Introdução à Ciência de Dados 2021-2022

Grupo A8

14 de novembro de 2021

—

77982 – João Raminhas

82836 – Renato Rosa

87361 – Guilherme Antunes

Índice

"Let the Dataset change your Mindset"

- Hans Ronsling

Introdução

O presente relatório insere-se no âmbito da unidade curricular de Introdução à Ciência de Dados. Descreve um processo de estudo de um conjunto de dados– *dataset*– através da metodologia CRISP-DM.

O conjunto de dados a que faz referência este trabalho, trata-se de uma listagem de computadores portáteis, com especificações e dados relevantes para apoio a uma decisão de compra.

Numa primeira abordagem, é analisado o contexto dos dados, avaliando a sua origem de forma a perceber melhor como são alimentados – *Business Understanding.* Esta abordagem foi relativamente célere, pois tratando-se de uma lista de especificações de computadores portáteis não foi considerado relevante conhecer muito mais sobre a origem desta informação.

Passando à segunda fase do método – *Data Understanding* –, considerou-se necessário manipular e reestruturar a informação, pelo facto de existirem muitos dados agregados que, aparentemente, seriam úteis para inclusão na análise analítica. Aqui, instintivamente, começaram a surgir questões relevantes para um processo de decisão de compra.

No passo seguinte – *Data Preparation* –, procedeu-se à correção e limpeza dos dados, desagregação de algumas colunas, com o objetivo de gerar mais e melhor informação para ser analisada.

Começando a explorar e a manipular os dados, foram surgindo novas ideias que forçaram várias vezes à reestruturação dos dados, pelo passo anterior. Pela modelação dos dados – *Modeling* –, foram adquiridas novas interpretações, que, surpreendentemente geraram respostas, a questões que anteriormente não haviam sido consideradas.

Finalmente, consolidando toda informação gerada pela modelação – *Evaluation* –, alcançaram-se valores concretos de extrema relevância para o objetivo pretendido.

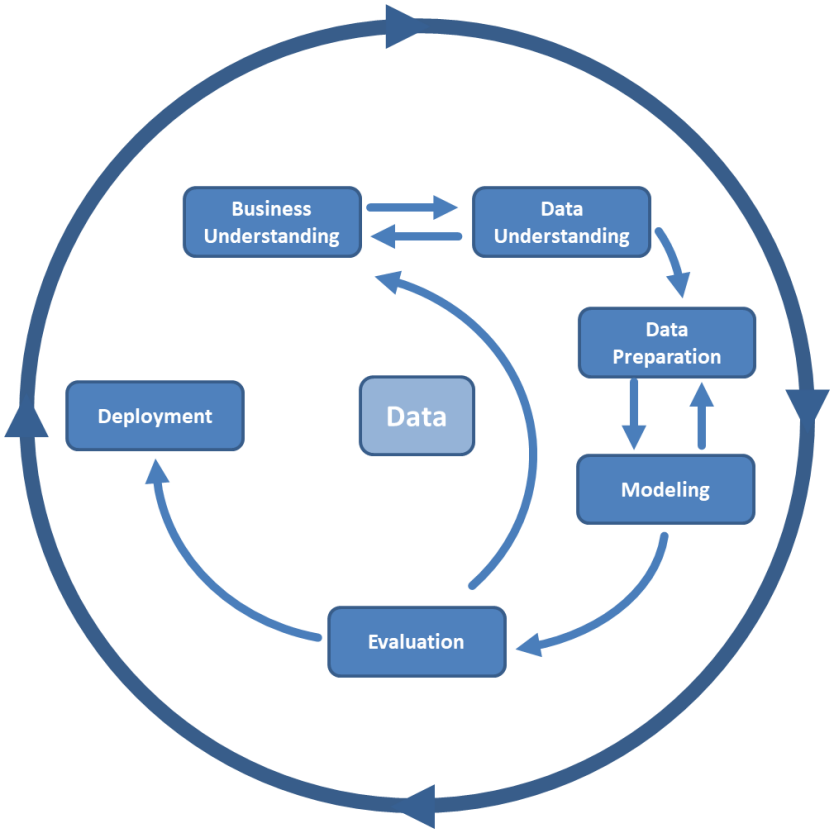
Business Understanding

Conforme constatado no PDF anexo ao dataset recebido, os computadores portáteis são considerados ferramentas de trabalho essenciais a várias atividades laborais, conjugando portabilidade e performance num dispositivo apenas.

No entanto, a vasta oferta muitas vezes origina confusão ao comprador, que poderá desconhecer ou não compreender a utilidade de cada especificação e como tirar proveito das mesmas para a sua atividade.

O presente estudo dos dados recebidos tem como objetivo organizar a informação de forma coerente e proporcionar ao comprador uma decisão consciente com base em estatísticas e métodos de análise, focando-se essencialmente nas características – *features* – que mais causam impacto no preço do computador.

Mediante variáveis que o utilizador tenha como garantidas - preço máximo, tipo de computador, peso, etc. - irá ser possível auxiliar a escolha do melhor dispositivo.



Data Understanding

1. Dados iniciais

O dataset em análise contém especificações de 1303 computadores portáteis. São apresentadas características com informação relevante, no entanto, a forma como são apresentadas requer alguma manipulação, nomeadamente pela extração de valores numéricos de colunas que apresentam as correspondentes unidades de medida, para que melhor possam representar padrões, tendências, etc.

Seguidamente descrevem-se as colunas originalmente existentes no dataset, e correspondentes informações que podem ser retiradas das mesmas:

Manufacturer	A marca que produz o dispositivo.
Model Name	O nome do modelo em questão.
Category	Categoria de computador atribuída pela marca de acordo com a finalidade de uso. (Notebook, Gaming, etc.)
Screen Size	Dimensões de ecrã, em polegadas.
Screen	Especificação de características do PC (Retina Display, IPS, etc.) e resolução de pixels.
CPU	Processador presente no computador. Contém informação sobre o seu fabricante, modelo e a sua frequência (em GHz).
RAM	Quantidade de memória RAM acoplada.
Storage	Capacidade de armazenamento do Portátil. Contém também especificação sobre o tipo: *Hard-Drive Disk (HDD)*, *Flash Storage* e *Solid-State Drives (SSD)*.
GPU	Placa Gráfica instalada no Computador.
Operating System	Sistema Operativo instalado na máquina.
Operating System Version	Versão do Sistema Operativo instalado na máquina.
Weight	Peso em Kilogramas do Computador.
Price (Euros)	Preço em Euros do Computador.

1. Limpeza e *Data Mining*

Neste âmbito da preparação dos dados, procedeu-se à correção, limpeza e transformação dos dados, por forma a tirar partido do máximo de informação possível.

O motor Power Query da Microsoft, foi o eleito para esta fase. Trata-se de uma tecnologia que permite o acesso e manipulação de dados através de várias fontes. Vem integrada em várias soluções da Microsoft, como Microsoft Excel e Power BI.

Optou-se pela utilização do Power BI Desktop, por permitir a manipulação de dados através de Scripting (Python e R), o que foi bastante útil, pois auxiliou a obtenção de padrões de informação, beneficiando do poder das expressões regulares.

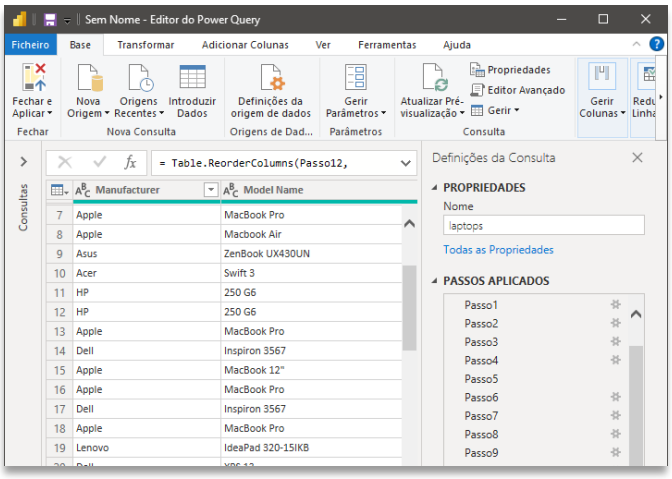


Figura 1 - Power Query Editor do Power BI Desktop

O Power Query Editor, permite, por interface gráfica, proceder à preparação dos dados através de "passos" que vão sendo registados na forma de script (Anexo 1).

Apresentam-se resumidamente, os mesmos:

1. Divisão da coluna [CPU], obtendo a frequência [CPU Freq]
2. Divisão da coluna [CPU], obtendo o tipo de fabricante [CPU Manufacturer]
3. (Script) Extração da expressão da resolução da coluna [Screen] e criação de uma coluna [Pixels], com o valor da multiplicação dos parâmetros da resolução.
4. (Script) Criação de uma coluna [Touchscreen] (Boolean), com base na coluna [Screen]
5. Divisão da coluna [Storage] em disco primário e secundário [Storage Prim] e [Storage Sec]

6. (Script) Extração do padrão numérico da coluna [Weight], omitindo as unidades (ip. lit. 'Kg' e 'Kgs').
7. Remoção da expressão 'GHz' da coluna [CPU], por forma a criar valores numéricos
8. Remoção da expressão 'GB' da coluna [RAM], por forma a criar valores numéricos
9. Remoção das aspas da coluna [Screen Size], por forma a criar valores numéricos
10. Divisão da coluna [Storage Prim] em Capacidade e Tipo, [Storage Prim] e [Storage Type]
11. Substituição de ',' por '.' na coluna [Price (Euros)]
12. Remoção de espaços adicionais nos valores das colunas modificadas

A divisão das colunas foi realizada ao verificar que existia demasiada informação em determinadas categorias que, mediante divisão, poderiam ser organizadas de melhor forma e contribuir positivamente para o presente estudo.

Durante as manipulações aos dados, foram identificados pequenos erros que se optou por serem corrigidos manualmente, permitindo aproveitar a informação, não exigindo grande esforço no seu tratamento.

Um caso tratou-se de uma linha em que um dos campos (Storage), foi omitido, levando a que os valores das colunas à esquerda tomassem os valores das colunas seguintes.

RAM	Storage	GPU	Operating System	Operating System Ve...	Weight	Price (Euros)
8GB	AMD Radeon RX 550	Windows	10	2.1kg	1144,50	
8GB	256GB SSD	Nvidia Quadro M620M	Windows	10	3.4kg	2999,00
8GB	256GB SSD	Nvidia Quadro M620	Windows	10	2.06kg	1763,00
16GB	256GB SSD	Nvidia Quadro M620	Windows	10	2.17kg	1975,00
8GB	1TB HDD	Nvidia Quadro M620	Windows	10	2.23kg	1778,00

A solução passou por reconsiderar o campo Storage nessa linha, ainda que o seu valor se tornasse omisso:

```
(…) ,8GB,AMD Radeon RX 550,Windows,10,2.1kg,"1144,50"
```

… para …

```
(…) ,8GB,,AMD Radeon RX 550,Windows,10,2.1kg,"1144,50"
```

Em relação aos restantes valores omissos, que se verificaram proporcionalmente em pouca quantidade – 16 linhas – foram mantidos e deixada essa avaliação para a modelação das variáveis em estudo. Consoante a análise específica, tratar-se-ão os mesmos, sendo a única exceção a esta regra as entradas que não possuam valor referente à feature *Manufacturer, as quais serão desprezadas.*

Para as fases que se seguem, utilizámos a aplicação *Orange* para visualizar estatísticas e proceder à análise de dados, o seu tratamento e respectiva estruturação. O *Orange* é uma ferramenta de visualização de dados, *machine learning* e *data mining*. Possui um formato de *front-end* de programação visual (*drag and drop*) para análise exploratória de dados qualitativa rápida e visualização de dados interativa.

A nível da transformação dos dados, conforme referido anteriormente, foi efetuada a alteração de níveis de variáveis categóricas para valores numéricos, de forma a conseguir apurar valores de correlação e aferir a linearidade das entradas.

Através do método de tentativa e erro e após discussão dos resultados, o grupo concluiu que não seria benéfico, mas sim um fator dificultador, incluir normalização dos valores de qualquer variável na visualização de dados.

2. Perguntas

Inicialmente, após uma análise superficial dos dados obtidos, gerámos uma pequena lista de perguntas, como ponto de partida, que nos pareceram ser interessantes para o problema apresentado. Nessa lista constam as questões:

Existirá relação entre o fabricante *Apple* e o preço?

Existirá relação entre as categorias e peso?

Qual será o componente que mais influencia o preço?

3. Data Visualization:

Feature Statistics & Distributions:

A utilização dos widgets *Feature Statistics e Distributions* foi pertinente para uma análise geral das variáveis. Através destes, deparámo-nos com algumas variáveis que tinham uma distribuição demasiado unilateral ou desequilibrada e que, portanto, não eram elegíveis para uma análise mais profunda. Assim, excluímos as seguintes variáveis: *TouchScreen, CPU Manufacturer, Storage Sec* e *Storage Prim Type*:

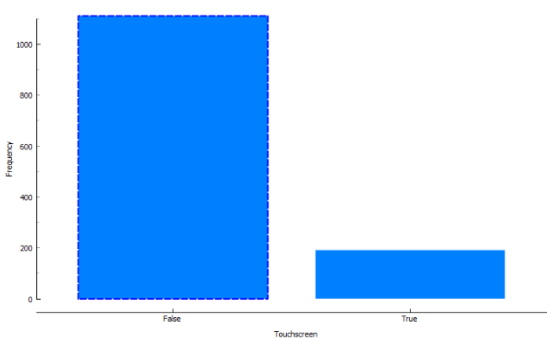


Figura 2 – Distributions TouchScreen

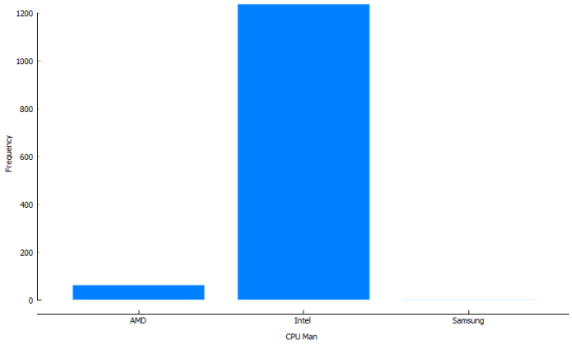


Figura 3 - CPU Manufacturer

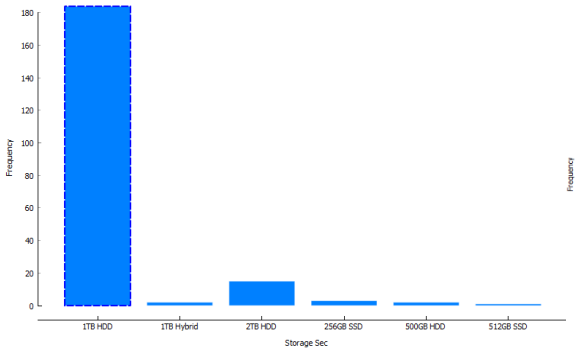


Figura 4 - Secondary Storage

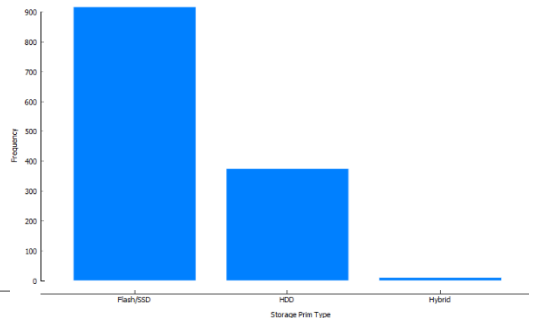


Figura 5 - Storage Primary Type

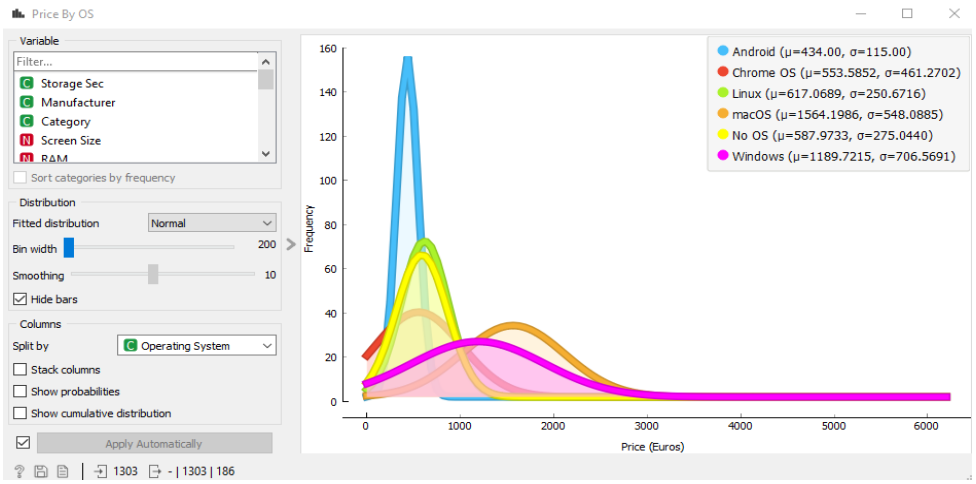


Figura 6 - Widget Box Plot

Através da Figura 6, que relaciona a frequência de amostragem com o preço por Sistema Operativo, denota-se uma maior amostragem de Laptops Android, sendo que a distribuição de preço revela um maior valor para o MacOS, da marca Apple

Correlations:

Após a análise inicial, procedemos à visualização das correlações entre as variáveis numéricas elegíveis, de maneira a perceber quais destas deveríamos incluir numa análise mais profunda.

Retirámos valor das seguintes correlações:

1	+0.826	Screen Size	Weight
2	+0.742	Price	RAM
3	+0.515	Pixels	Price
4	+0.430	CPU Freq	Price

Figura 2 Widget Correlations

Como foi possível observar, a correlação com maior valor (r = 0.83) é entre Screen Size e Weight, o que nos leva a concluir que o ecrã é uma das componentes que mais influência o peso de um portátil. É também notório o peso que três variáveis numéricas têm sobre a variável Price, sendo a variável RAM a mais influenciadora com r = 0.74.

Rank:

Através do widget Rank, é possível ver qual o acrescento de informação das variáveis categóricas, sendo que as categorias de Modelo de CPU, Manufacturer são as que apresentam um maior valor de ganho de informação.

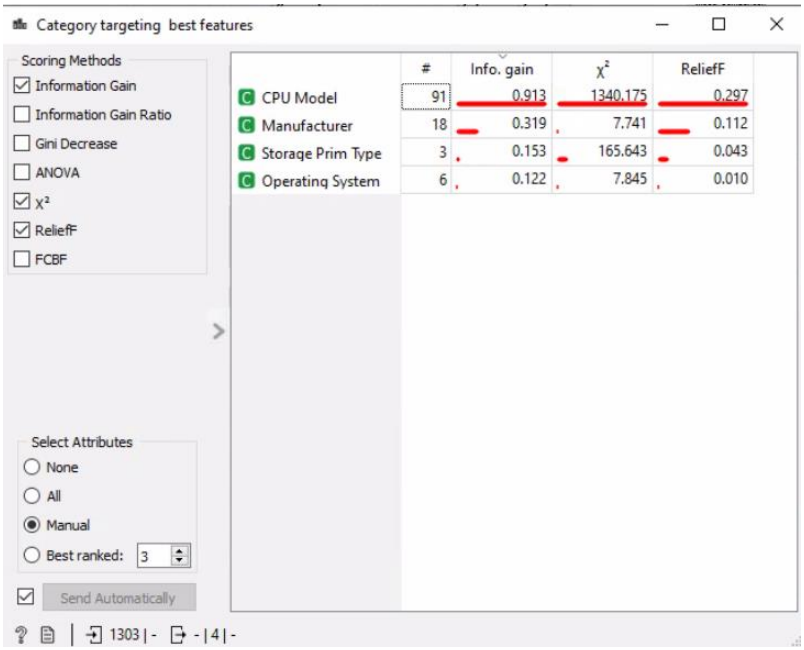


Figura 3 Widget Rank

Scatter Plot:

Para tentar uma abordagem mais visual das correlações, recorremos ao widget Scatter Plot, do qual retirámos os gráficos das duas correlações mais fortes:

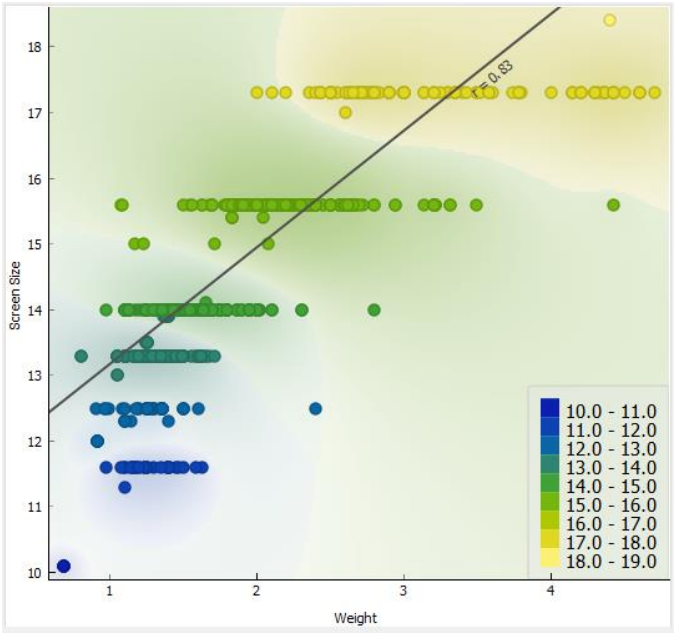


Figura 9 - Widget Scatter Plot

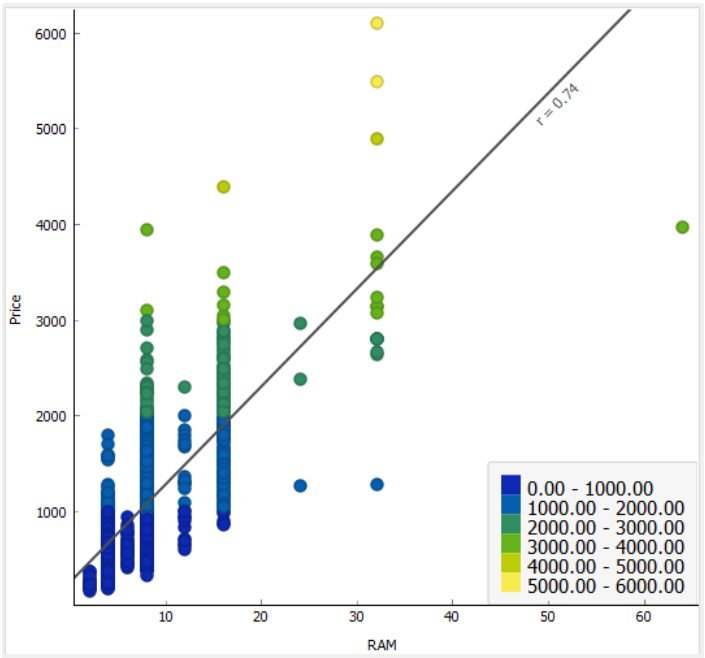


Figura 10 - Widget Scatter Plot

Violin Plot:

Este gráfico permite a agregação em grupos no eixo X e Y, resultado numa visualização ordenada de uma dada categoria mediante uma subcategoria.

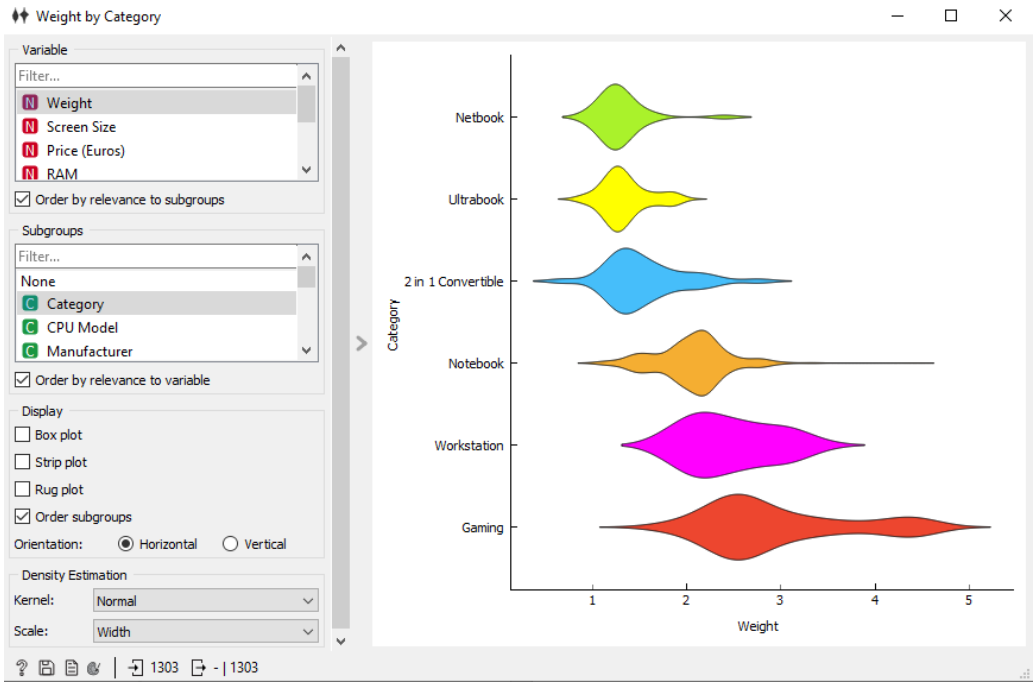


Figura 11 - Widget Violin-Plot

Mediante análise da Figura 11, é percetível a clara influência de tipo de computador no peso associado. Os Netbooks, computadores desenhados para serem facilmente transportáveis e pequenos, apresentam o menor valor de peso. Contrariamente, verificamos que um computador Gaming, que agrega características de maior performance e desempenho, apresentam os maiores valores de peso.

Box Plot:

Esta visualização permite verificar os intervalos de Quartis da amostra bem como a comparação dos valores de médias e medianas. Providência também informação gráfica de valor sobre Outliers.

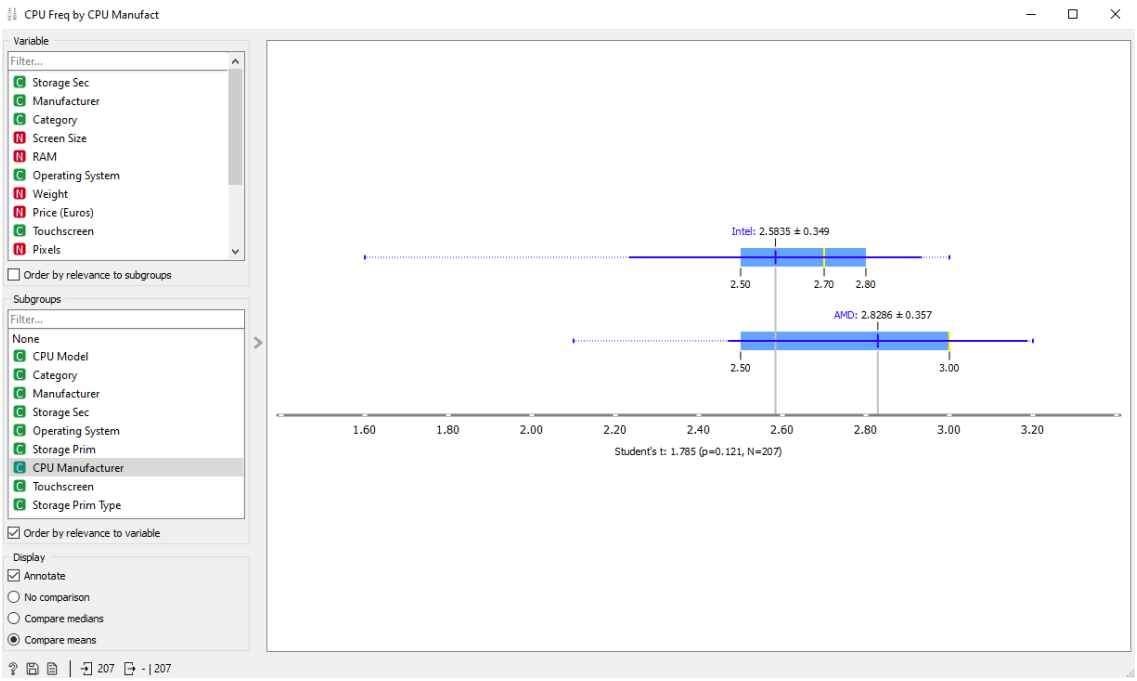
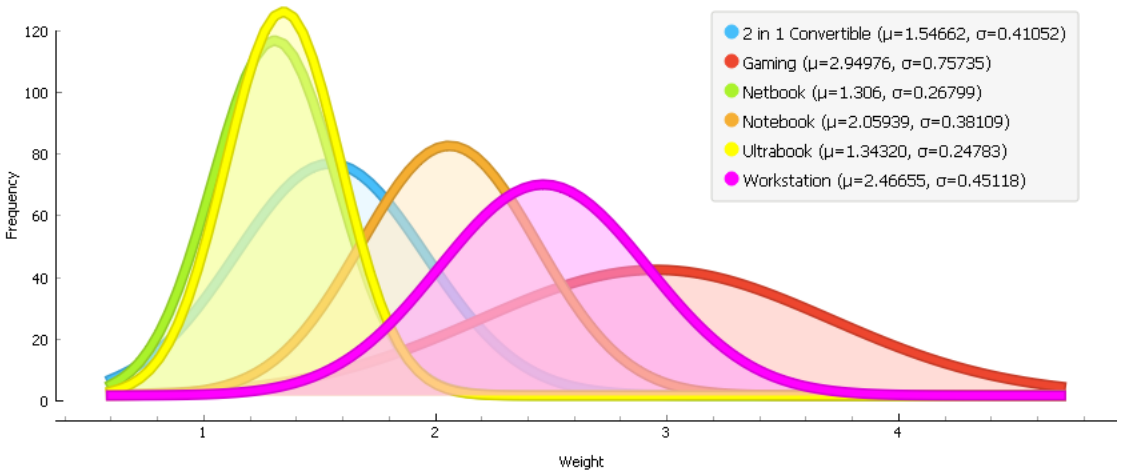


Figura 12 - Widget Box Plot

Na imagem é possível verificar os Quartis de frequência de CPU por Manufacturer, mostrando a comparação da respetiva média, bem como os outliers de cada uma das categorias. Podemos perceber a existência de um maior número de valores dispares no caso da marca Intel.

4. Respostas às Perguntas:

Após a fase de Data Visualization, a primeira questão foi imediatamente descartada visto que apenas um número muito reduzido de elementos da amostra possuía fabricante (Manufacturer) Apple (21 em 1289, cerca de 1.55%).



Atentando à segunda questão, e mediante o gráfico apresentado de distribuições de peso por categoria, é possível denotar que os computadores com maior peso, em média, são Gaming. Já os Netbooks e Ultrabooks, concebidos para providenciar uma maior mobilidade, apresentam valores de peso bastante menores.

Em relação à terceira questão, esta foi facilmente respondida, visto que, como mencionado anteriormente, o preço é a variável com o maior número de correlações pertinentes. Assim sendo, constámos que o componente (dos considerados elegíveis a estudo) que mais influencia o preço de um portátil é a memória RAM.

Modeling

Após a análise da informação nas fases anteriores, foram idealizadas duas modelações possíveis, uma delas com base nas categorias, outra com base em variáveis numéricas.

1. Variáveis numéricas

No que toca às variáveis numéricas, um desafio interessante seria perceber se, através das features numéricas mais significativas (com melhor correlação), se se conseguiriam atingir intervalos de valores de preços. Esta abordagem permitiria, por exemplo, através de três especificações desejadas pelo comprador, ter uma perceção do intervalo de valores que poderia esperar.

Uma vez que o preço do computador seria dificilmente determinado de forma precisa, optou-se por dividir o preço em intervalos de frequências equivalentes, por forma a obter um modelo preditivo, capaz de categorizar desta forma os vários computadores.

Com base nas correlações mais significativas, foram consideradas as colunas [RAM], [CPU Freq.], [Pixels] e [Weight] para determinar um intervalo de preços.

2. Variáveis categóricas

Em relação às variáveis categóricas delineámos uma modelação com vista a inferir a variável Category através de algoritmos de aprendizagem conhecendo as variáveis [Storage Prim], [CPU Freq], [Operating System] e [Price]. Com esta abordagem, procuramos fazer uma recomendação de categoria de portátil consoante as preferências do comprador, em termos de capacidade de disco, sistema operativo, potência de processador, e um intervalo monetário. Achámos pertinente gerar este modelo pois a categoria de computador é uma das primeiras especificações a ser dada a um profissional no momento da procura pelo produto certo.

Evaluation

1. Variáveis numéricas

Foram usados dois métodos de avaliação: *Test & Score* e *Predictions*.

Pelo método *Test & Score*, foram usados os algoritmos de aprendizagem *Random Forest* e *Grandient Boost*. Foram alcançados alguns valores interessantes, em particular os valores AUC.

Model	AUC	CA	F1	Precision	Recall
Gradient Boosting	0.842	0.428	0.426	0.424	0.428
Random Forest	0.815	0.416	0.413	0.411	0.416

Figura 4 Test & Score: Random Sampling

Nas *Predictions,* o valor AUC manteve-se alto embora ligeiramente mais baixos.

Model	AUC	CA	F1	Precision	Recall
Random Forest	0.804	0.422	0.415	0.425	0.422
Gradient Boosting	0.818	0.429	0.430	0.438	0.429

Figura 5 Test & Score: Random Sampling

Considera-se este modelo, credível para previsão de resultados.

2. Variáveis categóricas

Neste modelo foi usado o método *Test & Score*.

Por este método foram utilizados os algoritmos de aprendizagem *Naive Baites*, *Neural Network* e *Logistic Regretion*. Este modelo provou ser consideravelmente menos eficiente do que o primeiro, uma vez que analisando a matriz de confusão verifica-se alguma incerteza no que toca a algumas categorias.

Model	AUC	CA	F1	Precision	Recall
Logistic Regression	0.872	0.706	0.661	0.674	0.706
Neural Network	0.861	0.686	0.664	0.653	0.686
Naive Bayes	0.858	0.668	0.649	0.657	0.668

Figura 6 Test & Score Random Sampling

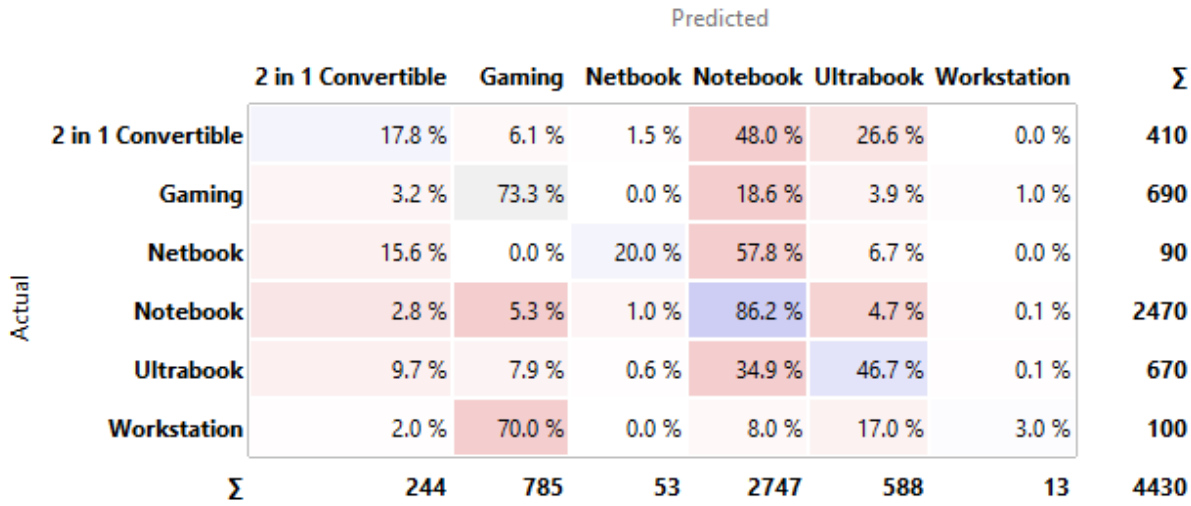


Figura 7 Matriz de confusão

Referências Bibliográficas

https://en.wikipedia.org/wiki/Orange_(software)

https://orangedatamining.com/widget-catalog/

https://docs.microsoft.com/en-us/powerquery-m/power-query-m-function-reference

https://support.microsoft.com/pt-pt/office/sobre-a-consulta-de-poder-no-excel-7104fbee-9e62-4cb9-a02e-5bfb1a6c536a

https://e-learning.iscte-iul.pt/ultra/courses/_15589_1/cl/outline

Anexo 1

Passos utilizados na ferramenta *Power Query Editor*, no programa *Power BI Desktop*:

```
let
    Origem = Csv.Document(File.Contents("F:\Users\Gui\Desktop\laptops.csv"),[Delimiter=",", Columns=13,
Encoding=1252, QuoteStyle=QuoteStyle.None]),

    CabeçalhosPromovidos = Table.PromoteHeaders(Origem, [PromoteAllScalars=true]),

    Passo1 = Table.SplitColumn(CabeçalhosPromovidos, "CPU", Splitter.SplitTextByEachDelimiter({" "},
QuoteStyle.Csv, true), {"CPU", "CPU Freq"}),

    Passo2 = Table.SplitColumn(Passo1, "CPU", Splitter.SplitTextByEachDelimiter({" "}, QuoteStyle.Csv,
false), {"CPU Manufacturer", "CPU Model"}),

    Passo3 = Python.Execute("import re#(lf)#(lf)def IsInt(s):#(lf)    try: #(lf)        int(s)#(lf)
return True#(lf)    except ValueError:#(lf)        return False#(lf)#(lf)def get_resolution(text):#(lf)
res = re.findall(r""[0-9]{3,4}x[0-9]{3,4}"", str(text))#(lf)    return
"";"".join(res)#(lf)#(lf)resolution_list=[]#(lf)for i in range(len(dataset)):#(lf)    text =
dataset.iat[i,4]#(lf)    res = get_resolution(text)#(lf)    if res != """":#(lf)        res =
res.split(""x"")#(lf)        try:#(lf)            w = res[0]#(lf)            h = res[1]#(lf)
if IsInt(w) and IsInt(h):#(lf)                resolution_list.append(int(int(w) * int(h)))#(lf)
except ValueError:#(lf)           resolution_list.append(None)#(lf)    else:#(lf)
resolution_list.append(None)#(lf)   #(lf)dataset['Pixels']=resolution_list",[dataset=Passo2]),

    Passo4 = Python.Execute(“ts_list=[]#(lf)for i in range(len(dataset)):#(lf)    text =
str(dataset.iat[i,4])#(lf)    eval = (text.find(""Touchscreen"", 0) > -1)#(lf)
ts_list.append(eval)#(lf)#(lf)dataset['Touchscreen']=ts_list",[dataset=Passo3{[Name="dataset"]}[Value]]
),

    Passo5 = Table.SplitColumn(Passo4{[Name="dataset"]}[Value], " Storage",
Splitter.SplitTextByDelimiter("+", QuoteStyle.Csv), {"Storage Prim", "Storage Sec"}),

    Passo6 = Python.Execute("import re#(lf)#(lf)def get_weight(text):#(lf)    res =
re.findall(r""\d+\.?\d*"", str(text))#(lf)    return "";"".join(res)#(lf)#(lf)weight_list=[]#(lf)for i
in range(len(dataset)):#(lf)    text = dataset.iat[i,14]#(lf)    res = get_weight(text)#(lf)
weight_list.append(res)#(lf)   #(lf)dataset['Weight']=weight_list",[dataset=Passo5]),

    Passo7 = Table.ReplaceValue(Passo6,"GHz","",Replacer.ReplaceText,{"CPU Freq"}),

    Passo8 = Table.ReplaceValue(Passo7,"GB","",Replacer.ReplaceText,{"RAM"}),

    Passo9 = Table.ReplaceValue(Passo8,""""","",Replacer.ReplaceText,{"Screen Size"}),

    Passo10 = Table.SplitColumn(Passo9, "Storage Prim", Splitter.SplitTextByEachDelimiter({" "},
QuoteStyle.Csv, false), {"Storage Prim", "Storage Prim Type"}),

    Passo11 = Table.ReplaceValue(Passo10,",",".",Replacer.ReplaceText,{"Price (Euros)"}),

    Passo12 = Table.TransformColumns(Passo11,{{"Pixels", Text.Trim, type text}, {"Storage Prim",
Text.Trim, type text}, {"Storage Prim Type", Text.Trim, type text}, {"Storage Sec", Text.Trim, type
text}, {"RAM", Text.Trim, type text}, {"CPU Freq", Text.Trim, type text}, {"CPU Manufacturer",
Text.Trim, type text}, {"CPU Model", Text.Trim, type text}, {"Screen Size", Text.Trim, type text}}),

    ColunasReordenadas = Table.ReorderColumns(Passo12,{"Manufacturer", "Model Name", "Category",
"Screen Size", "Screen", "Touchscreen", "Pixels", "CPU Manufacturer", "CPU Model", "CPU Freq", "RAM",
"Storage Prim", "Storage Prim Type", "Storage Sec", "GPU", "Operating System", "Operating System
Version", "Weight", "Price (Euros)"})
in
    ColunasReordenadas
```