo clasifica correctamente al **78.85%** de los datos del set de prueba.

Se puede ver entonces que cumplimos con los **criterios de éxito** que definimos para este proceso, siendo la precisión de la clasificación sobre el set de datos de prueba mayor al 70%, habiendo generado entre 10 y 15 reglas, y que todas las reglas tienen una confianza de más del 50%.

Así también, mediante el uso de estas reglas, podemos cumplir con el **objetivo de negocio**, que indica que las reglas deben ser simples y entendibles para que puedan ser utilizadas sin inconvenientes por el 90% de los individuos. En línea con esto, vemos que las reglas necesitan que el tasador evalúe como máximo 3 atributos para aplicarlas y que dichos atributos son fácilmente determinables.

### **Revisión del proceso**

Habiendo realizado una completa revisión del proceso y los datos, no hemos encontrado algún factor importante o tarea que se haya pasado por el alto, por lo tanto determinamos que en base a los resultados obtenidos y las conclusiones a las que llegamos en la etapa anterior al valorar los resultados, hemos cumplido con los objetivos de negocio y criterios de éxito que nos propusimos al iniciar este trabajo.

### **Próximos Pasos**

Como hemos cumplido con los objetivos de negocio y criterios de éxito del plan, y no es requerida la realización de la fase de Implementación, damos como concluido el trabajo.



# Bibliografía

- <https://www.argyor.com/es/informacion-diamantes.html>
- <https://www.joyeriasbizzarro.com/blog-inolvidable/que-es-el-fuego-de-un-diamante/>
- <https://www.diamonds.pro/education/cuts/>
- [https://en.wikipedia.org/wiki/Diamond\\_cut#Cut\\_grading](https://en.wikipedia.org/wiki/Diamond_cut#Cut_grading)
- <https://ige.org/gemologia/diamantes/peso/>
- <https://www.kaggle.com/shivam2503/diamonds>
- <https://es.wikipedia.org/wiki/Diamante>
- <https://es.wikipedia.org/wiki/Quilate>
- <https://beyond4cs.com/grading/depth-and-table-values/>
- <https://www.bluenile.com/pr/education/diamonds/depthpercentage>
- <https://www.kaggle.com/shivam2503/diamonds>