INSTITUTO UNIVERSITÁRIO DE LISBOA

Ano Letivo 2021/2022

10 de novembro de 2021

Dados de Automóveis

UC Dados na Ciência, Gestão e Sociedade

Licenciatura Ciência de Dados

Grupo 13, CDA1

Docentes
Ana Maria de Almeida
Elsa Alexandra Cardoso
José Miguel Dias
Nuno Alexandre Alves

André Silvestre Nº104532 | Diogo Catarino Nº104745
Eduardo Silva Nº104943 | Francisco Gomes Nº104944

Índice

Página

Sumário do Projeto

Este relatório foi proposto no âmbito da UC de *Dados na Ciência, Gestão e Sociedade* inserida na licenciatura de Ciência de Dados, lecionada pelos docentes Ana Maria de Almeida, Elsa Cardoso, José Dias e Nuno Alves, no ISCTE, em novembro de 2021, e pretende, após uma limpeza de dados, analisar, criticamente, tanto as correlações e causalidades entre as variáveis, como as estatísticas descritivas destas e as visualizações gráficas construídas posteriormente.

Pretende-se, igualmente, usar algoritmos de aprendizagem supervisionada e/ou *text mining* sobre o conjunto de dados com as variáveis escolhidas para, deste modo, constituir o *dataset*.

O conjunto de dados fornecidos provieram de um *dataset* público, e com estes, podemos aplicar os conhecimentos lecionados em aula, isto é, utilizar a metodologia CRISP-DM num caso real.

Assim, pretende-se com este trabalho, com o auxílio do programa *Open Source* utilizado para visualização de dados, *Machine Learning* e *Data Mining*, *Orange*, compreender, preparar, limpar, explorar e criar informação utilizando técnicas de aprendizagem supervisionada de classificação e de regressão.

Introdução

História dos carros nas décadas de 70 e 80

Os carros começaram a ser usados, mundialmente, a partir do século XX. As décadas de 70/80 ficaram marcadas pela modernização das linhas dos automóveis, que passaram a ser mais quadrados e funcionais.

Este período histórico ficou conhecido por duas crises de petróleo, que culminaram com um período de escassez de combustível, no qual, os carros japoneses brilharam no mercado americano, com modelos pequenos, baratos e económicos. Marcas como *Toyota*, *Honda* e *Datsun* cresceram rapidamente em vendas, superando as rivais alemãs, francesas e americanas.

Na Europa, os automóveis começaram a ficar muito modernos, com a utilização de injeção eletrónica, travões ABS e outras funcionalidades, produzindo carros mais baratos e mais seguros. O tamanho aumentou ligeiramente, e alteraram a transição do ar para a água, permitindo, a marcas como a *Volkswagen*, o desenvolvimento de carros potentes.

A economia com a crise do petróleo não era um problema para os europeus, uma vez que já estavam habituados à economia pós-guerra. A *Fiat* foi outra das marcas que dominou carros pequenos, refletindo-se também em Portugal com o modelo *Fiat 127,* que alcançou o recorde italiano de unidades produzidas de 3,8 milhões.

Contexto dos dados fornecidos

Os dados que foram utilizados para o desenvolvimento deste relatório e das suas respetivas conclusões, foram retirados da *Statlib Library*, um website disponibilizado pela *Carnegie Mellon University*.

Assim, com este trabalho procuramos compreender de que modo o consumo dos automóveis, expresso em *MPG*, está relacionado com as características intrínsecas aos mesmos.

Metodologia

Metodologia CRISP-DM

A metodologia utilizada neste projeto, CRISP-DM, é formada por seis fases: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation* e *Deployment*.

Primeiramente, na fase do *Business Understanding,* o objetivo é perceber qual o problema que queremos desenvolver e resolver. Para isso, precisamos de determinar a questão e objetivo do projeto, de modo a planear a sua execução.

A fase de *Data Understanding*, é onde nos focamos na compreensão dos dados que temos para responder ao problema anteriormente determinado. Nesta fase temos de recolher, descrever e explorar os dados, bem como verificar a sua qualidade.

Após compreender os dados que temos, passamos à próxima etapa, *Data Preparation*. É nesta fase onde preparamos os dados para a fase de modelação. Para tal, fazemos a seleção, limpeza, integração e formatação dos dados.

De seguida, chegamos à fase de *Modeling* na qual desenvolvemos um modelo de aprendizagem automática, quer através de técnicas de aprendizagem supervisionada (classificação e regressão), quer as não supervisionadas (*Clustering*, entre outras).

Por fim, *Evaluation* e *Deployment* são as etapas onde testamos os modelos desenvolvidos, verificamos se respondem ao problema inicial e apresentamos os resultados para posterior tomada de decisão.

1. *Business Understanding | Data Understanding*

Inicialmente, de modo a compreender o domínio a que os dados fornecidos estão inseridos, fizemos uma pesquisa do enquadramento dos mesmos, e das variáveis em estudo.

De seguida, definimos o problema que gostaríamos de resolver, em função dos dados, as variáveis que necessitamos de utilizar para o mesmo.

Conceitos das variáveis em estudo:

Ano do modelo – Ano de fabrico do modelo em causa;

Origem - As origens dos automóveis variam entre América, Europa e Japão, estando eles apresentados em 1, 2, 3 respetivamente;

MPG (milha por galão) - A milha por galão é a unidade de medida que nos foi dada relativamente aos consumos dos automóveis. Esta unidade de medida pode ser convertida para L/100 ou para Km/L (tal como apresentamos no nosso documento *Excel*) para que os dados sejam mais facilmente compreendidos;

Cilindros – Os cilindros são componentes constituintes do motor onde se deslocam os pistões e onde ocorre a mistura do combustível. Usualmente utilizado para caracterizar automóveis através da estrutura do bloco do motor (em casos onde exista mais do que 1 cilindro);

Displacement – Displacement, em português cilindrada, é extraído através da combinação do número de cilindros e da dimensão do motor. Os dados que nos foram facultados apresentam-se em *cubic inches* e mais uma vez fizemos a conversão dos dados para *centímetros cúbicos* (no documento *Excel*);

Horsepower- O horsepower de um automóvel, também conhecido como potência (ou número de cavalos) é a capacidade de um motor de executar uma tarefa num determinado tempo e também uma variável utilizada para indicar a força extraída do mesmo.

Peso- O peso de um automóvel é uma variável medida em *Pounds* na América e em Portugal em *Kilogramas* (medida para a qual convertemos os valores);

Aceleração- A aceleração de um automóvel é avaliada em segundos e designa o tempo que um veículo demora das 0 mph até às 60 mph (sendo "mph" – milhas por hora e convenciona-se que 60 mph = 100km/h). Sendo assim, a aceleração é o tempo, em segundos, que um veículo demora dos 0km/h até aos 100 km/h.

2. *Data Preparation*

O *dataset* utilizado para este projeto denominado *"auto-mpg"* tem, no total, 398 instâncias e 9 variáveis: *mpg, , displacement, horsepower* (com 6 valores ausentes)*, weight, acceleration,* como variáveis numéricas contínuas; *cylinders, model year, origin* como variáveis discretas e o *car name* como uma *string* que é único para cada instância.

O primeiro passo realizado para o desenvolvimento desta etapa foi o tratamento dos dados no *Excel*, com a divisão em colunas e a eliminação das linhas cujos valores estavam ausentes. Este procedimento teve como finalidade a preparação dos dados, para posteriormente, os exportamos para o software *Orange*, no sentido de puderem ser tratados da melhor forma, para o nosso interesse.

Após isso, criamos colunas extra no *Excel* nas quais convertemos as unidades de medida do sistema imperial, que diferem daquelas que usamos no quotidiano (sistema métrico), de modo a podermos compreender melhor o domínio dos dados fornecidos.

Ainda no *Excel*, procedemos à limpeza dos dados, eliminando as 13 linhas de um total de 406 linhas, pois apresentavam células com o termo *NA,* pelo que não tinham valor associado à variável respetiva. Tal se verificou apenas nos atributos *MPG* e *Horsepower,* porém tivemos de eliminar a linha completa pois, para efeitos de modelação, para dar resposta às questões levantadas, todas as variáveis/séries dos conjuntos de dados analisado devem ter o mesmo número de observações.

Observámos os dados no *Excel*, e visto serem apenas 9 variáveis em estudo, incluindo o *Car Name* que difere de instância para instância, decidimos utilizar todas para analisar e modelar no *Orange*, e assim poder concluir com o auxílio a este software quais é que se relacionam mais ou menos com o nosso *target* previamente escolhido, o MPG.

Ambiente *Orange*

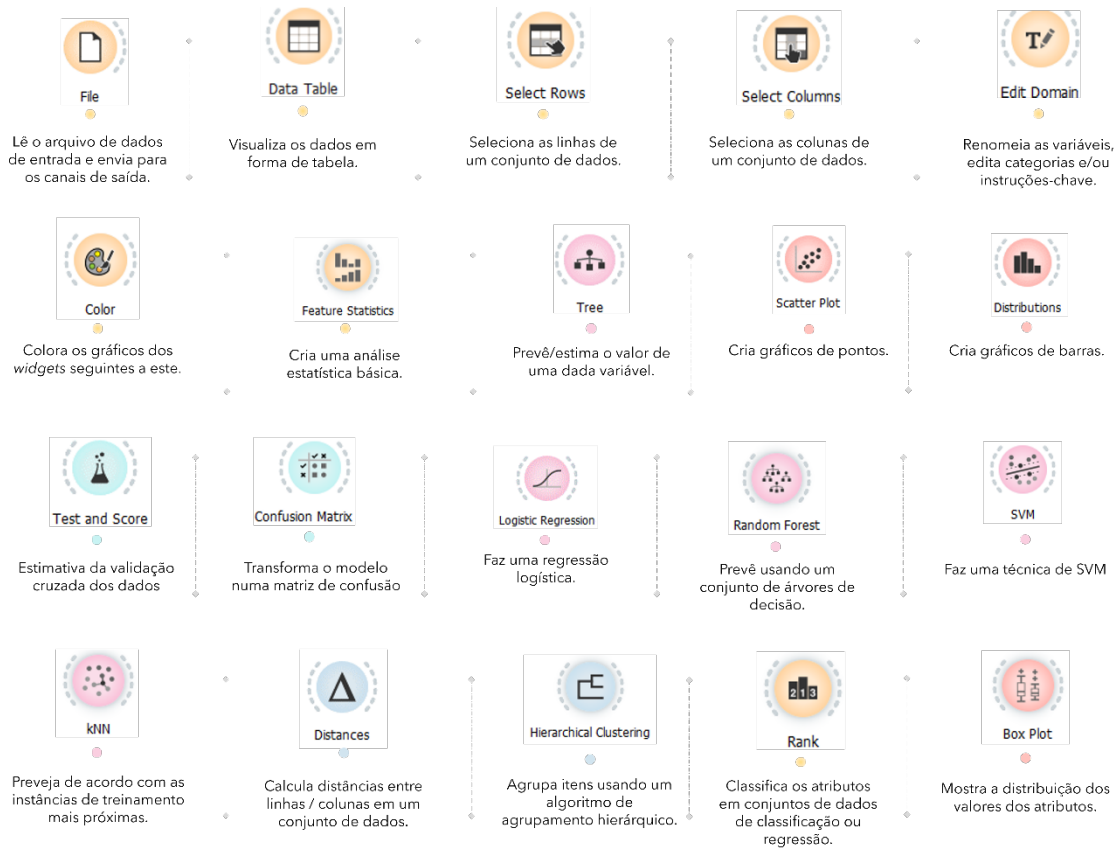

Figura 1 | Ferramentas do *Orange* e respetivas funções.

O programa *Orange* é constituído por vários ícones (ou *widgets*), tal como descritos na Figura 1, que representam uma ferramenta ou técnica que pode ser utilizada no processo de *Data Mining* e *Modeling*. É possível incluir dados ausentes, criar *features*, visualizar gráficos, aplicar modelos supervisionados, avaliar o desempenho dos mesmos, entre tantas outras funcionalidades.

Inicialmente, no *Orange,* colocámos o widget *File* de modo a importar o ficheiro *Excel* (.xlsx) previamente trabalhado. De seguida, ainda nesse widget selecionamos como *target* o *MPG*, e colocámo-lo como *categorical*, a fim de posteriormente, desenvolver o modelo para o mesmo.

Posteriormente, decidimos usar o widget *Correlations* (Fig.2), que relaciona todas as variáveis e devolve pares de variáveis, consoante o seu nível de correlação.

Através da observação do mesmo, podemos deferir que o MPG tem uma forte correlação com os atributos *Weight*, *Displacement*, *Horsepower* e *Cylinders*. Deste modo, analisámos as melhores relações que continham a nossa variável *target* e prosseguimos o trabalho.

De seguida, adicionámos os widgets *Data Table*, *Scatter Plot*, *Feature Statistics* e *Distributions*, a fim de conseguirmos visualizar de diferentes formas os dados já preparados, para assim, os podermos compreender melhor.

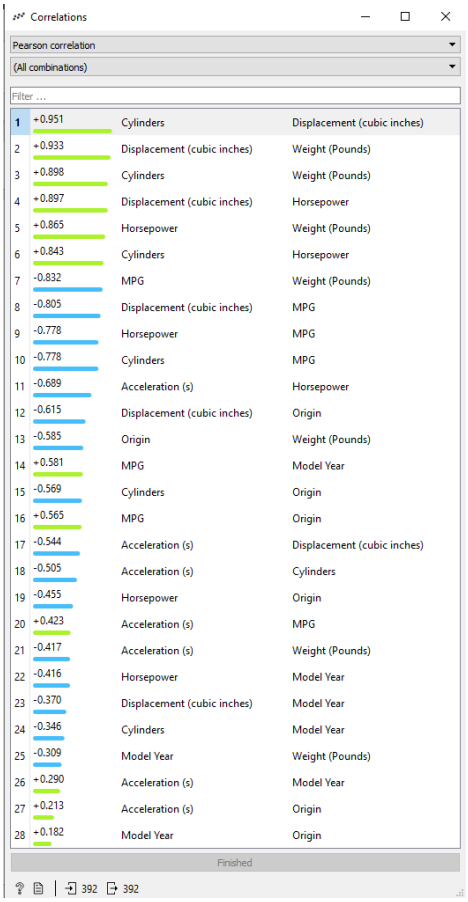

Figura 2 | Correlações das variáveis em estudo no software *Orange*.

Análise Exploratória de Dados

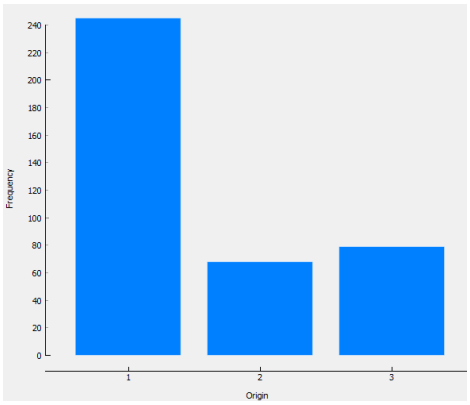

Gráfico 1 | Frequência da ***Origem***

Através da análise do **Gráfico 1**, pudemos concluir que na variável *Origem*, a maioria dos carros são americanos (1), sendo estes 245 no total. Estão também presentes 68 carros de origem europeia (2), e 79 de origem japonesa (3).

Possíveis razões que possam explicar a acentuada frequência de carros americanos são a economia mundial pós-guerra, que favorece fortemente os americanos e o seu puder de compra, face aos restantes países do mundo.

Outra razão, poderá ser apenas o facto de os dados fornecidos pertencerem a uma universidade americana, pelo que, poderá ter havido mais facilidade em recolher os dados dos automóveis refentes a este país.

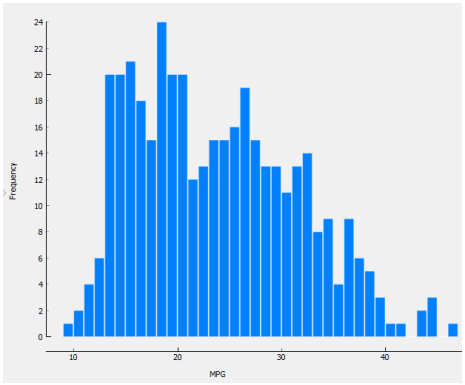

Gráfico 2 | Frequência do **MPG**

No **Gráfico 2** é visível que a predominante parte dos automóveis que encontramos (aproximadamente 84%) fariam consumos entre as cerca de 13 milhas por galão e 33 milhas por galão.

Isto é, os consumos que encontramos nos extremos existem com pouca frequência sendo que os automóveis com consumos inferiores a 13 milhas por galão representam apenas 3% e os automóveis com consumos acima das 33 milhas por galão cerca de 13%.

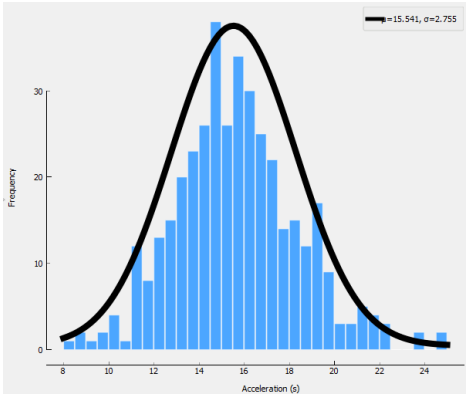

Gráfico 3 | Frequência da ***Aceleração***

Relativamente ao **Gráfico 3**, podemos retirar uma conclusão idêntica à do gráfico anterior (**Gráfico 2**). A conclusão a que nos referimos é de que os valores das acelerações que nos são dados nos extremos tem uma frequência muito inferior à dos valores das acelerações que encontramos ao centro do gráfico.

Cerca de 92% dos automóveis presentes no nosso conjunto de dados chegam aos 100km/h (significado de aceleração) ao fim de 11 a 20 segundos.

Os automóveis que têm acelerações superiores, ou seja, demoram menos tempo a chegar aos 100km/h, representam uma pequena percentagem do total de automóveis (,aproximadamente, 3%).

Por fim, os automóveis que demoram para lá dos 20 segundos até chegarem aos 100km/h que são cerca de 5%.

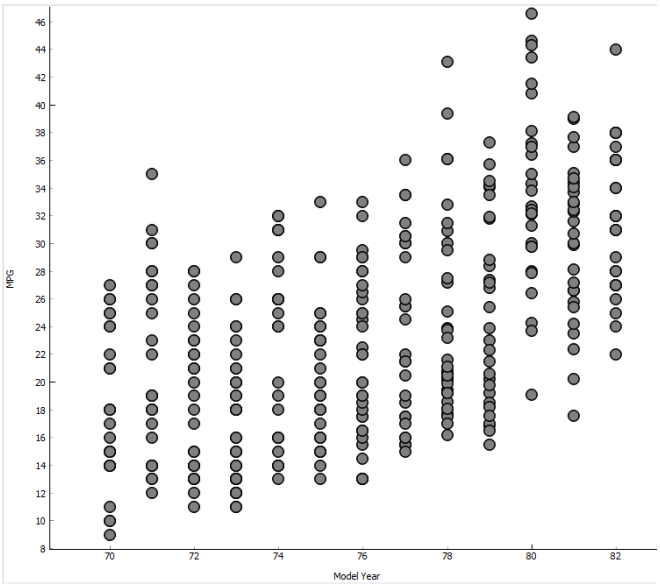

Gráfico 4 | Gráfico da evolução do ***MPG*** em função dos ***Anos dos modelos***

No Gráfico 4, expresso em consumo para os vários anos dos modelos, é visível uma evolução gradativa da variável que traduz o consumo automóvel (MPG). Uma vez que quanto maior o valor associado à variável, menor será o consumo, no geral, do motor.

Em primeira instância, e pela análise do gráfico, é possível observar que ao longo dos anos o consumo tem vindo a diminuir, pelo que a autonomia a aumentar, bem como a eficiência dos motores. Em 70 o consumo variava entre 8-27 MPG, ao passo que, em meados de 80 este consumo já variava entre 21-46 MPG, o que nos permite inferir que se verificaram crescimentos exponencias ao nível da eficiência dos motores com registos de mais 50% a curto prazo, cerca de 5 anos, e a longo prazo mais de 75% (10 anos). Ainda que possa não ser da mesma marca, nem o mesmo modelo, no entanto, verifica-se que, em comparação, a evolução se verificou.

É-nos possível concluir ainda, através do gráfico, que, mesmo que de forma reduzida, este retrata o que foi a evolução dos veículos a motor ao longo dos anos. A competitividade, aliada ao desenvolvimento tecnológico, foram dos aspetos mais preponderantes, que levaram a esta busca incessante pelo que poderia ser, não só, um motor mais eficiente do ponto de vista dos consumos, como a todos os outros níveis.

Em suma, este gráfico traduz com clareza o que foi, não só a evolução no que diz respeito à autonomia, mas também aos automóveis em si e a todos os constituintes que o compõem. Uma evolução crescente e volátil, que advém de algumas tentativas de inovação por parte de algumas empresas, que nem sempre contribuíram para esta evolução, mas que serviram de percussores para outras adoções mais eficientes.

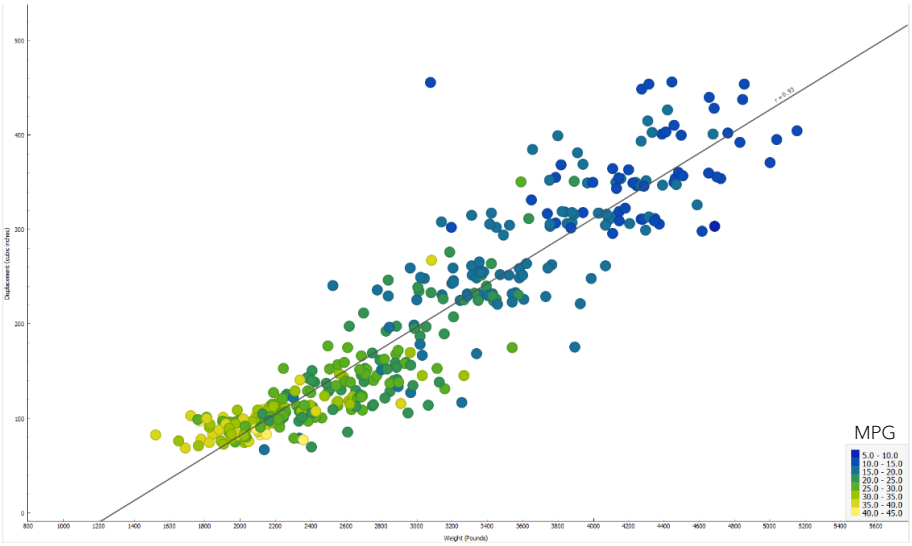

Gráfico 5 | Relação entre o ***Peso***, em *Kg*, a ***Cilindrada***, em *inch*^{3}, e o ***MPG***

Através da análise do **Gráfico 5**, que é um *Scatter Plot*, pudemos relacionar duas variáveis (o peso e a cilindrada), com a nossa variável *target* (o MPG). A conclusão retirada é que quanto maior for o peso de um determinado automóvel, maior tende também a ser a sua cilindrada, mas o MPG é menor, ou seja, há um maior consumo.

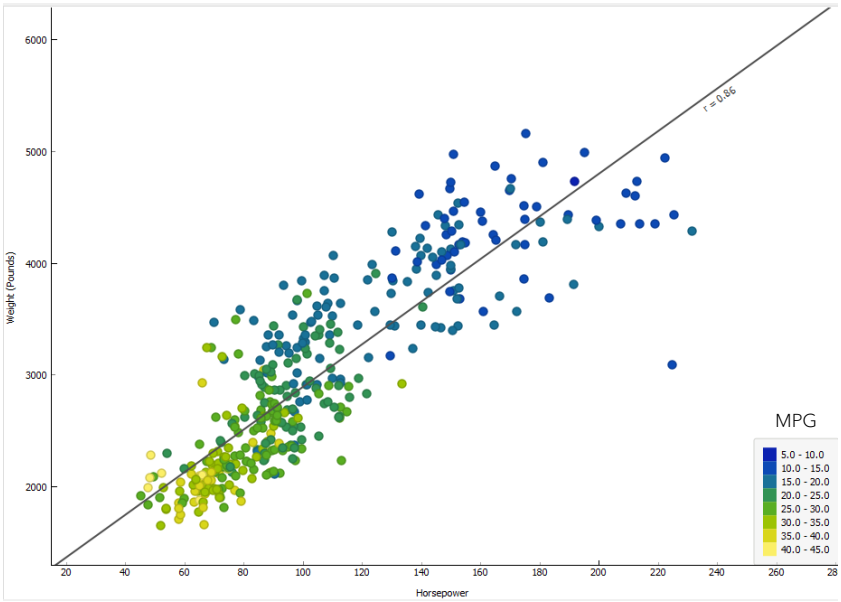

Gráfico 6 | Relação entre os ***Cavalos***, o ***Peso***, em *pounds*, e o ***MPG***

O **Gráfico 6** é semelhante ao gráfico 5 visto que partilham duas das variáveis em estudo (o MPG e o peso), mas introduz uma nova, a potência (horsepower). Após análise, conclui-se que com um aumento da potência, os carros tendem também a ser mais pesados e consumir mais, tendo por isso um menor MPG.

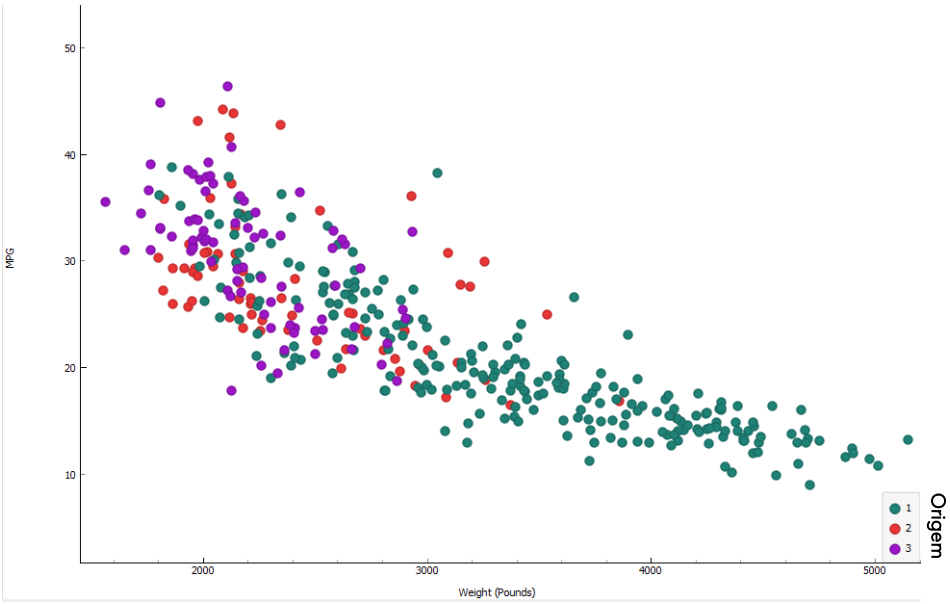

Gráfico 7 | Relação entre os ***Peso***, em *pounds*, o ***MPG***, e a ***Origem*** dos automóveis

À semelhança dos gráficos anteriores, o Gráfico 7 relaciona a variável peso, expresso em *Pounds*, com o consumo, expresso em *MPG*, associado ainda à origem.

Pela análise é possível, ainda que não seja linear, inferir que consoante o aumento do peso, maior será o consumo uma vez que a variável MPG diminui.

Verificamos ainda que a maior parte dos gráficos em que relacionam o peso a conclusão é semelhante, os carros de origem Americana (1) são os que possuem maior tonelagem, o que advém, como inferimos anteriormente, piores consumos.

Os carros de origem Japonesa e Europeia eram carros com melhores consumos, desta forma , e como analisámos previamente, consequentemente com peso um pouco mais reduzido, o que se justifica pelo facto de o acesso aos combustíveis se verificar mais condicionado nestas regiões. Em contrapartida, na América, o acesso aos mesmos era manifestamente mais facilitado, uma vez que era abundante e de baixo custo.

3. *Modeling*

Depois de analisar e compreender as diferentes variáveis e as suas correlações, no *Orange,* começamos a criar o fluxo de dados de modo a modelar o nosso *target.*

Assim, exploramos mais alguns widgets, para ver possíveis relações entre os dados, tais como, os widgets *Distances*, *Hierarchical Clustering*¸ *Rank* e *Box Plot.*

Após essa exploração de dados no software *Orange, adicionámos* os widgets *Test and Score, Logistic Regression, Random Forest, SVM, kNN.*

Na **Figura 3** , encontra-se o ambiente do Orange do nosso projeto.

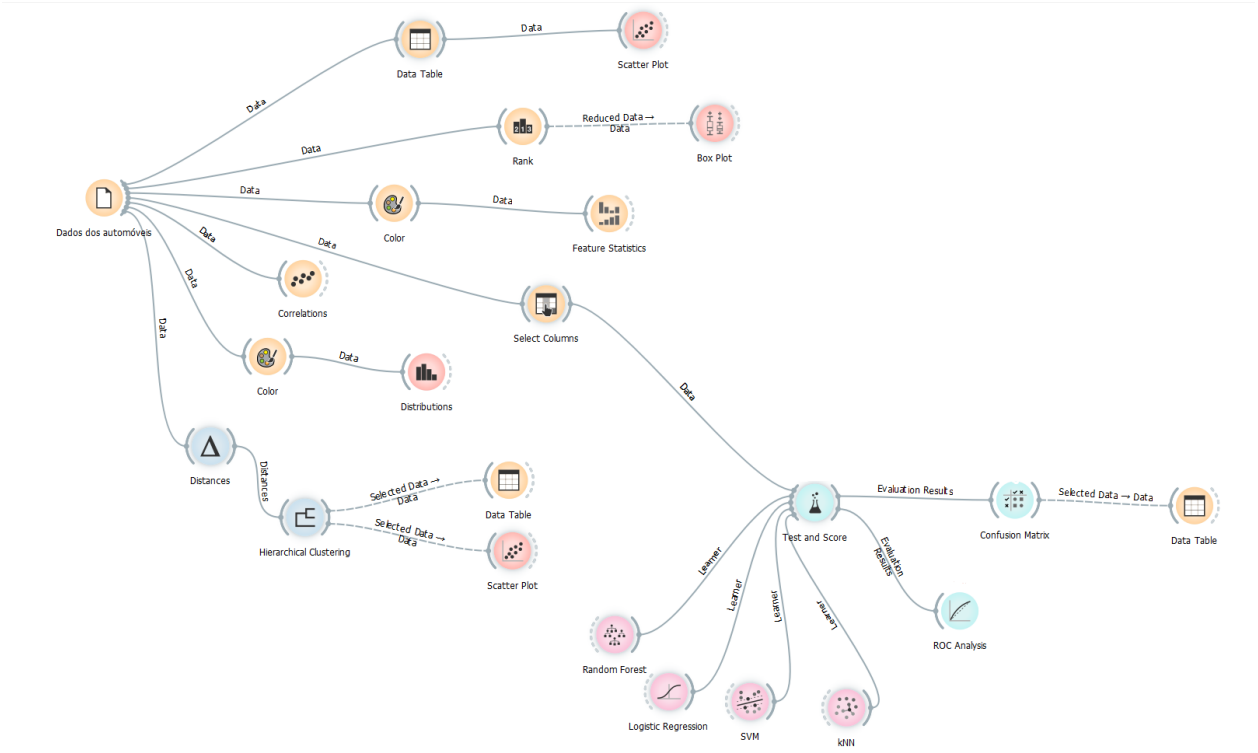

Figura 3 | Fluxo criado do *Orange.*

De modo, a aplicar o conhecimento adquirido em aula, no *Test and Score,* selecionamos a opção *Random Sampling,* na qual optamos por treinar com 70% dos dados e testar com os restantes 30%, sendo que repetimos este processo 10 vezes.

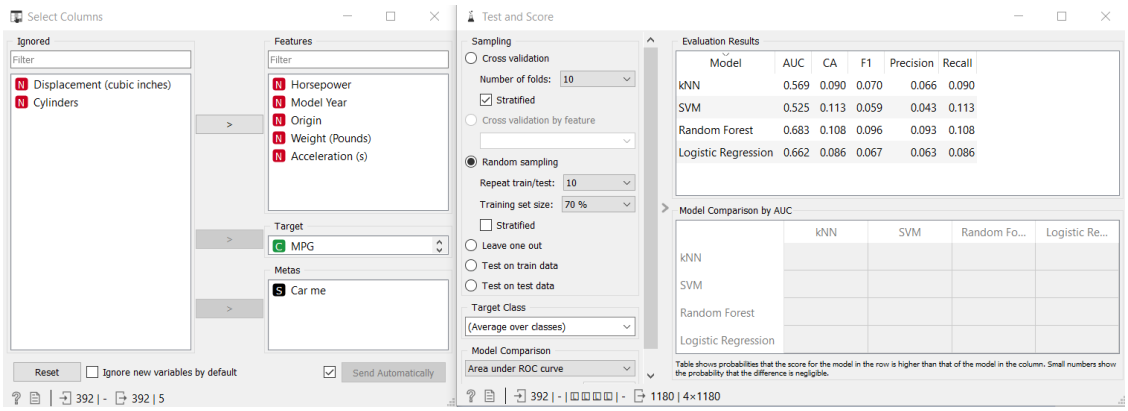

Figura 4 | Melhor modelo com as respetivas Colunas criado do *Orange*.

Para escolher o melhor modelo possível para o nosso *target*, o MPG, fomos selecionando diferentes colunas no widget *Select Columns*, presente antes do *Test and Score,* e de entre várias combinações possíveis, aquela que nos deu um resultado mais favorável ao modelo de *Logistic Regression,* foi a visível na Figura 4, ou seja, com as varáveis de *Horsepower*, *Model Year, Origin, Weight* e *Acceleration.*

Sendo que o nosso *target* é numérico, no entanto, devido à limitação do *Orange* de apenas permitir testar se este for *categorical,* os resultados da *Logistic Regression,* parecem-nos adequados para a complexidade e variedade dos dados reais analisados.

Assim, obtemos uma CA (que relata a proporção de instâncias de dados classificados corretamente) de 0,086. Este valor tem em conta que estão em causa: dados reais, associados à complexidade de um automóvel e o facto de os dados serem antigos, o que leva à existência mais desvios (provenientes da menor precisão dos equipamentos utilizados na época).

Já no que diz respeito aos restantes modelos colocados no *Orange*, foram apenas para uma exploração e visualização de mais modelos, para além do principal, no qual nos baseamos no trabalho, a *Logistic Regression*.

4. *Evaluation | Deployment*

Por fim, com os widgets *Confusion Matrix* e *Roc Analysis*, observámos os resultados do modelo desenvolvido e extraímos as conclusões acerca do mesmo.

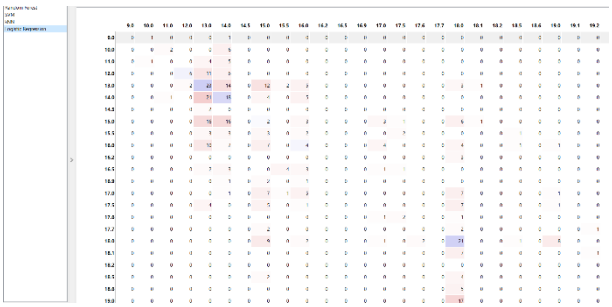

Figura 5 | Parte da Matriz de Confusão do *Logistic Regression* desenvolvido no *Orange*.

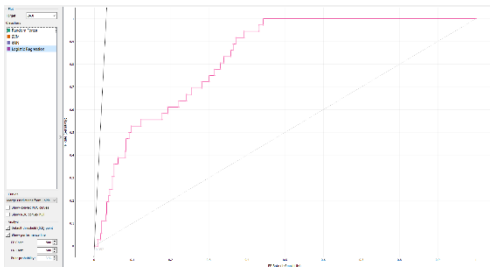

Figura 6 | *ROC Analysis* relativa ao MPG 21,0.

Uma vez mais, sendo o nosso *target* numérico, devido à limitação do *Orange, que* apenas nos permite testar se o *target* for *categorical,* a *Confusion Matrix,* resultante do modelo desenvolvido anteriormente (Figura 5), é de uma elevada dimensão, tornando complexa a sua análise.

Já no que diz respeito à *ROC Analysis* (Figura 6), que é fornecedora de uma representação gráfica de uma curva (quanto mais longe da reta $y = x$ se apresenta, melhor é o valor associado a essa curva), sucede-se o mesmo ao ocorrido na *Confusion Matrix*, isto é, apresentam inúmeras curvas diferentes para cada valor de MPG. Deste modo, não nos permite fazer uma análise dos resultados como um todo, mas apenas de cada valor do MPG individualmente. Uma vez que este está compreendido num intervalo de 9 a 48, a compreensão dos resultados é de difícil execução.

Conclusões

Com os resultados anteriormente obtidos, podemos constatar que o *dataset desenvolvido, associado ao* algoritmo realizado através de técnicas supervisionadas, desenvolvido neste projeto, apresenta uma baixa *Calculation Accucary,* podendo este facto estar relacionado com diversas causas.

Em resposta à pergunta que propusemos responder, *"De que modo o consumo dos automóveis, expresso em MPG, está relacionado com as características intrínsecas aos mesmos."*, inferimos que as variáveis em estudo não são suficientes para que o algoritmo desenvolvido seja preciso no valor que, automaticamente, associa à nossa variável *target* por análise das restantes.

A variável MPG apresenta um valor oscilante que varia de carro para carro, uma vez que são raras as repetições e os padrões que se podem associar e criar. Para que, através da análise deste conjunto de dados, se possa criar um algoritmo capaz de satisfazer uma necessidade tão complexa, como a de prever um consumo automóvel, será necessário recolher mais dados referente a outras variáveis, assim como o acréscimo de mais instâncias.

Não obtivemos as conclusões esperadas, uma vez que não se conseguiu estabelecer uma relação efetiva entre o *target* e o algoritmo, ainda assim foi nos possível extrair conclusões a outros níveis e procurar saber a razão do sucedido. A análise efetiva dos dados aliada à dos gráficos e elementos disponíveis no programa *Orange,* onde conseguimos enfatizar as relações da nossa variável *target* com todas as outras, explorando as mais relevantes, permitiu-nos expô-los anteriormente neste relatório.

Adicionalmente ao trabalho desenvolvido, explorámos os dados fornecidos, numa tentativa de ver outras propostas de problemas, refentes a distintos *targets.* Com este trabalho complementar, observámos resultados bem dispares dos apresentados no presente relatório.

A título de exemplo, o desenvolvimento de modelos cujo *target* seja a *Origem* ou os *Cilindros*, devido ao facto de os valores não variarem nas dimensões que o MPG varia, o *CA* tem elevado valor, tal como a *Precision.* O que seria perfeitamente expectável de ocorrer, uma vez que a facilidade para um algoritmo de apurar um valor para um *target* que tenha apenas 2 a 3 possibilidades, não é análogo à complexidade de apurar um valor para um *target* que tem um intervalo compreendido entre 9 e 46,6. Apesar de ter sido constatado que assim fosse, preferimos, ainda assim, manter a variável MPG como *target* uma vez que foi a variável que nos propusemos a estudar e analisar.

Em suma, a efetividade do algoritmo desenvolvido é baixa, uma vez que como referido, o target se verifica muito incerto e com um grande intervalo de valores, sendo, por isso, difícil de encontrar uma relação que se possa estabelecer para que se associem valores de consumo mediante as outras variáveis.

Todas estas conclusões estão corroboradas pelos resultados obtidos no modelo de previsão que tinha como *target* MPG*,* tal como anteriormente, apresentado e descrito.

Tendo em conta o trabalho realizado, teria sido pertinente ter havido mais tempo no desenvolvimento do projeto, de modo a obtivermos mais modelos de diferentes correlações que podem existir entre variáveis que não fossem expectáveis à partida, assim como apresentar tipos de gráficos mais heterogêneos.

Referências Bibliográficas

https://www.noticiasautomotivas.com.br/como-eram-os-carros-antigamente/, consultado a 2/11/2021

https://pt.wikipedia.org/wiki/Autom%C3%B3vel, consultado a 2/11/2021

https://www.automundo.pt/consumo/fiat-127-um-dos-carros-que-conquistou-os-portugueses-nos-anos-70/, consultado a 2/11/2021

https://pt.wikipedia.org/wiki/Cilindrada, consultado a 2/11/2021

http://www.portalautoshopping.com.br/blog/potencia-de-motor-de-carro-conheca-os-principais-tipos/, consultado a 2/11/2021

https://towardsdatascience.com/crisp-dm-methodology-for-your-first-data-science-project-769f35e0346c, consultado a 4/11/2021

https://towardsdatascience.com/crisp-dm-methodology-for-your-first-data-science-project-769f35e0346c, consultado a 4/11/2021

https://www.kaggle.com/uciml/autompg-dataset, consultado a 4/11/2021