Shape

Description automatically generated with medium confidence

# 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, Nº 104841

João Botas, Nº 104782

Maria João Lourenço, Nº 104716

Umeima Mahomed, Nº 99239

## Índice

[1-Introdução 2](#_Toc87351004)

[2-Data understanding: Estudo das variáveis 3](#_Toc87351005)

[3-Data Preparation: Limpeza 5](#_Toc87351006)

[4-Modeling: Modelo 7](#_Toc87351007)

[Métodos de estimativa para classificação 15](#_Toc87351008)

[Conjuntos de treino e de teste do dataset 17](#_Toc87351009)

[5-Evaluation: Avaliação e interpretação 18](#_Toc87351010)

[6-Webgrafia 19](#_Toc87351011)

# 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 influência 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.

Uma imagem com mesa

Descrição gerada automaticamente

Figura 1

# 3-Data Preparation: Limpeza

Começámos 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.

Uma imagem com mesa

Descrição gerada automaticamenteUma imagem com texto

Descrição gerada automaticamente

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.

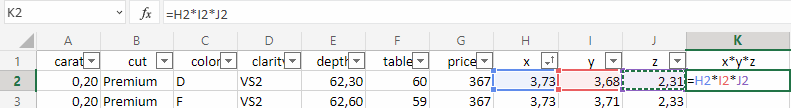


Figura 3 – criação variável x\*y\* z

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.

Uma imagem com texto

Descrição gerada automaticamenteUma imagem com mesa

Descrição gerada automaticamente

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.

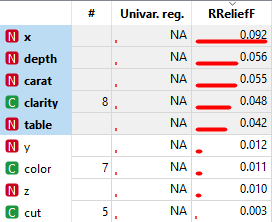


Figura 6 – “Rank”

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

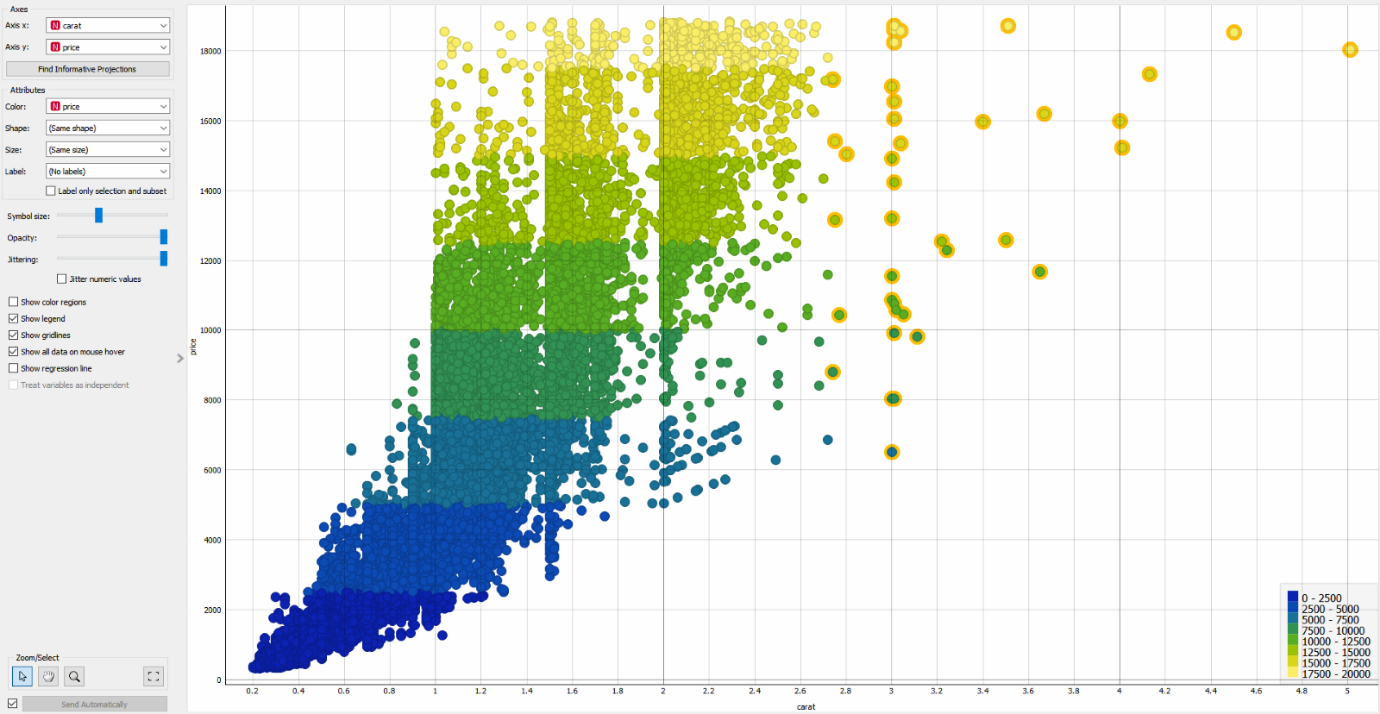


Figura 7 - “Scatter Plot”

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, selecionámos esses casos e fomos explorá-los na “Data Table”.

Uma imagem com mesa

Descrição gerada automaticamente

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.

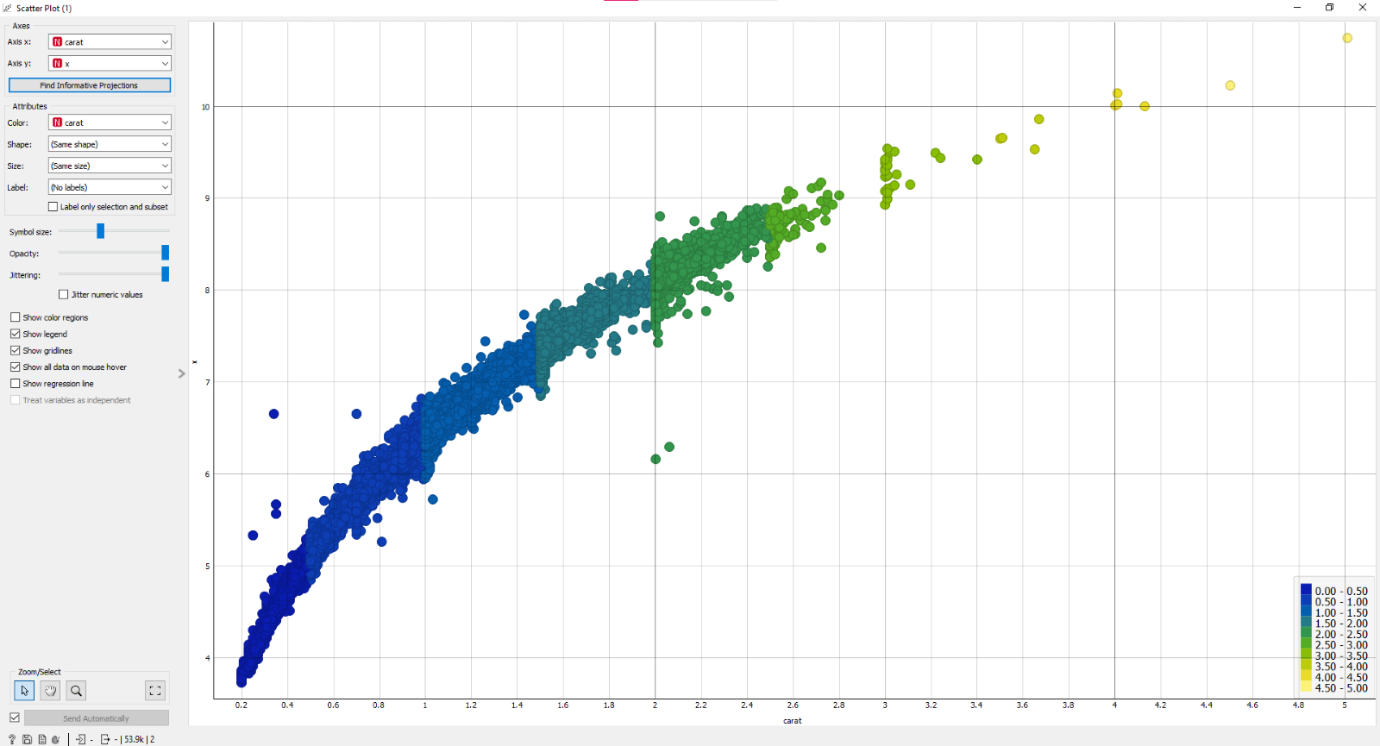
Uma imagem com mesa

Descrição gerada automaticamente

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

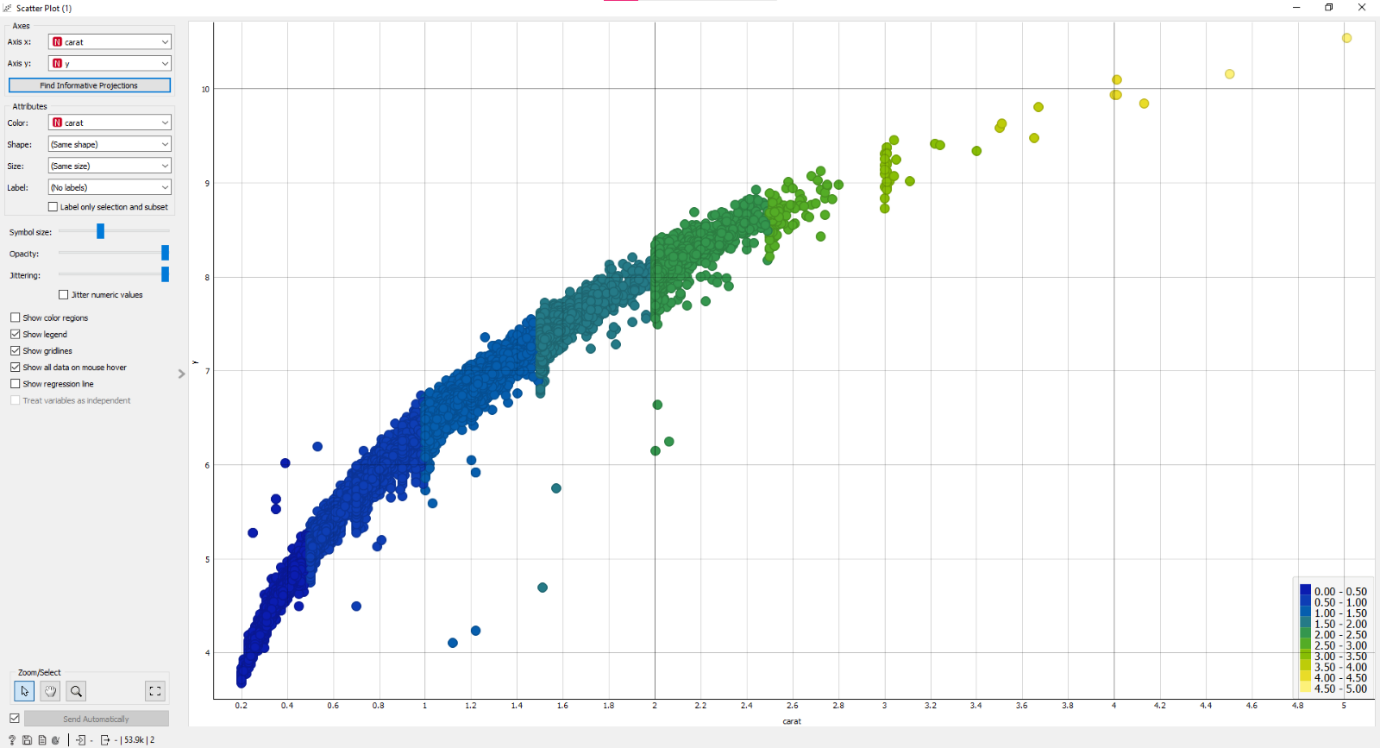


Figura 11

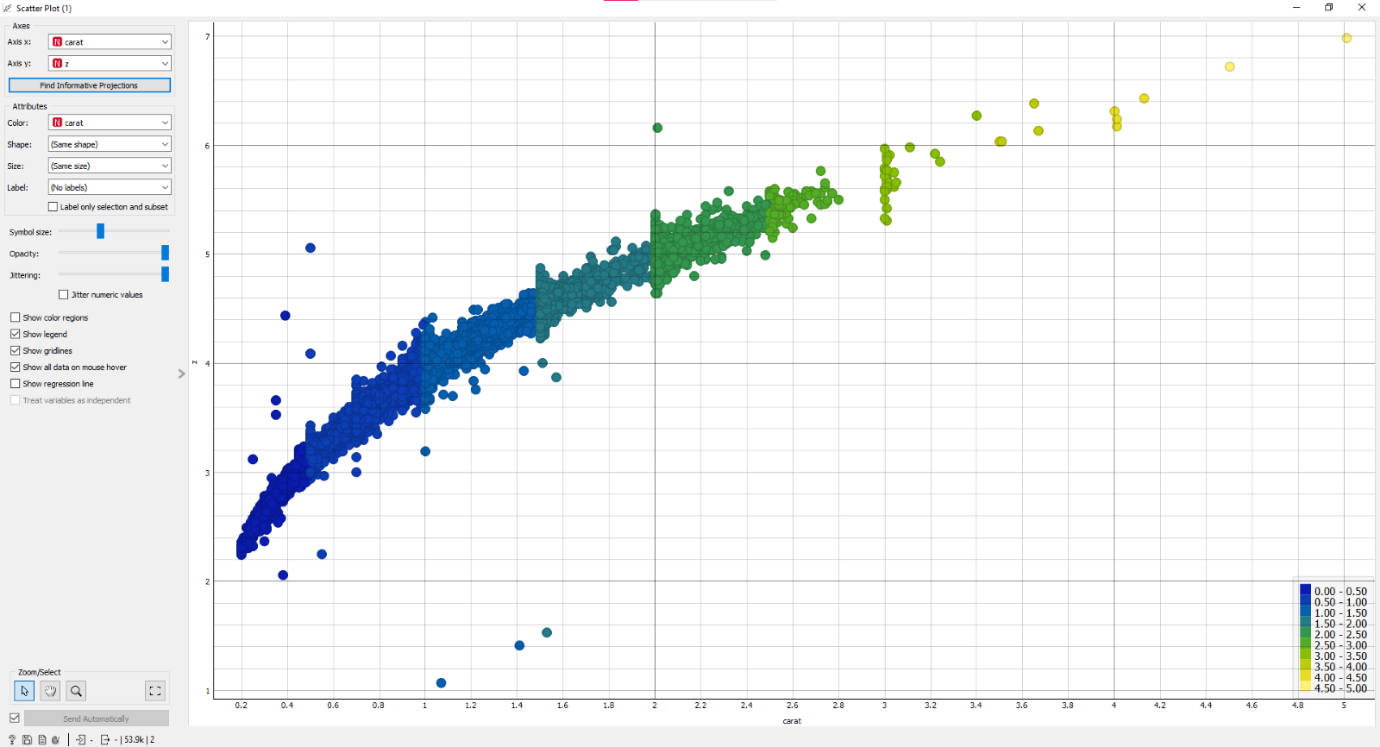


Figura 12

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.

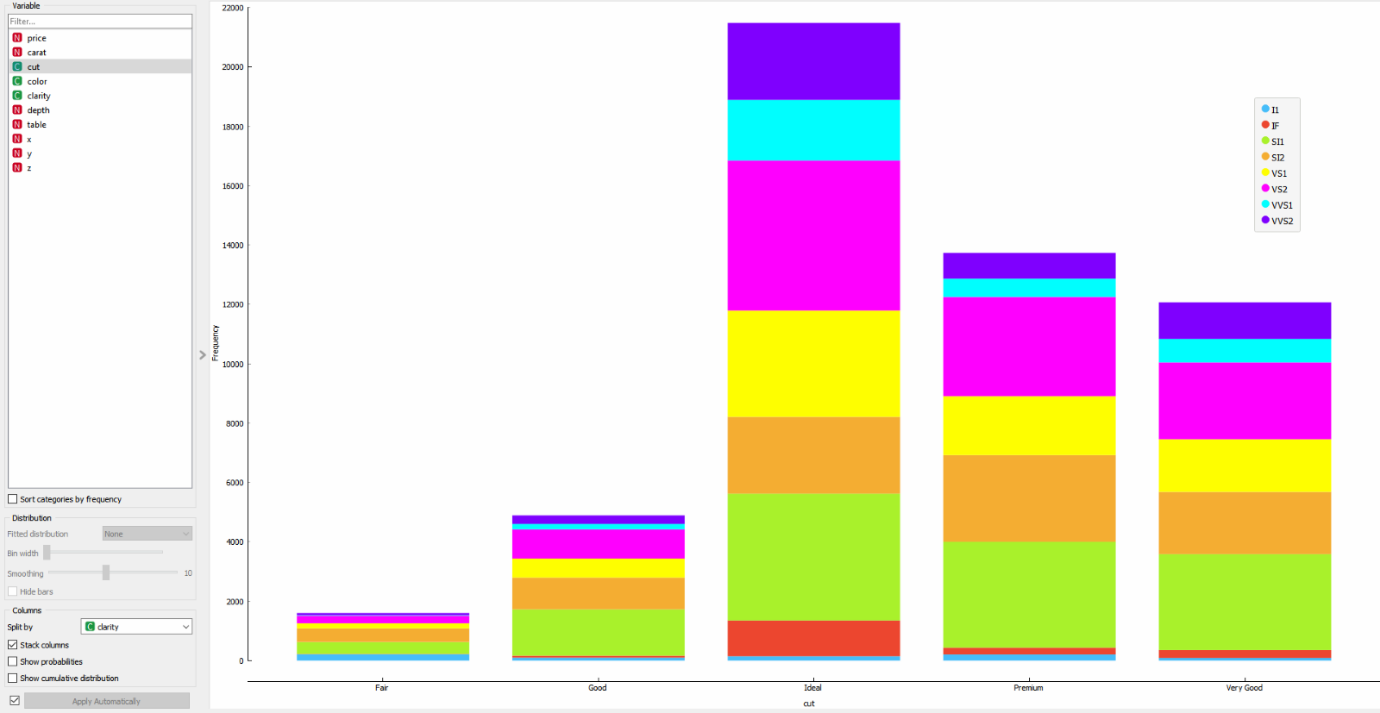


Figura 13 – “Cut”

Uma imagem com texto, captura de ecrã

Descrição gerada automaticamente

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.

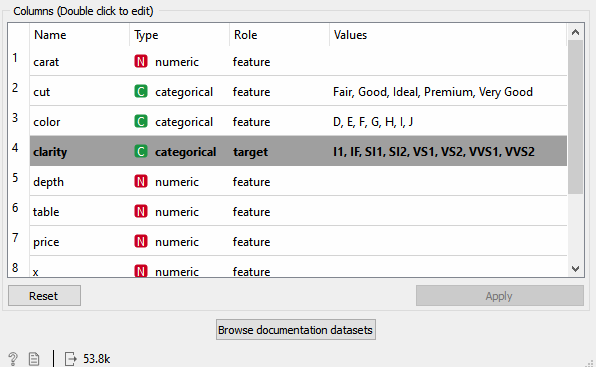


Figura 15 – Target: “clarity”

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.

Uma imagem com mesa

Descrição gerada automaticamente

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.

Uma imagem com texto

Descrição gerada automaticamente

Figura 17

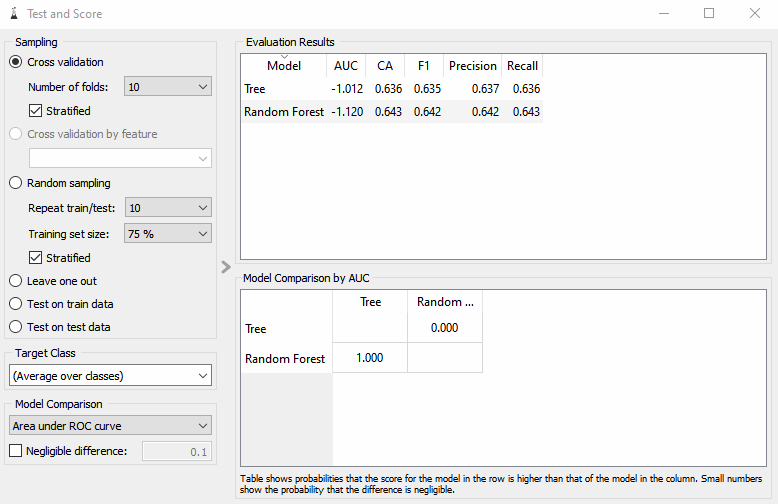


Figura 18

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 :

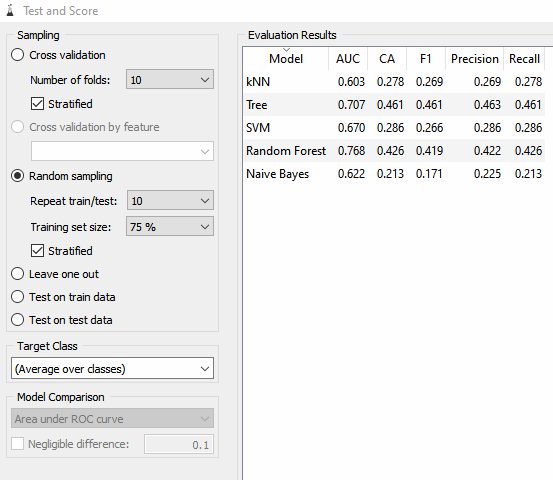
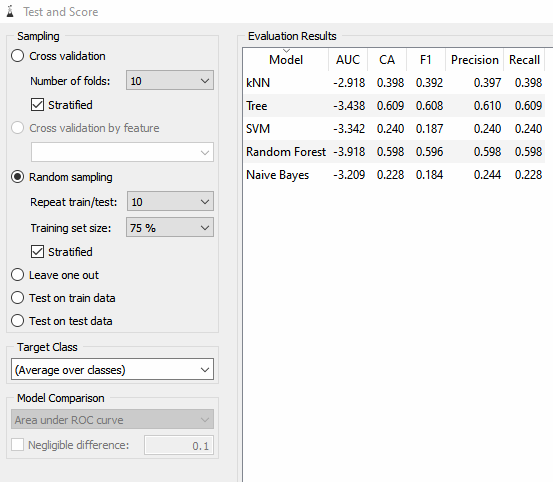


Figura 19

### 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.

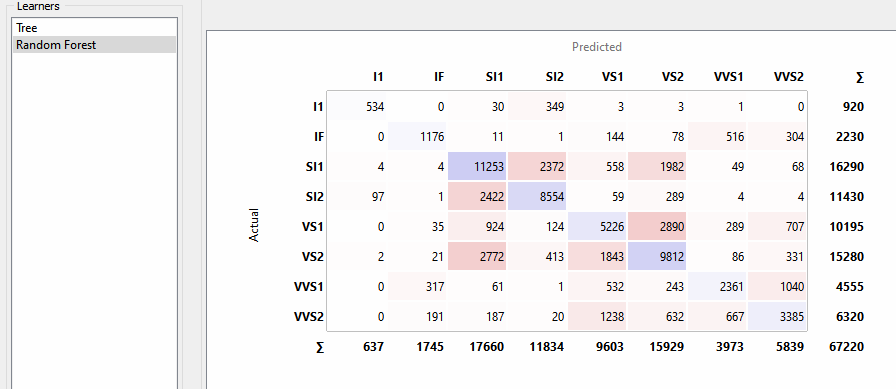


Figura 20 – “Confusion Matrix” com o “random forest”

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 errar.

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.

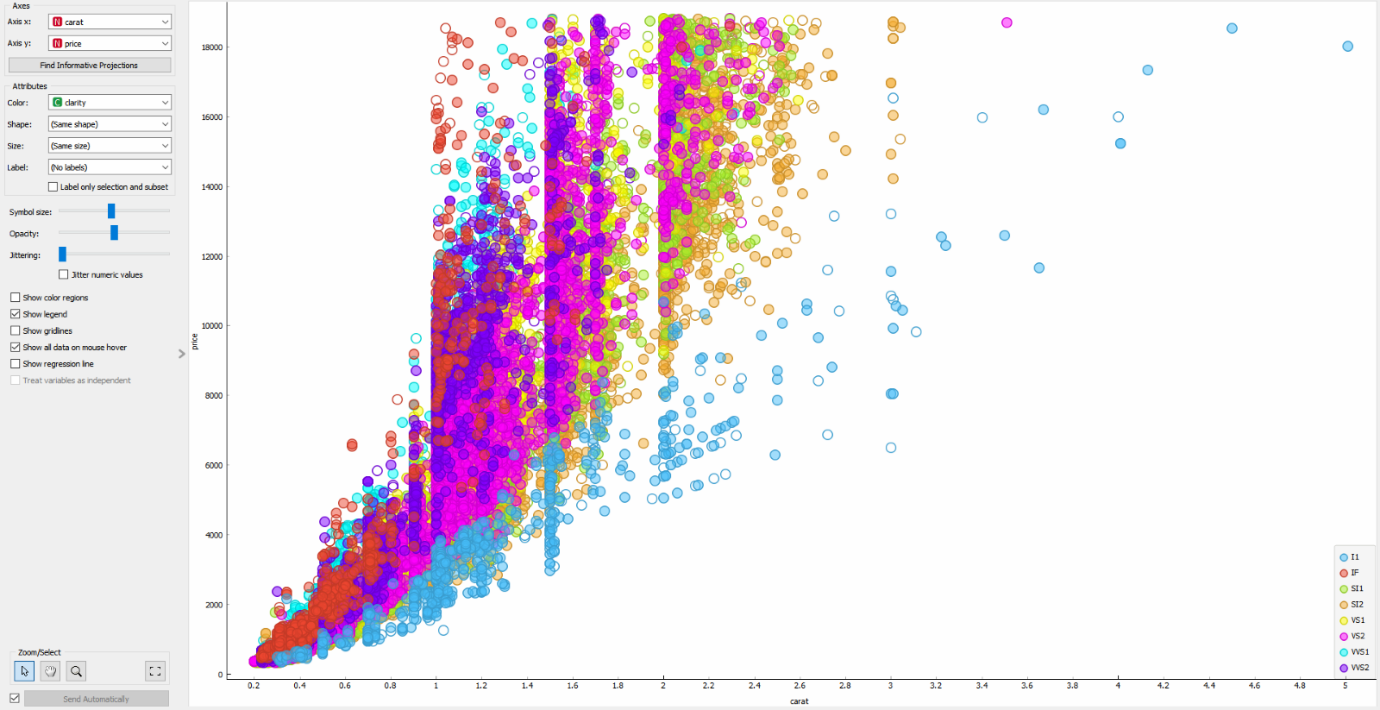


Figura 21 – “Scatter Plot”

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



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%,

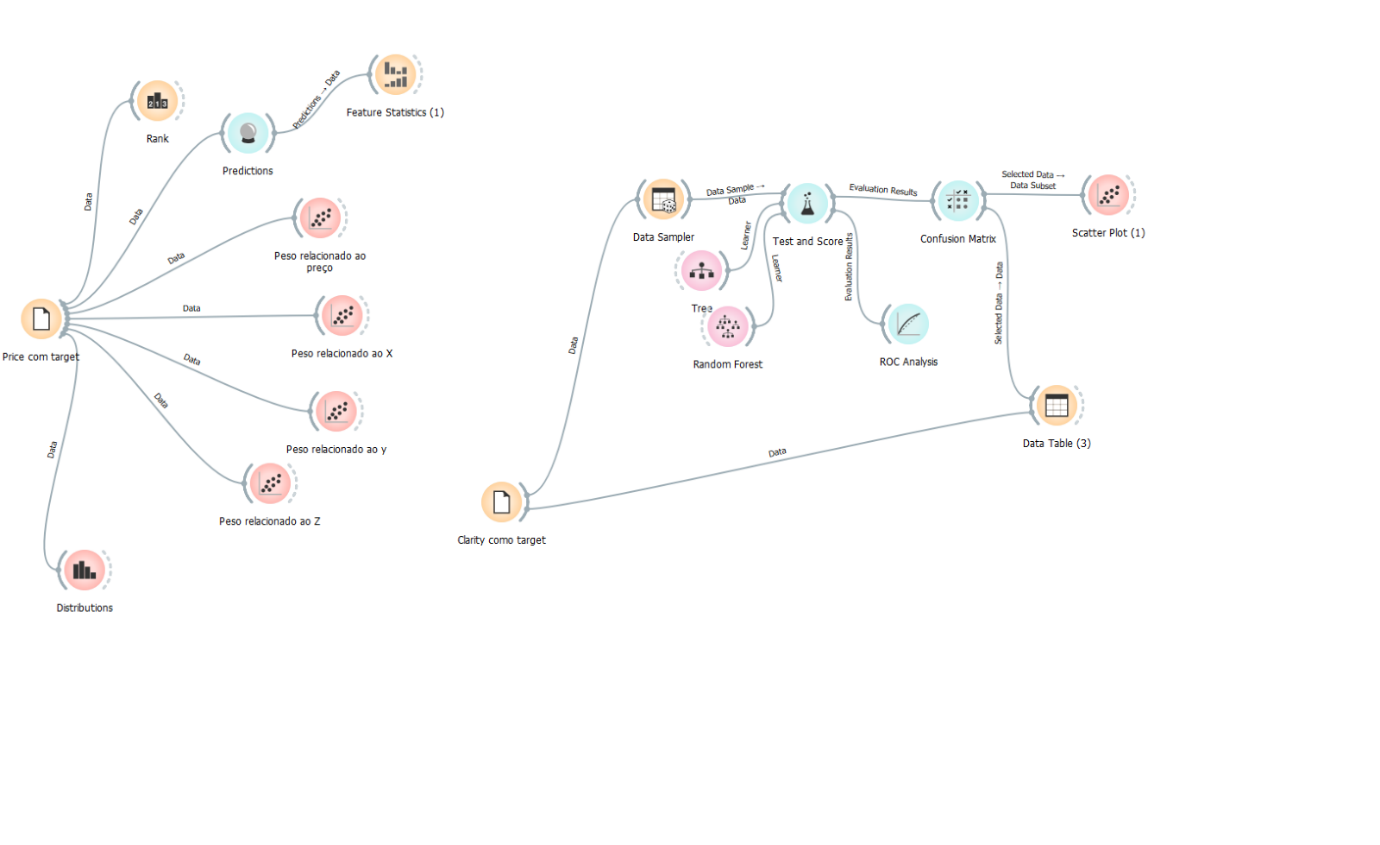
# 5-Evaluation: Avaliação e interpretação

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 dõs 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