

Dados na Ciência, Gestão e Sociedade
Como é que o preço é influenciado pelas demais
variáveis?



Professores:

Ana Maria Almeida

Elsa Cardoso

Miguel Sales Dias

Nuno Alves

Trabalho elaborado por:

Diogo Freitas, N° 104841

João Botas, N° 104782

Maria João Lourenço, N° 104716

Umeima Mahomed, N° 99239

Índice

1-Introdução.....	2
2-Data understanding: Estudo das variáveis	3
3-Data Preparation: Limpeza	5
4-Modeling: Modelo	7
Métodos de estimativa para classificação	15
Conjuntos de treino e de teste do dataset	17
5-Evaluation: Avaliação e interpretação	18
6-Webgrafia	19

1-Introdução

Este relatório resulta de um trabalho de grupo, no domínio da Unidade Curricular Dados na Ciência, Gestão e Sociedade, no qual nos foi atribuída uma base de dados com o propósito de aplicar a metodologia CRISP-DM e responder a uma questão relacionada com a informação disponibilizada.

Inicialmente, o objetivo do grupo era responder à pergunta quais as variáveis que mais influenciam o preço e como é que os 4 C's influenciam o preço dos diamantes.

Deste modo, após um estudo detalhado sobre os dados disponibilizados na base de dados "Diamonds", decidimos que a questão em que nos iríamos focar durante este projeto seria a forma como diferentes variáveis, como o "carat", "cut", "color", "clarity", "depth", "table", X, Y e Z, influenciam o preço dos diamantes.

Utilizando a metodologia CRISP-DM, começámos por selecionar, limpar, pré-processar e transformar os dados, de seguida, procedemos à exploração, à modelação e à avaliação dos mesmos e, por último, apresentámos as conclusões retiradas acerca do modo como as variáveis acima mencionadas influenciam o preço.

Para realizarmos este trabalho utilizámos duas ferramentas: o Orange e o Excel.

O Orange, enquanto ferramenta que permite a visualização de dados, foi utilizada para o tratamento e análise dos dados disponibilizados, que se relacionam com os diamantes.

Por sua vez, o Excel foi fundamental para experimentarmos e percebermos as variáveis que iríamos precisar a fim de estudar o nosso objetivo. Além disso, foi uma ferramenta essencial que permitiu a limpeza dos dados em que iríamos trabalhar.

2-Data understanding: Estudo das variáveis

A base de dados disponibilizada apresentava dez variáveis diferentes, tais como, o “carat”, “cut”, “color”, “price”, “clarity”, “depth”, “table”, X, Y e Z.

Após algumas pesquisas em relação às diversas variáveis, que nos foram apresentadas na base de dados fornecida, conseguimos perceber o significado das mesmas, cuja explicação se encontra abaixo.

O peso do diamante - “carat” - tem influência direta no preço deste, sendo que um “carat” equivale a 200 miligramas.

A lapidação do diamante - “cut” - está relacionada com a forma como a luz é refratada quando em contacto com a superfície e com o interior deste. Uma boa lapidação implica simetria, polimento e proporção.

A pureza do diamante - “clarity” - mede a pureza e a raridade deste, sendo que esta pode ser visualizada sob uma lupa com um poder de ampliação de 10x. Esta variável pretende perceber se existem inclusões, no interior do diamante, e manchas, no exterior deste, já que o diamante é formado no interior da Terra a altas pressões e temperatura. As inclusões vão afetar a forma como a luz é refratada no diamante, diminuindo o valor deste. Os diamantes classificados como IF não têm inclusões no seu interior. Por sua vez, os diamantes que apresentam inclusões são classificados de forma crescente como VVS1 e VVS2, VS1 e VS2 e SI1 e SI2. Os diamantes I1 são considerados imperfeitos.

A profundidade do diamante - “depth” - é a distância desde a ponta do diamante até ao topo do mesmo.

A mesa do diamante - “table” - é a área da superfície deste, quanto virado para cima.

A variável “color” - cor - refere-se à cor que o diamante possui, após a sua formação, sendo que quanto mais transparente este for mais raro é. Esta variável influencia o aspeto do diamante. A escala vai de D até J, sendo que D é o melhor diamante porque é o mais transparente.

As variáveis X, Y e Z estão relacionadas com as dimensões do diamante.

A variável “price” - preço - é autoexplicativa.

Ao longo da nossa pesquisa apercebemo-nos que algumas variáveis são mais importantes que outras, pois influenciam o preço do diamante de forma direta,

como o “carat”, o “cut”, a “clarity” e a “color”, uma vez que estas formam o chamado 4 C’s.

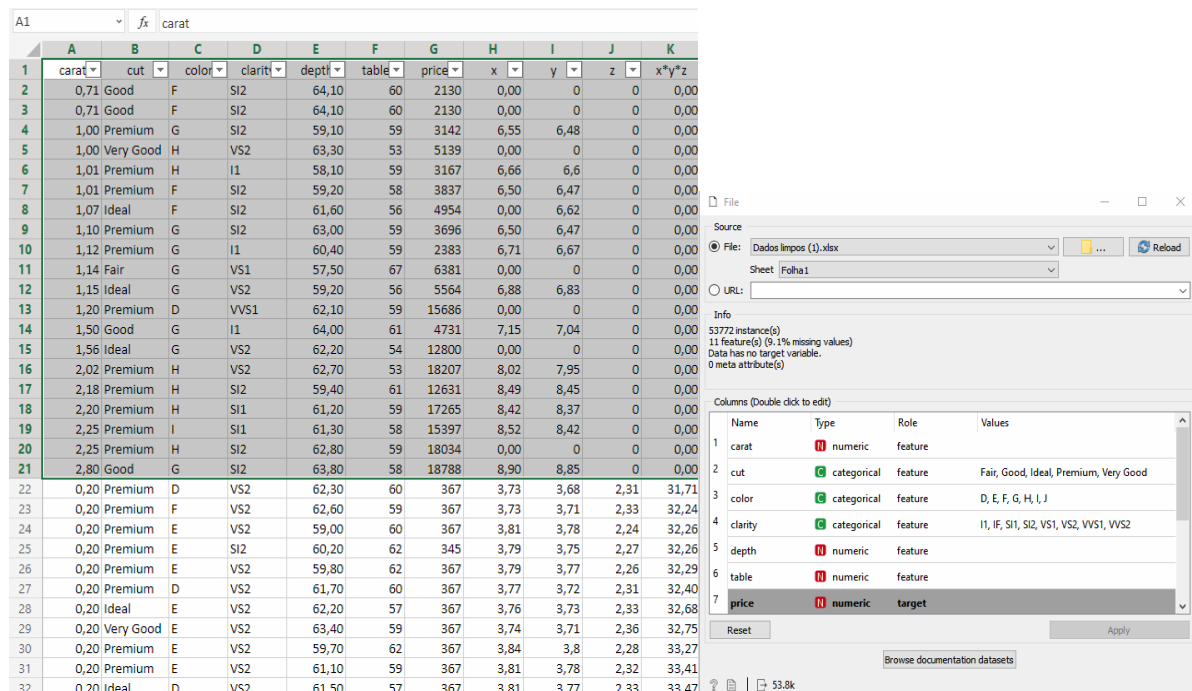
Além disso, é importante salientar que existe uma percentagem referente ao “table” e ao “depth” ideal para cada “shape” diferentes do diamante, como se pode ver na figura 1.

Ideal Depth & Table Percentages for Every Diamond Shape		
Diamond Shape	Ideal Table Percentage	Ideal Depth Percentage
Round Brilliant Cut	54 to 57%	59 to 62.6%
Princess Cut	69 to 75%	68 to 75%
Cushion Cut	< 68%	61 to 68%
Emerald Cut	60 to 68%	61 to 68%
Asscher Cut	60 to 68%	61 to 68%
Oval Cut	53 to 63%	< 68%
Pear Shape	53 to 65%	< 68%
Radiant Cut	61 to 69%	< 67%
Heart Shape	56 to 62%	56 to 66%
Marquise Cut	53 to 63%	58 to 62%

FIGURA 1

3-Data Preparation: Limpeza

Começamos por analisar em profundidade as diferentes variáveis, os seus valores e a sua classificação para percebermos a coerência dos valores apresentados na base de dados fornecida.

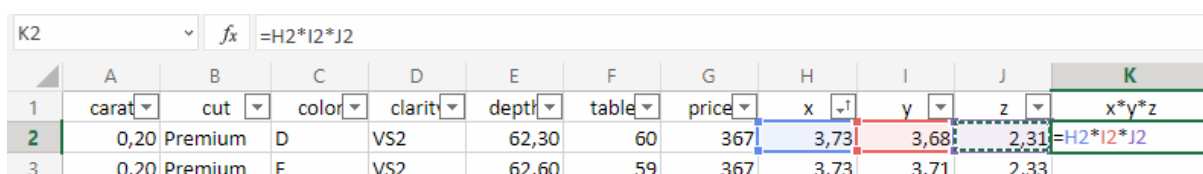


	A	B	C	D	E	F	G	H	I	J	K
1	carat	cut	color	clarity	depth	table	price	x	y	z	x*y*z
2	0,71	Good	F	SI2	64,10	60	2130	0,00	0	0	0,00
3	0,71	Good	F	SI2	64,10	60	2130	0,00	0	0	0,00
4	1,00	Premium	G	SI2	59,10	59	3142	6,55	6,48	0	0,00
5	1,00	Very Good	H	VS2	63,30	53	5139	0,00	0	0	0,00
6	1,01	Premium	H	I1	58,10	59	3167	6,66	6,6	0	0,00
7	1,01	Premium	F	SI2	59,20	58	3837	6,50	6,47	0	0,00
8	1,07	Ideal	F	SI2	61,60	56	4954	0,00	6,62	0	0,00
9	1,10	Premium	G	SI2	63,00	59	3696	6,50	6,47	0	0,00
10	1,12	Premium	G	I1	60,40	59	2383	6,71	6,67	0	0,00
11	1,14	Fair	G	VS1	57,50	67	6381	0,00	0	0	0,00
12	1,15	Ideal	G	VS2	59,20	56	5564	6,88	6,83	0	0,00
13	1,20	Premium	D	VVS1	62,10	59	15686	0,00	0	0	0,00
14	1,50	Good	G	I1	64,00	61	4731	7,15	7,04	0	0,00
15	1,56	Ideal	G	VS2	62,20	54	12800	0,00	0	0	0,00
16	2,02	Premium	H	VS2	62,70	53	18207	8,02	7,95	0	0,00
17	2,18	Premium	H	SI2	59,40	61	12631	8,49	8,45	0	0,00
18	2,20	Premium	H	SI1	61,20	59	17265	8,42	8,37	0	0,00
19	2,25	Premium	I	SI1	61,30	58	15397	8,52	8,42	0	0,00
20	2,25	Premium	H	SI2	62,80	59	18034	0,00	0	0	0,00
21	2,80	Good	G	SI2	63,80	58	18788	8,90	8,85	0	0,00
22	0,20	Premium	D	VS2	62,30	60	367	3,73	3,68	2,31	31,71
23	0,20	Premium	F	VS2	62,60	59	367	3,73	3,71	2,33	32,24
24	0,20	Premium	E	VS2	59,00	60	367	3,81	3,78	2,24	32,26
25	0,20	Premium	E	SI2	60,20	62	345	3,79	3,75	2,27	32,26
26	0,20	Premium	E	VS2	59,80	62	367	3,79	3,77	2,26	32,29
27	0,20	Premium	D	VS2	61,70	60	367	3,77	3,72	2,31	32,40
28	0,20	Ideal	E	VS2	62,20	57	367	3,76	3,73	2,33	32,68
29	0,20	Very Good	E	VS2	63,40	59	367	3,74	3,71	2,36	32,75
30	0,20	Premium	E	VS2	59,70	62	367	3,84	3,8	2,28	33,27
31	0,20	Premium	E	VS2	61,10	59	367	3,81	3,78	2,32	33,41
32	0,20	Ideal	D	VS2	61,50	57	367	3,81	3,77	2,33	33,47

FIGURA 2 – TARGET: PREÇO

Após uma breve análise dos dados, reparámos, através do Orange, que não existiam valores omissos.

No entanto, através do Excel, conseguimos descobrir a existência de dados inválidos nas variáveis X, Y e Z, que representam a dimensão do diamante. Como é possível verificar na figura 2, o valor zero (0) está presente nestas variáveis, algo que é impossível de acontecer, pois, um diamante não pode ter altura, largura ou comprimento igual a zero (0). Deste modo, decidimos criar uma nova coluna que permitisse ver se alguma das três variáveis era 0, usando a multiplicação entre elas, como vemos na figura 3.



	A	B	C	D	E	F	G	H	I	J	K
1	carat	cut	color	clarity	depth	table	price	x	y	z	x*y*z
2	0,20	Premium	D	VS2	62,30	60	367	3,73	3,68	2,31	=H2*I2*J2
3	0,20	Premium	F	VS2	62,60	59	367	3,73	3,71	2,33	

FIGURA 3 – CRIAÇÃO VARIÁVEL X*Y*Z

Deste modo, decidimos apagar as linhas que continham o valor zero (0) na dimensão do diamante, pois não só reduziríamos o número elevado de dados, facilitando a sua análise, como também o seu processamento no Orange, sem influenciar os resultados finais.

De seguida, criámos um filtro no Excel de modo a remover as linhas em que todas as variáveis eram iguais através da remoção dos valores duplicados – figura 4.

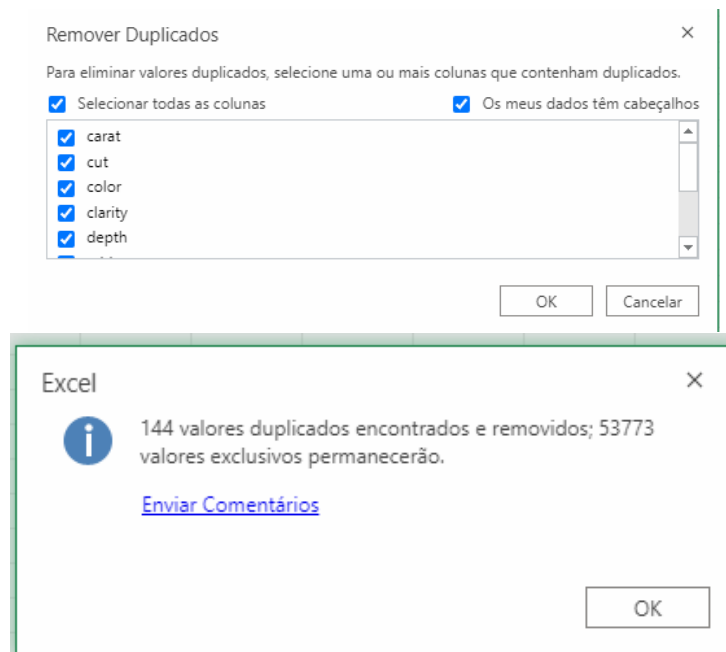


FIGURA 4 – REMOÇÃO DOS VALORES DUPLICADOS

4-Modeling: Modelo

Após a limpeza dos dados decidimos averiguar a forma como as diferentes variáveis interagem umas com as outras, de maneira a conseguirmos retirar conclusões fiáveis. Deste modo, prosseguimos para o estudo e análise das mesmas.



FIGURA 5 – “FEATURED STATISTICS”

Antes de mais, optámos por verificar num “Featured Statistics” se os dados estavam corretos e sem falhas, o que acabou por se verificar, tal como é possível visualizar na figura 5.

	#	Univar. reg.	RReliefF
N x	.	NA	0.092
N depth	.	NA	0.056
N carat	.	NA	0.055
C clarity	8	NA	0.048
N table	.	NA	0.042
N y	.	NA	0.012
C color	7	NA	0.011
N z	.	NA	0.010
C cut	5	NA	0.003

FIGURA 6 – “RANK”

Começamos por explorar no “Rank” - figura 6 - quais das variáveis influenciavam mais o preço, que é o nosso target, e chegamos à conclusão de que eram o “carat”, X, “table”, “depth” e “clarity”, sendo o “table” e o “depth” variáveis não exploradas nem analisadas.

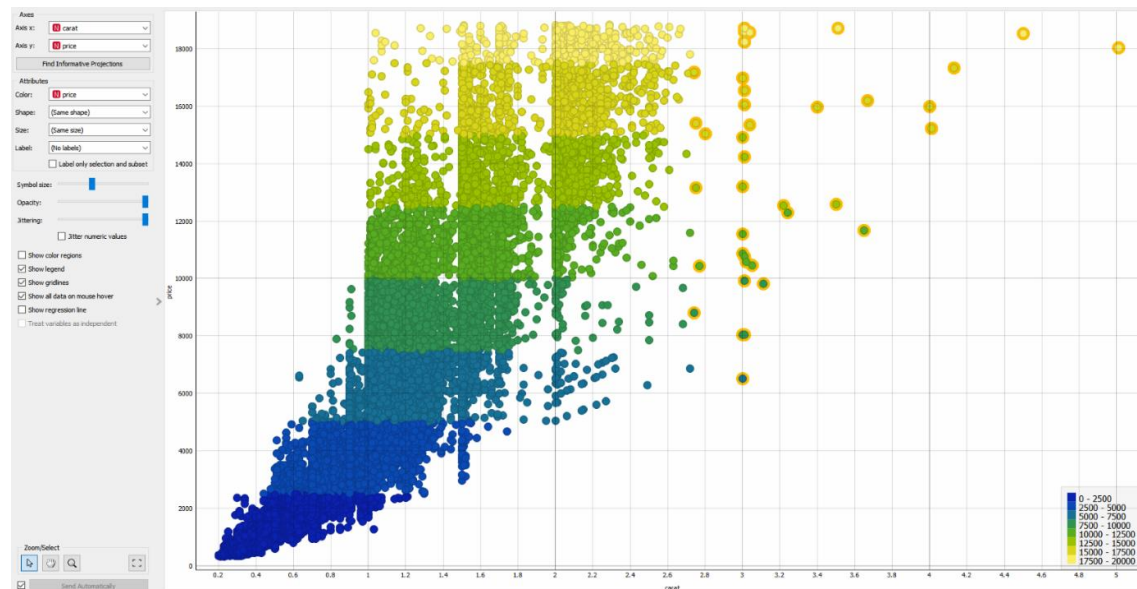


FIGURA 7 - “SCATTER PLOT”

Estando o “carat” ou o peso do diamante no primeiro lugar das variáveis que mais influenciam o preço, fomos ver a relação entre o peso e o preço.

Podemos assim concluir que, de forma geral, quanto mais pesado o diamante, maior será o seu valor, mas também podemos reparar que em alguns casos essa relação não ocorre e, por isso, selecionamos esses casos e fomos explorá-los na “Data Table”.

	price	Group	carat	cut	color	clarity	depth	table	x	y	z
40	18010	G1	5.01	Fair	J	I1	65.5	59.0	10.74	10.54	6.98
44	19531	G1	4.50	Fair	J	I1	63.8	58.0	10.23	10.16	6.72
46	17106	G1	4.13	Fair	H	I1	64.8	61.0	10.80	9.85	6.40
42	15223	G1	4.01	Premium	I	I1	61.0	61.0	10.14	10.10	6.17
42	15223	G1	4.01	Premium	J	I1	62.5	62.0	10.02	9.94	6.24
41	15994	G1	4.00	Very Good	I	I1	63.3	58.0	10.01	9.94	6.31
39	16186	G1	3.67	Premium	I	I1	62.4	56.0	9.86	9.81	6.15
35	11668	G1	3.65	Fair	H	I1	61.1	53.0	9.53	9.48	6.38
38	16701	G1	3.51	Premium	J	VS2	62.5	59.0	9.66	9.63	6.03
17	12587	G1	3.50	Ideal	H	I1	62.8	57.0	9.65	9.59	6.03
29	15946	G1	3.40	Fair	D	I1	66.8	52.0	9.42	9.34	6.27
32	12308	G1	3.24	Premium	H	I1	62.1	58.0	9.44	9.40	5.85
33	12545	G1	3.22	Ideal	I	I1	62.6	55.0	9.49	9.42	5.82
17	9623	G1	3.11	Fair	J	I1	63.9	57.0	9.15	9.02	5.98
21	10452	G1	3.05	Premium	E	I1	63.9	58.0	9.26	9.25	5.66
34	16106	G1	3.04	Premium	I	S2	59.3	60.0	9.51	9.46	5.62
36	13354	G1	3.04	Very Good	I	S2	63.2	59.0	9.14	9.07	5.75
19	10177	G1	3.02	Fair	I	I1	63.2	56.0	9.11	9.02	5.81
38	10791	G1	3.01	Fair	H	I1	59.1	62.0	9.54	9.38	5.31
31	16986	G1	3.01	Good	H	S2	57.8	64.0	9.44	9.38	5.42
30	14238	G1	3.01	Premium	G	S2	59.8	58.0	9.44	9.37	5.62
27	16710	G1	3.01	Premium	J	S2	59.7	58.0	9.41	9.32	5.59
13	16340	G1	3.01	Premium	I	S2	60.2	59.0	9.36	9.31	5.62
24	16710	G1	3.01	Premium	J	S2	60.7	59.0	9.35	9.22	5.64
20	16031	G1	3.01	Ideal	J	S2	61.7	58.0	9.25	9.20	5.69
12	9623	G1	3.01	Premium	F	I1	62.2	56.0	9.24	9.13	5.75
14	8848	G1	3.01	Premium	I	I1	62.7	58.0	9.10	8.97	5.67
12	16340	G1	3.01	Good	I	S2	63.9	60.0	9.06	9.01	5.77
8	16242	G1	3.00	Fair	I	S2	63.8	56.0	8.99	8.94	5.90
7	16326	G1	3.01	Ideal	J	I1	63.4	60.0	8.99	8.93	5.86
28	13303	G1	3.00	Premium	G	I1	59.7	60.0	9.42	9.26	5.58
26	10863	G1	3.00	Good	I	I1	57.0	64.0	9.38	9.31	5.33
23	14910	G1	3.00	Good	J	S2	59.3	64.0	9.32	9.19	5.50
11	16910	G1	3.00	Premium	I	S2	62.7	59.0	9.30	9.14	5.60
18	6512	G1	3.00	Very Good	H	I1	63.1	55.0	9.23	9.10	5.77
13	11548	G1	3.00	Good	E	I1	64.2	65.0	9.08	8.96	5.79
6	8848	G1	3.00	Fair	H	I1	67.1	57.0	8.93	8.84	5.67
4	16976	G1	3.00	Fair	I	S2	64.8	59.0	8.93	8.73	5.72
10	15020	G1	2.80	Premium	I	S2	61.1	59.0	9.03	8.88	5.50
5	13406	G1	2.77	Premium	H	I1	62.6	62.0	8.93	8.83	5.56
11	11106	G1	2.75	Ideal	D	I1	60.9	57.0	9.04	8.88	5.49
9	13415	G1	2.75	Premium	H	S2	65.5	61.0	8.99	8.87	5.48
3	17194	G1	2.74	Very Good	H	S2	63.3	58.0	8.88	8.84	5.61
2	17194	G1	2.74	Very Good	J	S2	61.5	62.0	8.87	8.80	5.46
1	8882	G1	2.74	Fair	J	I1	64.9	61.0	8.76	8.66	5.65

FIGURA 8 - DIAMANTES QUE NÃO SEGUEM A RELAÇÃO PESO - PREÇO

Na figura 8 podemos ver os casos onde esta relação de peso e preço não ocorreram totalmente da forma que estávamos à espera, o que levou à retirada de algumas conclusões.

Apesar do peso do diamante ser a variável que mais influencia o diamante, o preço continua a ser influenciado por outras variáveis, tais como o corte, a pureza e a cor. Se olharmos com atenção, é notório a má qualidade, de forma geral, dos diamantes presentes na tabela, que têm ou um mau corte ou pouca pureza ou uma cor amarela.

Temos por exemplo, o primeiro diamante presente na figura 8, que corresponde ao diamante mais pesado, mas não é o mais caro, pois, apresenta a pior cor, um mau corte e pureza.

Enquanto discutíamos o próximo tópico, surgiu uma dúvida por efeito secundário, ou seja, ficámos com a sensação de que quanto mais pesado o diamante fosse, maior era a sua dimensão.

Com isto, fomos investigar se isso realmente era verdade.

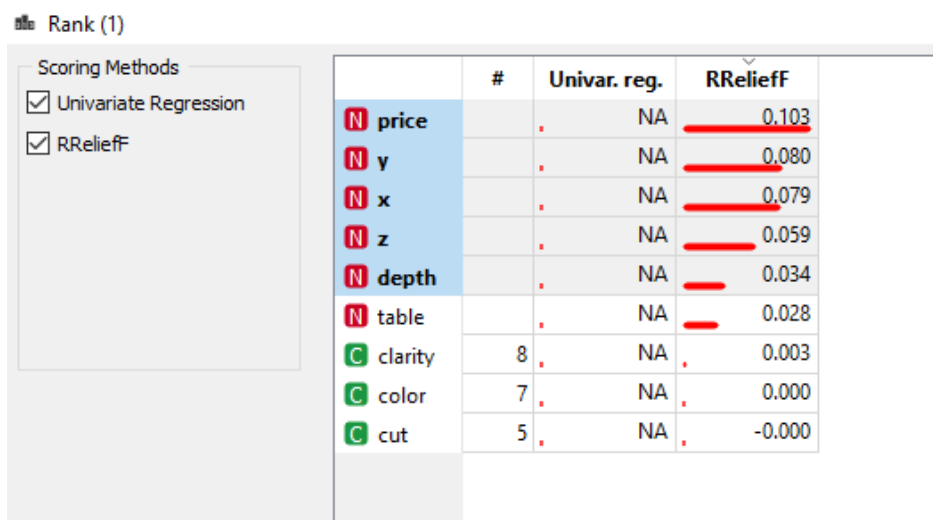


FIGURA 9 – “RANK”

Primeiro começámos por verificar quais as variáveis que mais influenciavam o peso do diamante e, pela figura 9, é possível verificar que tanto o Y, o X e o Z encontram-se no Top5.

Também concluímos que, como o peso era o que mais influenciava o preço, decidimos verificar se o inverso também seria verdadeiro, ou seja, se a variável preço também é a que mais influencia a variável peso.

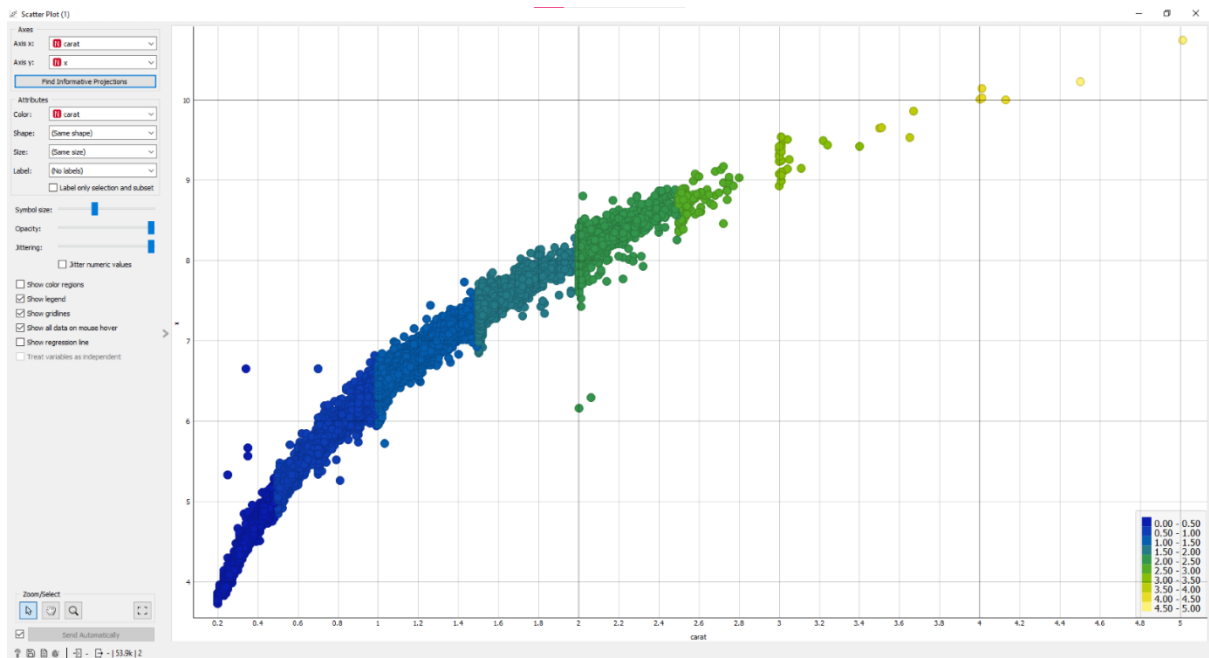


FIGURA 10

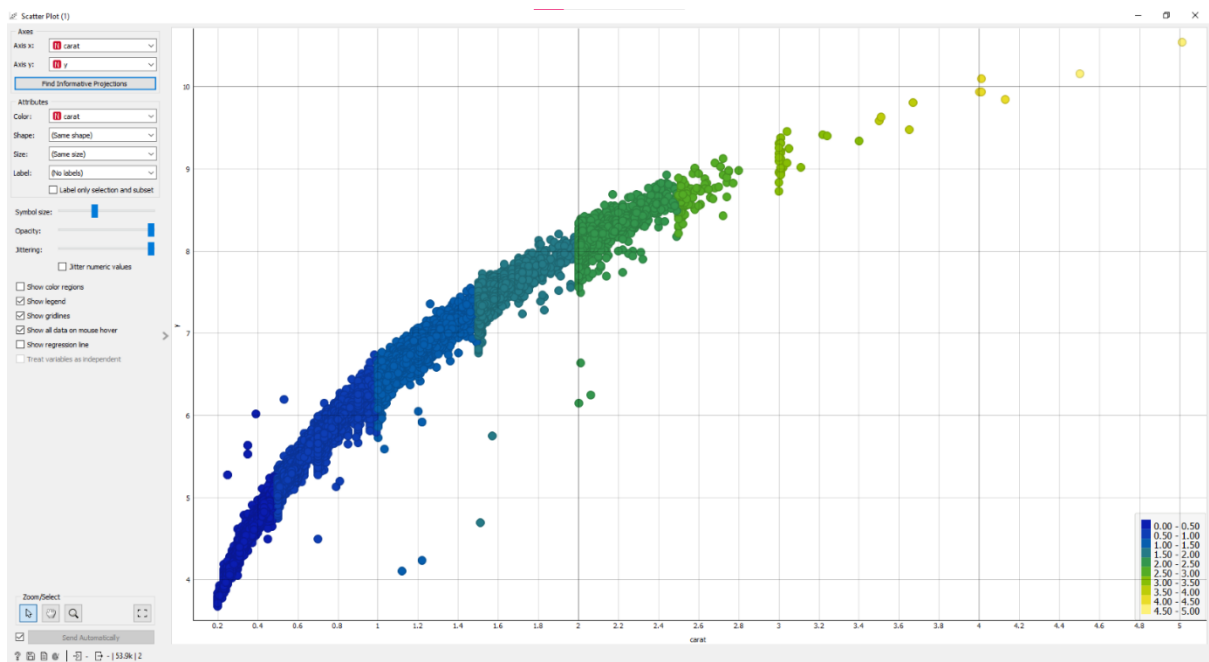


FIGURA 11

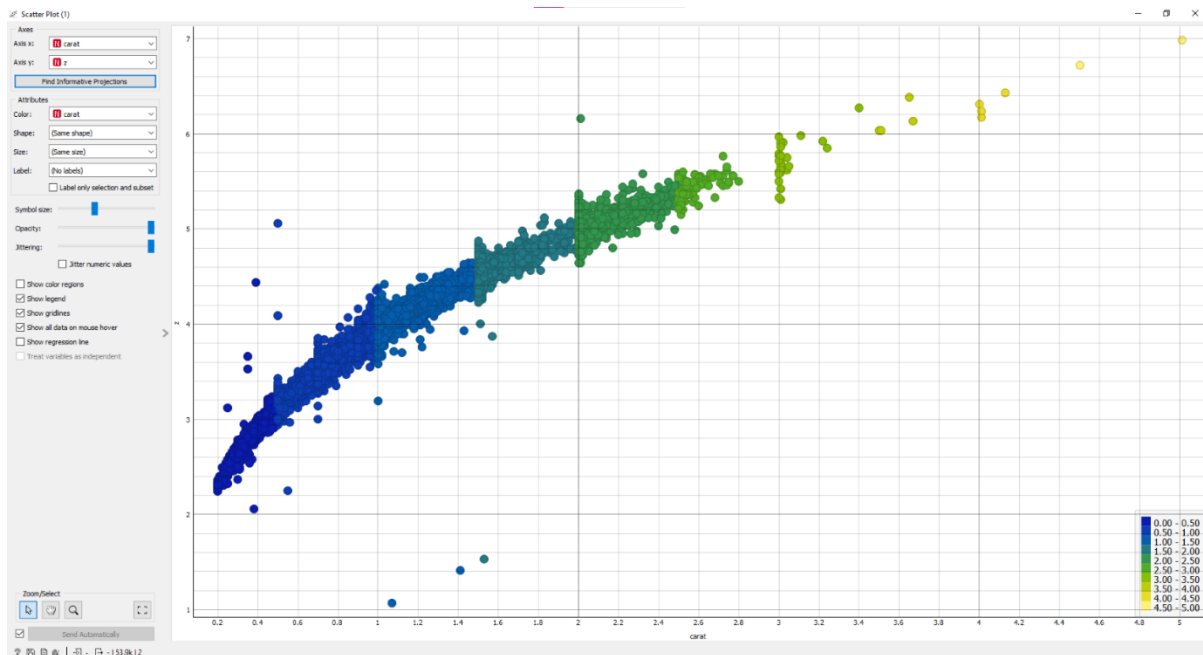


FIGURA 12

Após analisarmos os gráficos acima, onde podemos relacionar o peso com o X (figura 10), o Y (figura 11) e o Z (figura 12), chegámos à conclusão que, quando uma destas variáveis aumentava, o peso também aumentava.

Podemos assim concluir que quanto maior o diamante, ou seja, quanto maior o X, Y ou o Z do próprio diamante, maior será o peso do diamante contribuindo para o aumento do seu preço.

Concluída a nossa pesquisa e análise sobre as variáveis que mais influenciavam o preço dos diamantes, reparámos que havia uma variável, pertencente aos 4C's, que tinha uma grande influência no preço do diamante. Das categóricas a "clarity" é a que influenciava mais o preço.

Deste modo, decidimos pôr o "clarity" ou pureza como "split by" de colunas numa tabela de distribuição e relacionar com os outros dois C's categóricos, o "cut" (figura 13) e o "color" (figura 14). Assim, conseguimos retirar a quantidade do nível de pureza nas categorias do "cut" e no "color", sendo a pureza I1 e a IF os tipos menos predominantes, e os VVS1 e VVS2 os mais predominantes na relação com as outras duas variáveis.

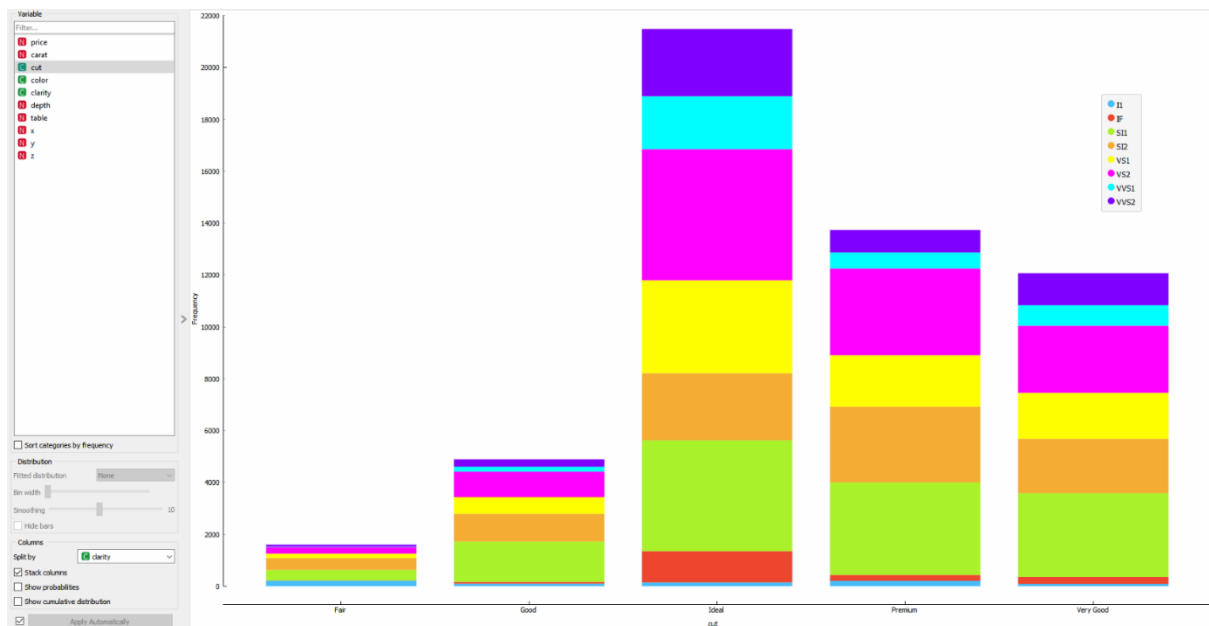


FIGURA 13 – “CUT”

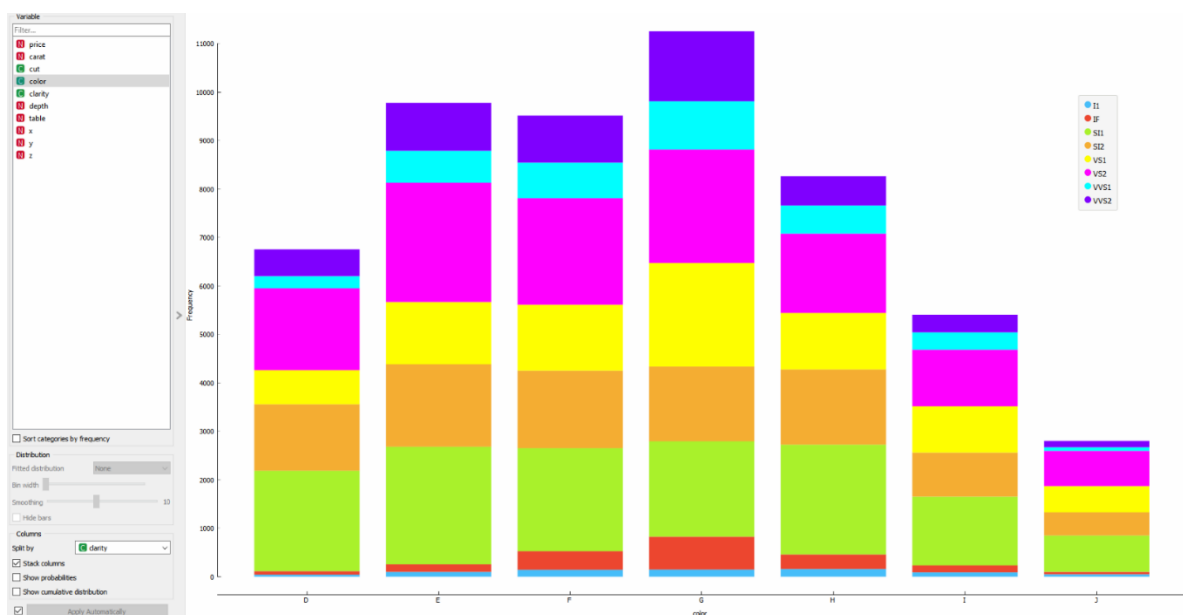


FIGURA 14 – “COLOR”

Com este pensamento em vista, o grupo decidiu analisar esta variável para verificar como ela se comportava. Por se tratar de uma variável categórica, passamos a poder utilizar ferramentas e algoritmos de aprendizagem supervisionada, enriquecendo a exploração e a análise dos nossos dados.

Columns (Double click to edit)

	Name	Type	Role	Values
1	carat	N numeric	feature	
2	cut	C categorical	feature	Fair, Good, Ideal, Premium, Very Good
3	color	C categorical	feature	D, E, F, G, H, I, J
4	clarity	C categorical	target	I1, IF, SI1, SI2, VS1, VS2, VVS1, VVS2
5	depth	N numeric	feature	
6	table	N numeric	feature	
7	price	N numeric	feature	
8	x	N numeric	feature	

Reset Apply

Browse documentation datasets

53.8k

FIGURA 15 – TARGET: “CLARITY”

Sendo a variável “clarity”, a variável categórica mais importante, quando relacionada com o preço, decidimos escolher esta variável com o objetivo de estudar e analisá-la no “Test and Score” a fim de usar algoritmos de aprendizagem supervisionada.

	#	Gain ratio	Gini
N carat		0.057	0.018
N x		0.056	0.017
N z		0.056	0.017
N y		0.054	0.017
N price		0.037	0.014
C cut	5	0.023	0.006
C color	7	0.010	0.005
N d...h		0.010	0.003
N table		0.010	0.003

FIGURA 16 – “RANK”

Com o objetivo de facilitar o processamento dos dados, fomos ver quais as variáveis que menos influenciavam o “clarity” para as retirarmos e não as usarmos na análise.

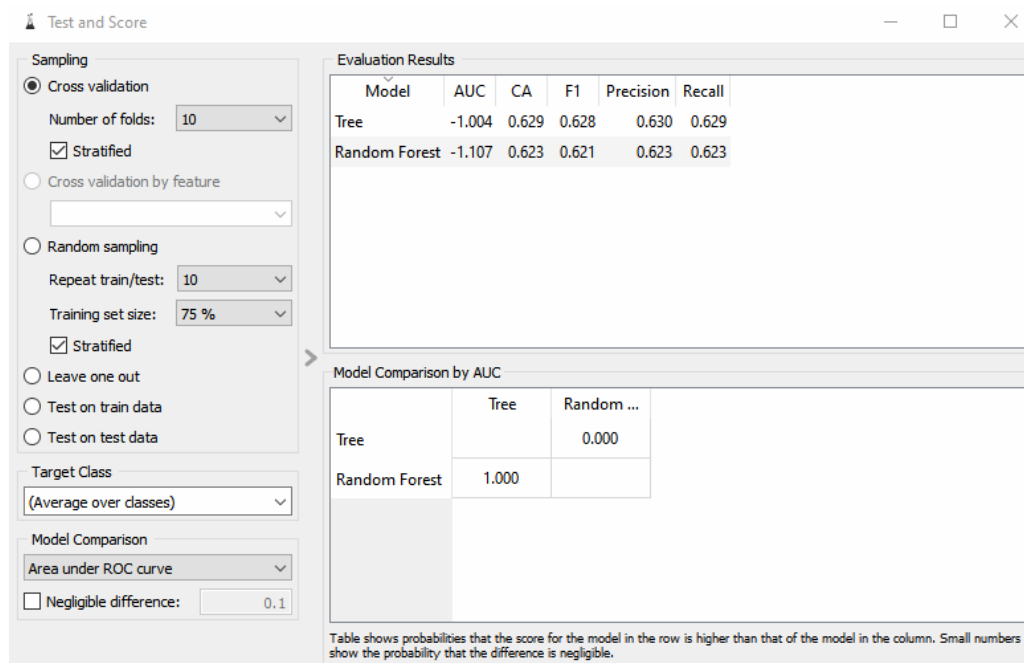


FIGURA 17

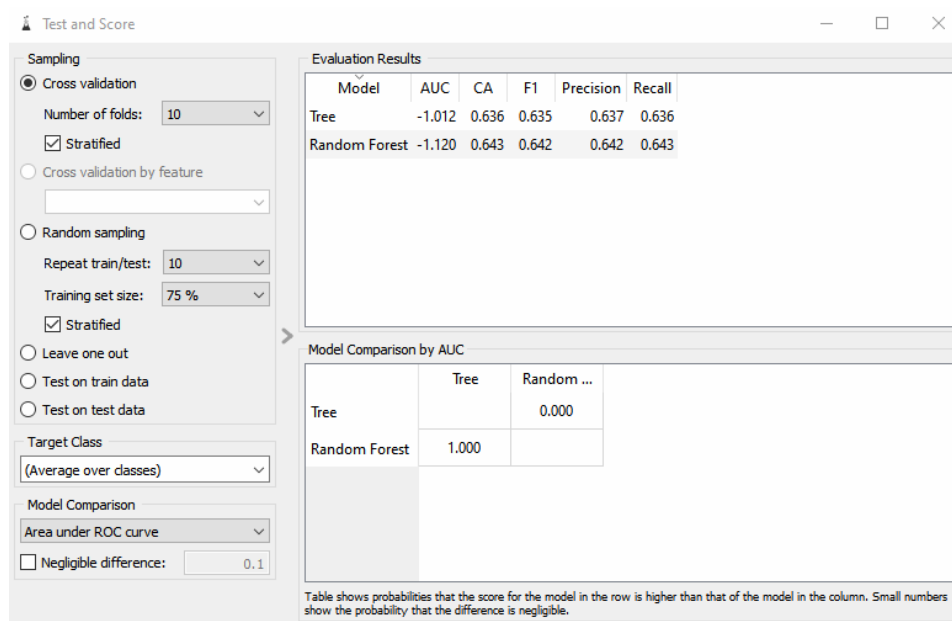


FIGURA 18

Para demonstrar a veracidade da situação, podemos verificar na figura 17 que quando as variáveis “table” e “depth” estavam em uso, a precisão se situava em 63 % - quando analisada na “Tree”, enquanto na figura 18, após retirarmos estes dados, a precisão subiu para 64,2 % - na “Random Forest”.

É de notar que estes resultados foram retirados com a utilização de todos os dados, sem uso de um “Data Sampler”. Sendo a variável a estimar categórica, utilizámos um método de aprendizagem supervisionada - a classificação.

Explicação Abaixo :

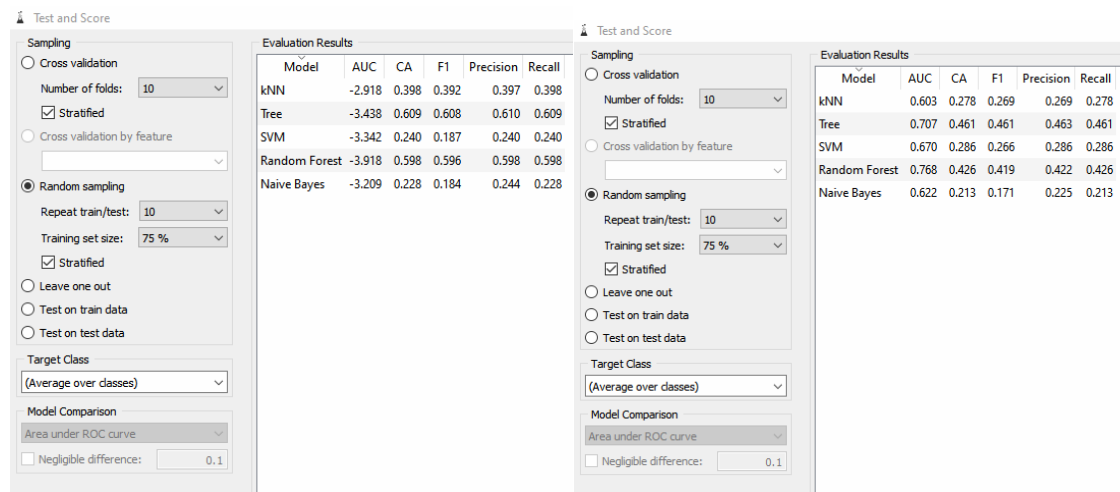


FIGURA 19

Métodos de estimativa para classificação

Nos seguintes modelos referidos na figura 19 (kNN, SVM, Naive Bayes) podemos observar que a precisão de acerto foi bastante inferior quando comparada com a da “Tree” e a da “Random Forest”, pelo que decidimos só utilizar estes dois modelos nas experiências utilizadas daqui para a frente.

Learners		Predicted								
		I1	IF	SI1	SI2	VS1	VS2	VVS1	VVS2	Σ
Actual	I1	534	0	30	349	3	3	1	0	920
	IF	0	1176	11	1	144	78	516	304	2230
	SI1	4	4	11253	2372	558	1982	49	68	16290
	SI2	97	1	2422	8554	59	289	4	4	11430
	VS1	0	35	924	124	5226	2890	289	707	10195
	VS2	2	21	2772	413	1843	9812	86	331	15280
	VVS1	0	317	61	1	532	243	2361	1040	4555
	VVS2	0	191	187	20	1238	632	667	3385	6320
	Σ	637	1745	17660	11834	9603	15929	3973	5839	67220

FIGURA 20 – “CONFUSION MATRIX” COM O “RANDOM FOREST”

Decidimos conectar a “Confusion Matrix” ao “Random Forest”, por ser o com maior taxa de precisão.

De seguida, conectamos o “Test and Score” à “Confusion Matrix”, para podermos analisar a eficácia do programa em reconhecer a “clarity” dos diamantes,

ou seja, se o programa conseguia indicar a “clarity” de um diamante de forma correta, sem erro.

Para que o programa funcionasse de forma perfeita, seria necessário que a diagonal principal da matriz fosse a única com valores. Quanto mais valores tiverem fora da diagonal, pior é a eficácia do programa.

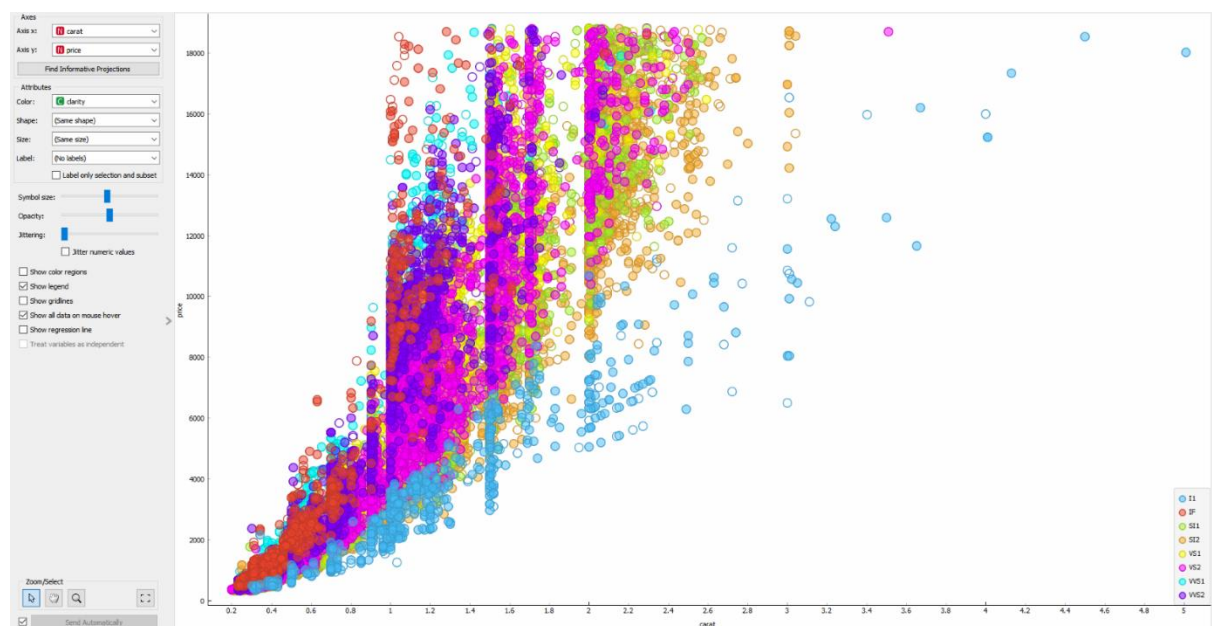


FIGURA 21 – “SCATTER PLOT”

Finalmente com a criação do programa, conectámos o “Scatter Plot” ao “File” e à “Confusion Matrix” para comparar quantos dados estavam corretos e não eram mal interpretados pelo programa. Na figura 21 podemos verificar o resultado dessa ligação, sendo que utilizámos o mesmo gráfico que relaciona o preço com o peso, onde podemos verificar bolinhas totalmente pintadas, e outras não. As que não estão totalmente pintadas representam as falhas e podemos verificar que existem em um número elevado.

Conjuntos de treino e de teste do dataset

Data Sample					Remaining data				
	90%	80%	70%	60%		10%	20%	30%	40%
Tree	0,611	0,602	0,594	0,588	Tree	0,444	0,511	0,545	0,558
Random Forest	0,596	0,591	0,574	0,568	Random Forest	0,419	0,479	0,520	0,526

FIGURA 22 - DADOS RETIRADOS A PARTIR DE 5 FOLDS E DE 5 REPETIÇÕES DE TESTE

Usando um “Data Sampler” pudemos fazer vários ensaios no “Test and Score”. Primeiro testámos para uma percentagem de 90% dos dados retirados aleatoriamente do “Data Sampler” e, em seguida, testámos para os restantes 10% (“Remaining Data”), verificando a precisão dos mesmos. De seguida, fizemos o mesmo estudo, mas para as proporções de 80%-20%,

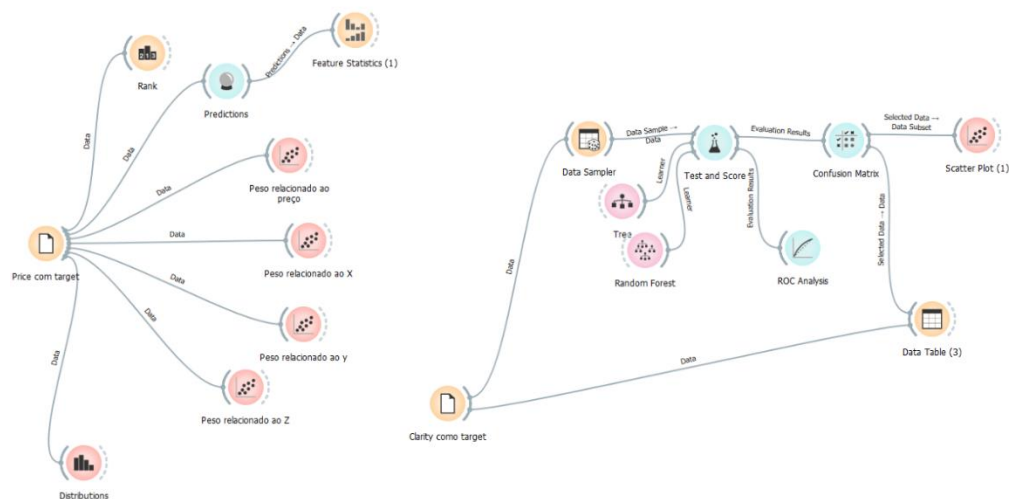
5-Evaluation: Avaliação e interpretação

Após termos todos os dados limpos e gráficos realizados na ferramenta Orange decidimos retirar conclusões acerca de quais as variáveis influenciam o preço, respondendo à nossa pergunta inicial, através da avaliação e interpretação dos resultados obtidos.

Em suma, podemos verificar a partir da análise dos dados que, de alguma forma, conseguimos refutar a informação fornecida pelas ourivesarias e sites de diamantes que afirmam que os 4 C's ("cut", "color", "clarity" e "carat") são as características que mais influenciam o preço dos diamantes.

A partir da nossa análise, e dos nossos resultados, conseguimos perceber que o "carat" e a "clarity" são as únicas características presentes nos 4C's que realmente têm alguma influência no preço dos diamantes.

Fora dos 4C's, temos também as variáveis X, "table" e "depth" que têm uma grande influência nos preços, mais influencia que o "cut" e "color" que pertencem aos 4C's, que são consideradas as características mais importantes dos diamantes.



6-Webgrafia

7 Factors to Consider When Buying a Diamond. (s.d.). Obtido em 5 de 11 de 2021, de Beldiamond: <https://www.beldiamond.com/blogs/guidance/7-factors-to-consider-when-buying-a-diamond>

Diamantes x valor. (n.d.). Retrieved 11 5, 2021, from Giulietta: <https://www.giuliettajoias.com.br/diamantes-x-valor/>

Diamond Education. (n.d.). Retrieved 11 3, 2021, from With Clarity: <https://www.withclarity.com/education/diamond-education/diamond-cut/what-is-diamond-depth-or-diamond-education>

Harris, D. (2017, March 30). *Techniques for Data Cleaning and Integration in Excel.* Retrieved from Software Advice: <https://www.softwareadvice.com/resources/excel-data-cleaning-integration-techniques/>

How to select diamond. (n.d.). Retrieved 11 3, 2021, from Diamond Collection: <http://www.diamondc.com.hk/us/how-to-select-diamond>

Pureza do diamante. (n.d.). Retrieved 11 3, 2021, from Tiffany: <https://www.tiffany.com.br/engagement/the-tiffany-guide-to-diamonds/clarity/>

Rocha, I. (2016, 6 2). *Diamante: Conheça a origem e o valor desta pedra preciosa!* Retrieved 11 5, 2021, from Blog Pedras Preciosas: <http://blog.pedrasriscas.pt/dicas-preciosas/diamante-origem-valor/>

Tomich, A. (n.d.). *Você sabe quais são os 4 cs do diamante?* Retrieved 11 4, 2021, from Anatomich: <https://www.anatomich.com/voce-sabe-quais-sao-os-4-cs-do-diamante/>

Your complete diamond characteristics guide. (n.d.). Retrieved 11 4, 2021, from Yadav: <https://www.yadavjewelry.com/info/diamond-education/your-complete-diamond-characteristics-guide>