INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA DE INTELIGÊNCIA ARTIFICIAL




Sistema de Previsão de Preço de Carros Usados através de Machine Learning

TOMÁS SILVA DE MAGALHÃES
Julho de 2023

P.PORTO

Sistema de Previsão de Preço de Carros Usados através de Machine Learning

Tomás Silva de Magalhães

Aluno nº: 1151182

Dissertação para obtenção do Grau de

Mestre em Engenharia de Inteligência Artificial

Orientador: Doutor Luiz Felipe Rocha de Faria, Professor Coordenador do Instituto Superior de Engenharia do Instituto Politécnico do Porto

Júri:

Presidente:

Doutora Ana Maria Neves Almeida Baptista Figueiredo, Professora Coordenadora do Instituto Superior de Engenharia do Instituto Politécnico do Porto

Vogais:

Doutor Luiz Felipe Rocha de Faria, Professor Coordenador do Instituto Superior de Engenharia do Instituto Politécnico do Porto

Doutor António Constantino Lopes Martins, Professor Adjunto do Instituto Superior de Engenharia do Instituto Politécnico do Porto

Porto, Julho de 2023

Ao meu eterno avô Tony!

Resumo

O avanço da Inteligência Artificial tem fomentado o lançamento de automóveis com especificações cada vez mais inovadoras e, consequentemente, a preços mais elevados.

Tal aumento de preços conduz a uma maior procura na compra/venda de carros usados. Esta procura leva, muitas vezes, à atribuição de preços irrealistas aos mesmos, aumentando o número de fraudes neste setor, e a uma elevada discrepância nos preços praticados.

Neste âmbito, a área de *Machine Learning* pode ter um papel preponderante, nomeadamente na elaboração de modelos de previsão de preços de carros usados. Assim, o objetivo do presente trabalho prendeu-se com a análise dos modelos já desenvolvidos neste contexto, do grau de precisão dos mesmos e com a criação de um modelo que colmatasse as falhas nos já existentes, de forma a se aumentar o referido grau de precisão.

Neste contexto, foram testados os algoritmos RF, XGBoost, LightGBM, RL, MLP e CNN em quatro conjuntos de dados A, B, C e D. O conjunto de dados A possui 50 características e 57038 carros, o conjunto de dados B possui 30 características e 70253 automóveis, o conjunto de dados C possui 10 características e 192799 veículos e o conjunto de dados D possui as 13 características mais preponderantes e 144702 carros.

Os algoritmos aplicados aos conjuntos de dados A, B e C foram testados duas vezes, com hiperparâmetros padrão e hiperparâmetros modificados. Todos os algoritmos dos quatro conjuntos de dados foram sujeitos a uma metodologia de 80% de treino e de 20% de testes e avaliados, maioritariamente, através das métricas R2, MSE, RMSE e MAE.

Os algoritmos testados com os conjuntos de dados A, B e C obtiveram melhores resultados aquando da alteração de hiperparâmetros padrão, com a exceção do algoritmo MLP no conjunto de dados A e o algoritmo RL nos quatro conjuntos de dados.

Dentro dos algoritmos testados, os algoritmos XGBoost e LightGBM foram os que apresentaram melhores resultados, tendo os mesmos sido muito idênticos entre si nos 4 conjuntos de dados. Entre os dois algoritmos, o XGBoost foi o que apresentou melhores resultados.

Por fim, o algoritmo XGBoost do conjunto de dados A (MAE=0.12892, RMSE=0.18947, MSE=0.03590, R2=0.96432) e D (MAE=0.12389, RMSE=0.18913, MSE=0.03577, R2=0.96404) foram os que apresentaram melhores resultados entre os algoritmos testados, bem como quando comparados com os algoritmos estudados aquando da revisão do estado da arte.

Palavras-chave: Inteligência Artificial, Machine Learning, Deep Learning, Sistema de Previsão, Carros usados, RF, RL, XGBoost, LightGBM, MLP, CNN.

Abstract

The development of Artificial Intelligence has fostered the launch of cars with increasingly innovative specifications and, consequently, at higher prices.

Such price increases lead to a bigger demand for the purchase/sale of used cars. This demand often leads to the attribution of unrealistic prices to used cars, increasing the number of frauds in this setor, and a high discrepancy in prices.

In this context, the area of Machine Learning can play a preponderant role, namely in the elaboration of used car price-prediction models. Thus, the goal of this study was to analyze the models already developed in this context, their precision level as well as the creation of a model that would fill the gaps in the existing models, to increase the referred precision level.

In this context, the algorithms RF, XGBoost, LightGBM, RL, MLP, and CNN were tested on four data sets A, B, C, and D. Dataset A has 50 features and 57038 cars, dataset B has 30 features and 70253 cars, dataset C has 10 features and 192799 vehicles, and dataset D has the 13 most prevalent features and 144702 cars.

The algorithms applied to datasets A, B, and C were tested twice, with default hyperparameters and modified hyperparameters. All algorithms of the four datasets were submitted to an 80% training and 20% testing methodology and mostly evaluated using the R2, MSE, RMSE, and MAE metrics.

The algorithms tested with datasets A, B, and C obtained better results when changing default hyperparameters, except for the MLP algorithm of dataset A and RL algorithm of datasets, A, B, C, and D.

XGBoost and LightGBM algorithms were the most successful ones, being their results very similar to each other in all 4 datasets. Among the two algorithms, XGBoost was the one that presented the best results.

The algorithm XGBoost on datasets A (MAE=0.12892, RMSE=0.18947, MSE=0.03590, R2=0.96432) and D (MAE=0.12389, RMSE=0.18913, MSE=0.03577, R2=0.96404) were the ones that presented better results among the tested algorithms, as well as when compared with the algorithms studied when reviewing the state of the art.

Keywords: Artificial Intelligence, Machine Learning, Deep Learning, Prediction System, Used Cars, RF, RL, XGBoost, LightGBM, MLP, CNN.

Agradecimentos

Em primeiro lugar, quero agradecer ao meu orientador, Dr. Luiz Faria, por toda a disponibilidade demonstrada ao longo de todo o período da dissertação. Por ter estado sempre presente e por ter sempre uma palavra de incentivo. Agradeço-lhe ainda o facto de ter proporcionado um ambiente de trabalho descontraído, de me possibilitar manter sempre focado no que é mais essencial e por todos os conhecimentos e conselhos transmitidos e partilhados. Agradeço também a todos os professores do MEIA, por toda a motivação e passagem de conhecimento ao longo de todo o mestrado.

Agradeço à empresa Standvirtual, em especial ao Daniel Rocha e Miguel Lucas, por toda a disponibilidade e por me terem facultado os dados que me permitiram realizar esta dissertação, de um modo muito mais completo e orientado para o mercado nacional de automóveis.

Aos meus três pilares, mãe, pai e irmão, obrigado pelo apoio incontestável, por toda a força, por terem estado sempre presentes, por me proporcionarem chegar tão longe, por me fazerem acreditar que irei muito mais além, pelos valores que sempre me transmitiram, por serem o meu exemplo, pela inquestionável motivação e, acima de tudo, pela educação que me deram e me proporcionaram ter.

Um agradecimento a toda a minha família e amigos por todas as palavras de força e motivação que me deram ao longo de todo este percurso.

Um agradecimento muito especial à minha namorada, por toda a paciência, por todos os ensinamentos, por todas as palavras de carinho e motivação, por toda a disponibilidade em me ajudar em todos os momentos, por me fazer crescer tanto todos os dias e principalmente pela amizade e amor que me dá.

Por fim, não podia deixar de agradecer ao maior responsável por todo este trabalho. O homem que, sem saber, todos os dias me deu a força, o foco e a determinação para nunca desistir dos meus objetivos. Ele que todos os dias tinha mais um ensinamento ou uma história para contar, para me fazer aprender, crescer e lutar pelo que ambiciono. Sem os acima referidos, mas acima de tudo sem ele fica a certeza de que nada disto seria possível. Obrigado, avô Tony!

Índice

Lista de Figuras

Lista de Tabelas

Acrónimos e Símbolos

Lista de Acrónimos

AD	Árvores de Decisão
ANFIS	*Adaptive Neuro-Fuzzy Inference Systems*
CE	Critério de Exclusão
CI	Critério de Inclusão
CNN	*Convolutional Neural Network*
DL	*Deep Learning*
FD	Fonte de dados
GBDT	*Gradient Boosting Decision Tree*
IA	Inteligência Artificial
KNN	*K-nearest Neighbors*
LightGBM	*Light Gradient Boosting Machine*
LSTM	*Long Short-Term Memory*
MAE	*Mean Absolute Error*
MAPE	*Mean Absolute Percentage Error*
ML	*Machine Learning*
MSE	*Mean Squared Error*
MTE	*Mean Tendency Error*
NB	*Naive Bayes*
O	Objetivo
PLN	Processamento de Linguagem Natural
Q	Questão
R2	*R2 Score*

RF	*Random Forest*
RL	Regressão Linear
RMSE	*Root Mean Square Error*
RNA	Redes Neuronais Artificiais
RNN	*Recurrent Neural Networks*
RP	Regressão Polinomial
SHAP	Shapely Additive exPlanations
SVM	*Support vector machines*
XGBoost	*Extrem Gradient Boosting*

1 Introdução

O presente documento descreve o trabalho preparatório elaborado no âmbito do processo decorrido durante a elaboração de previsão de preços para carros usados aplicado ao mercado nacional, bem como a comparação do mesmo com os já existentes.

Este documento foi escrito no contexto da dissertação de Mestrado em Inteligência Artificial (MEIA) do Instituto Superior de Engenharia do Porto (ISEP), o qual pertence ao Instituto Politécnico do Porto (IPP).

Neste capítulo apresenta-se a contextualização do tema, os problemas mais relevantes identificados no seguimento do mesmo e os principais objetivos para colmatar tais dificuldades. Por último, este capítulo termina com a descrição da estrutura do documento em causa.

1.1 Contextualização

A Inteligência Artificial (IA) é uma área que está cada vez a ter mais impacto no nosso mundo (Poola, 2017). Tais impactos podem ser visíveis em diversas áreas, nomeadamente na educação, na indústria, na saúde, nas finanças, no marketing, nos transportes, entre outros (Shubhendu & Vijay, 2013).

Na área dos transportes, um dos grandes setores que tem vindo a ser explorado pela IA é o do setor automóvel (Gandhi et al., 2022). De facto, é inquestionável o avanço existente em tecnologias como controlo de voz; assistência à condução e estacionamento; sensores digitais; sistemas de segurança inteligentes; sistemas de diagnóstico; etc. (Suhaib Kamran et al., 2022).

Tais avanços levam, indiscutivelmente, a que a indústria acabe por se tornar mais competitiva, lançando automóveis com especificações cada vez mais inovadoras. Estas especificações geram um impacto na definição do preço dos referidos veículos, o qual tem vindo a aumentar significativamente (Pal et al., 2019).

Tal aumento tem fomentado a procura de outras opções, nomeadamente a compra/venda de carros usados (Pal et al., 2019). De facto, desde 2013 que os carros em segunda mão têm cada

vez mais peso no mercado nacional, sendo que a quota de mercado mais do que duplicou nos últimos sete anos, passando de 15,6% para 39,9% (Figura 1) (Nunes, 2021).

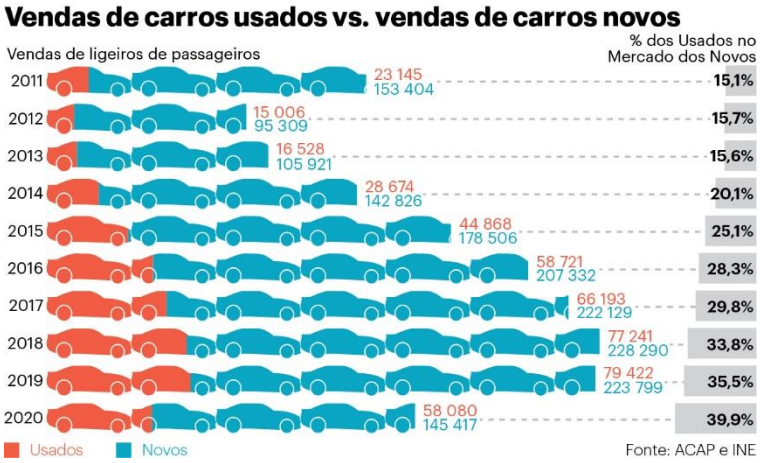


Figura 1 – Vendas de carros usados vs. vendas de carros novos de 2011 a 2020 (Nunes, 2021)

Assim, o mercado dos veículos usados tem cada vez desempenhado um papel mais preponderante no panorama geral do setor automóvel.

1.1.1 Descrição do Problema

A elevada procura de carros em segunda mão leva a que os seus vendedores aproveitem este cenário para atribuir preços irrealistas aos automóveis em causa (Pal et al., 2019), aumentando o número de casos de fraude na venda de um carro usado e condicionando a credibilidade do próprio setor.

Além disso, um dos maiores problemas associados a este mercado prende-se com a dificuldade em perceber qual o preço justo a pagar/vender por um determinado veículo, devido à disparidade de preços praticados dentro da mesma gama ou até dentro do mesmo modelo de carro.

De facto, o preço dos carros usados depende de diversos fatores, sendo que os mais significativos acabam por ser marca e o modelo, a idade, a potência e a quilometragem. O tipo de combustível utilizado no veículo, bem como o consumo de combustível por quilómetro, afetam fortemente o preço de um carro. Além disso, diferentes características como a cor do exterior, o número de portas, o tipo de transmissão, as dimensões do automóvel, a segurança, a existência ou não de ar condicionado, as condições do interior, etc., também influenciam o preço do um carro (Gegic et al., 2019). Contudo, e infelizmente, a informação sobre todas estas características nem sempre está disponível, obrigando o comprador a tomar uma decisão com base num número limitado de fatores (Pudaruth, 2014).

Dado o exposto, torna-se imprescindível a definição de um modelo de avaliação para prever o preço justo para compra/venda de carros usados (Samruddhi & Ashok Kumar, 2020).

Neste âmbito, a IA pode ter um papel preponderante (Idris et al., 2020). Mais concretamente, os algoritmos de *Machine Learning (ML)*, uma subárea da IA, podem ser utilizados para prever o valor de venda a retalho de um automóvel, com base num determinado conjunto de características.

De facto, embora haja um vasto número de aplicações da *ML* na vida real, uma das aplicações mais proeminentes recai sobre o problema de previsão. Os modelos de previsão são processos nos quais são utilizados métodos e tecnologias estatísticas para analisar determinados dados históricos, a fim de fornecer novos *insights* para planificar o futuro em conformidade (Narayana et al., 2021).

O aparecimento de portais *online,* através dos quais é possível estimar o valor de compra/venda de carros usados, tem levado a que tanto o cliente como o vendedor estejam melhor informados sobre as tendências e padrões que determinam o valor de um carro usado no mercado. No entanto, diferentes *websites* aplicam diferentes algoritmos para gerar o preço de venda a retalho de carros usados (Ganesh, 2019). Tais algoritmos podem ter por base diferentes modelos de previsão, cujos números têm vindo a aumentar significativamente, com várias centenas de modelos a serem desenvolvidos todos os dias (Collins & Moons, 2019). Assim, dependendo do portal utilizado, poder-se-ão obter diferentes valores para uma mesma marca e modelo de carro usado.

Dado o exposto, torna-se fundamental analisar os modelos já desenvolvidos neste contexto, perceber o seu grau de precisão e conceber um modelo que colmate as falhas nos já existentes, de forma a se aumentar o referido grau de precisão.

1.1.2 Objetivos

Dado o supracitado, o principal objetivo do presente trabalho prende-se com a comparação do nível de precisão de diferentes modelos de previsão de preço para aplicação no mercado de carros usados e, posteriormente, o desenvolvimento de um modelo mais preciso com base nesse estudo. A concretização do objetivo mencionado pretende, assim, contribuir para a negociação de carros usados a preços mais justos e razoáveis, reduzindo os riscos associados às transações neste setor.

A especificação dos restantes objetivos deste trabalho foi realizada com base em quatro questões principais, as quais visam ajudar a orientar a investigação em curso, encontrando-se descritas de seguida:

- **Q1** – Como se encontra atualmente o estado de arte no que respeita a sistemas de previsão, particularmente no âmbito da previsão de preços para carros usados?
- **Q2** – Poder-se-á considerar um sistema de previsão de preço de carros usados preciso, utilizando apenas um certo número de características e um conjunto de dados relativos a transações anteriores?

- **Q3** – Será o resultado de um modelo de previsão diferente consoante os algoritmos utilizados?
- **Q4** – É um algoritmo mais preciso se fizer uso da combinação de várias técnicas?

Assim, os objetivos identificados com base nas questões acima formuladas são os seguintes:

- **O1** – Investigar o atual estado de arte no âmbito de modelos de previsão e o seu enquadramento no setor automóvel;
- **O2** – Detetar as características que impactam a criação de um modelo de previsão de preço de carros usados;
- **O3** – Criar um *dataset* com a informação necessária para aplicação num modelo de previsão de preço de carros usados;
- **O4** – Comparar e avaliar a performance dos diferentes modelos e técnicas criadas;
- **O5** – Desenvolver um modelo para previsão de preços de carros usados, com base nos resultados obtidos das comparações efetuadas.

Para facilitar a compreensão de que objetivos contribuem para a clarificação de cada questão, a Tabela 1 apresenta a correspondência entre as perguntas de investigação elaboradas e os objetivos que são úteis para lhes dar resposta:

Tabela 1 – Questões de investigação e objetivos correspondentes

Questões	Objetivos
Q1	O1
Q2	O2; O3; O4; O5
Q3	O4; O5
Q4	O4; O5

1.2 Estrutura do documento

O presente trabalho encontra-se estruturado em seis partes principais: Introdução; Estado de Arte; Dados e Metodologia; Análise e Discussão de Resultados; Conclusões Finais e Referências.

O capítulo referente à Introdução contempla uma breve contextualização do panorama no qual o tema da dissertação se insere; o problema que o trabalho desenvolvido pretende colmatar e os objetivos propostos para concretizar com sucesso tal tarefa. Por fim, o primeiro capítulo termina com a descrição da estrutura do documento em causa.

A seguir à Introdução apresenta-se o capítulo Estado de Arte, o qual se foca, essencialmente, na metodologia de pesquisa utilizada, nos modelos de previsão mais utilizados na área de ML e enquadramento dos mesmos na previsão de preço de automóveis usados.

No capítulo seguinte, revela-se o conjunto de dados a utilizar para o treino do modelo a ser desenvolvido, assim como a análise dos referidos dados. Adicionalmente, apresentam-se os algoritmos que vão ser implementados, bem como a metodologia de testes e validação utilizada para a conceção do modelo a desenvolver.

O capítulo subsequente é dedicado à implementação dos modelos desenvolvidos, nomeadamente no que respeita à sua experimentação, treino, teste e validação. Posteriormente, são explicitados, analisados e discutidos os resultados alcançados com o modelo desenvolvido e a comparação dos mesmos com os obtidos com outros já existentes.

A dissertação termina com um capítulo de considerações finais, o qual engloba as conclusões do estudo, principais limitações e sugestões de investigação futura.

2 Estado de Arte

O presente capítulo visa apresentar o estado de arte relativamente a sistemas de previsão para compra/venda de carros usados.

Para tal, começa-se por apresentar a metodologia de pesquisa utilizada na revisão sistemática, nomeadamente no que respeita às fontes de dados usadas, questões e termos de pesquisa, critérios de inclusão e exclusão aplicados, bem como técnicas de extração da informação pretendida.

Posteriormente, elabora-se uma contextualização do panorama geral da Inteligência Artificial (IA), bem como das suas subáreas. Neste contexto, explora-se a área de *Machine Learning* (ML), detalhando-se as técnicas por ela mais utilizadas, nomeadamente de aprendizagem supervisionada, não supervisionada, semi-supervisionada, por reforço e redes neuronais artificiais.

De seguida, apresenta-se uma visão global da aplicabilidade das técnica de ML em sistemas de previsão de preços em diversas áreas, bem como alguns exemplos de artigos elaborados neste âmbito. Subsequentemente, são detalhados os estudos efetuados no contexto da previsão de preços de carros usados, bem como as principais observações deles retiradas.

O presente capítulo termina com uma breve conclusão das principais ideias passíveis de serem retiradas no seguimento do mesmo.

2.1 Metodologia de Pesquisa

Para a revisão sistemática do estado de arte foi utilizado o PRISMA (Page et al., 2021) como metodologia de pesquisa. O objetivo da investigação foi recolher os estudos que, de alguma forma, desenvolveram sistemas de previsão em várias áreas, incluindo a área de sistemas de previsão para compra/venda de automóveis usados.

2.1.1 Fontes de dados

O presente subcapítulo descreve todas as fontes de dados que foram utilizadas para a pesquisa efetuada. A Tabela 2 apresenta as bases de dados que foram utilizadas para este estudo.

Tabela 2 – Tabela de fontes de dados utilizadas

Descrição	Fonte
FD1	IEEE Explore Digital Library
FD2	ScienceDirect
FD3	eBook Index
FD4	Complementary Index
FD5	Academic Search Complete
FD6	Business Source Complete

2.1.2 Questões de pesquisa

Para o propósito da revisão sistemática foram colocadas questões de pesquisa, já referidas no capítulo dos Objetivos, as quais se focam essencialmente no ponto atual do estado de arte no que respeita ao tema, na precisão de um sistema de previsão utilizando apenas um conjunto de dados limitado sobre vendas anteriores, na diferença de resultados entre diversas técnicas e algoritmos, na precisão de um algoritmo englobando várias técnicas e modelos. Todas estas questões foram elaboradas com o intuito de delinear uma linha de raciocínio, de modo a facilitar a compreensão e análise do referido estado de arte.

2.1.3 Termos de Pesquisa

O presente subcapítulo descreve o conjunto de domínios e as respetivas palavras-chave que foram utilizadas para a pesquisa efetuada. Para tal, foi elaborada uma pesquisa desde o âmbito geral até ao particular acerca da IA, dos sistema de previsão e da sua utilização no contexto de carros usados. A Tabela 3 apresenta os domínios e respetivas palavras-chave que foram utilizadas para este estudo.

Tabela 3 – Tabela de Domínios e Palavras-chave.

Domínio	Palavras-chave
Prediction Systems	(“Prediction Systems” OR “Price Prediction”)
Cars	(“Cars” OR “Car”)
Machine learning	(“Machine learning” OR “ML”)

2.1.4 Critérios de Inclusão e Exclusão

A presente secção descreve todos os critérios de inclusão e exclusão incluídos na pesquisa. Nesta foram incluídos todos os artigos que obedeceram aos critérios de inclusão e

automaticamente excluídos todos os que fizeram parte da lista dos critérios de exclusão. As Tabela 4 e Tabela 5 apresentam os critérios de inclusão e exclusão, respetivamente, que foram utilizados para este estudo.

Tabela 4 – Critérios de inclusão

Descrição	Critério
CI1	A fonte foca-se no tema de sistemas de previsão
CI2	A fonte refere várias técnicas ou algoritmos utilizados
CI3	A fonte refere os resultados obtidos para cada um dos algoritmos utilizados
CI4	A fonte descreve as vantagens e desvantagens dos algoritmos utilizados
CI5	Para além dos resultados obtidos, a fonte refere quais as dificuldades encontradas no desenvolvimento dos sistemas de previsão apresentados

Tabela 5 - Critérios de exclusão

Descrição	Critério
CE1	Fontes relativas a sistemas de previsão de carros usados foram redigidas há mais de cinco anos
CE2	A fonte tem mais de 10 anos
CE3	A fonte não refere técnicas ou algoritmos utilizados
CE4	A fonte não apresenta estudos referentes a sistemas de previsão
CE5	A fonte é duplicada
CE6	O artigo não contém o texto integral disponível

2.1.5 Extração de dados

O presente subcapítulo visa demostrar a *query string* que serviu de apoio à pesquisa. Esta *query* [Tabela 6] tem como objetivo juntar todos os domínios e respetivas palavras-chave, representados no anterior subcapítulo 2.1.3. numa só frase, de modo a poder incluir todos os artigos que contenham esta informação. Posteriormente, é mostrado o diagrama do PRISMA (Figura 2) com todo o detalhe sobre a seleção dos artigos estudados.

Tabela 6 – Query String

Query String
((((“Prediction Systems” OR “Price Prediction”) AND (“Cars” OR “Car”)) OR (“Prediction Systems” OR “Price Prediction”)) AND (“Machine learning” OR “ML”))

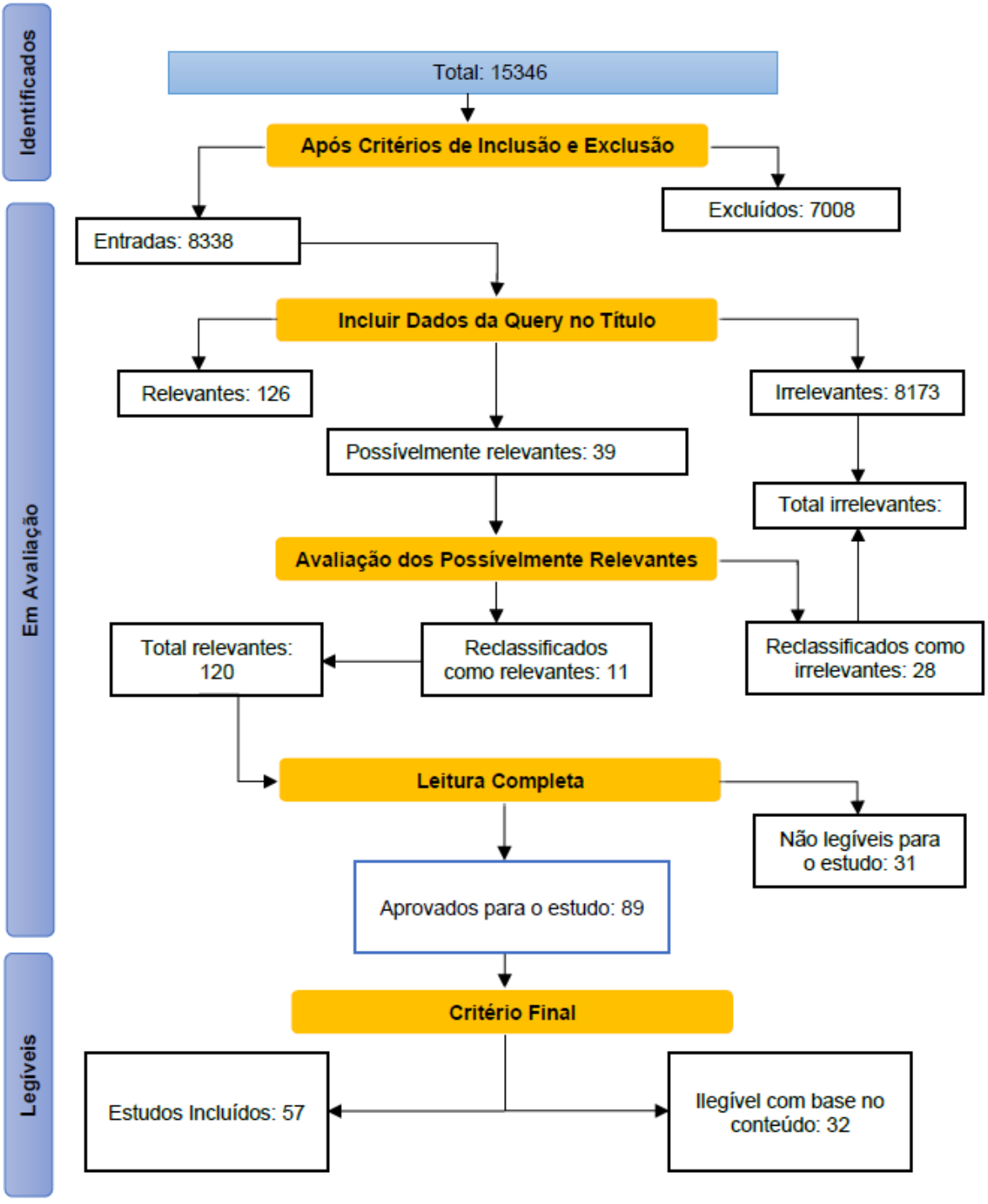


Figura 2 – Diagrama PRISMA

2.2 Introdução à Inteligência Artificial

A Inteligência Artificial (IA) é a capacidade que uma máquina tem para reproduzir competências semelhantes às humanas, como é o caso do raciocínio, a aprendizagem, o planeamento e a criatividade. Para além de isso, estas máquinas têm a capacidade, pela sua observação, de poderem também aprender com elas próprias (Mccarthy, 2007).

As subáreas da IA são variadas. No entanto, existe uma lista de algumas que podem ser consideradas como principais, nomeadamente:

- **Reconhecimento da Fala**: também conhecido por *Speech Recognition* é um processo que, através de um algoritmo, traduz sinais de voz numa sequência de palavras (Anusuya & Katti, 2009).
- **Sistemas Multi Agente:** consistem em entidades conhecidas por agentes que resolvem tarefas com grande flexibilidade, devido à sua capacidade inerente de tomarem decisões autonomamente. Fazem-no através de interações com agentes vizinhos ou com o ambiente ao seu redor, de modo a aprenderem novos contextos ou a executarem determinadas ações (Dorri et al., 2018).
- **Visão por Computador:** é um campo responsável por fazer uma máquina conseguir "ver". Utiliza uma câmara e um computador, os quais substituem o olho humano, com o objetivo de identificar, seguir ou medir alvos, de modo a conseguir realizar o processamento de uma imagem (Tian et al., 2020).
- **Sistemas Periciais:** conhecidos também por *Expert Systems,* são sistemas concebidos por um programa com a capacidade de resolver problemas complexos, proporcionando capacidade de decisão, à semelhança de um perito humano. Funcionam através de informação extraída de uma base de conhecimentos, utilizando raciocínio e regras de inferência, de acordo com as informações obtidas pelo utilizador (Liao, 2005).
- **Processamento de Linguagem Natural (PLN):** é um sistema capaz de analisar, entender e sintetizar a linguagem humana. Existem várias aplicações onde este sistema foi usado ao longo dos anos, tais como: reconhecimento de fala; traduções de linguagem; recuperação de informação; sumarização de textos (Ranjan et al., 2016).
- **Robótica:** é uma área responsável por confecionar, fabricar e operar robôs, onde estes são programados para representar/imitar ações, com objetivo de ajudar os seres humanos de várias maneiras (Asada, 2003).

Para além das subáreas inerentes à inteligência artificial referidas acima, é crucial referir-se uma das mais importantes e mais utilizadas, a subárea de ML.

2.3 Machine Learning

Machine Learning (ML), uma subárea da IA, é um ramo que gira em torno da evolução de sistemas computacionais, baseados em algoritmos, os quais são desenhados para simular a inteligência do ser humano, através do conhecimento cada vez mais aprofundado do ambiente

que o rodeia (el Naqa & Murphy, 2015). Tais sistemas podem aprender de forma autónoma e tomar decisões sem serem explicitamente programados (Mahesh, 2018).

Para que tal seja possível, é necessário treinar um sistema de ML com um conjunto de dados, o qual é usado para ajustar os parâmetros do modelo aplicado. Depois de treinado, esse modelo pode ser usado para fazer previsões ou tomar decisões com base em novos conjuntos de dados.

De facto, pode dizer-se que um sistema de ML aprende com a sua experiência se o seu desempenho na execução de uma determinada tarefa melhorar à medida que o sistema ganha constantemente mais experiência na execução dessa mesma tarefa (Ray, 2019).

Dado o exposto, as diversas fases que formam o processo de ML podem ser resumidas aos passos representados na Figura 3, cujo objetivo final é o de aumentar a precisão do algoritmo e consequentemente a sua performance.

Figura 3 – Fases de um modelo de ML (adaptado de Su et al., 2018)

Existem vários tipos de técnicas de ML, nomeadamente aprendizagem supervisionada (*supervised learning*), não-supervisionada (*unsupervised learning*), semi-supervisionada (*semi supervised learning*) e aprendizagem por reforço (*reinforcement learning*). Cada uma destas técnicas envolve diferentes abordagens para treinar os modelos de ML, as quais são usadas para resolver diferentes tipos de problemas. De seguida, apresenta-se uma breve explicação sobre cada uma das referidas técnicas.

2.3.1 Aprendizagem Supervisionada (*Supervised Learning*)

A aprendizagem supervisionada é uma tarefa na qual um dado modelo é treinado através de dados pré-definidos, os quais são formados por conjuntos de pares *input* e *output* (Mahesh, 2018). Assim, através da análise dos mesmos, o algoritmo aprende a prever o *output* para novos exemplos de *input*, mesmo que estes lhe sejam completamente desconhecidos, através de padrões encontrados nos conjuntos de dados previamente fornecidos. Por este motivo, a aprendizagem supervisionada também é conhecida por aprendizagem via exemplos (Su et al., 2018).

A aprendizagem supervisionada é utilizada para dois tipos de problemas: classificação e regressão (Yan & Wang, 2022). No caso da classificação, o *output* são valores discretos, enquanto no caso da regressão o *output* são valores contínuos (Stulp & Sigaud, 2015).

Existem vários algoritmos que utilizam técnicas de aprendizagem supervisionada, as quais podem ser aplicadas no contexto da previsão de preços de carros usados, sendo os mais relevantes apresentados de seguida (Caruana, 2006):

- ***Support vector machines* (SVM):** podem ser implementadas em classificação e regressão. Estes algoritmos visam encontrar híper planos, os quais irão separar um certo conjunto de dados de uma forma fiável nas distintas classes de dados (Hopke, 2003);
- ***Long Short-Term Memory* (LSTM):** é um tipo de *Recurrent Neural Network* (RNN) que tem a capacidade de obter relações entre dados em longos períodos temporais. Tem a habilidade de esquecer informações antigas e irrelevantes, fazendo com que possa manter uma memória de longo prazo, aumentando a sua eficiência nas tarefas (Yin et al., 2017);
- **Regressão Linear (RL):** é um algoritmo de previsão utilizado para obter valores contínuos baseado nas características de entrada. Assume que existe uma relação entre os valores de entrada e saída. Distingue a influência das variáveis independentes através das variáveis dependentes.(Maulud & Abdulazeez, 2020);
- **Regressão Polinomial (RP):** Ao contrário da regressão linear, a regressão polinomial tem duas variáveis com equações de relação curvilínea (Jin, 2021). É útil quando há razões para acreditar que as relações entre duas variáveis são curvilíneas(Ostertagová, 2012);
- ***Naive Bayes* (NB):** é também um conhecido algoritmo de classificação, o qual utiliza a regra de *Bayes,* onde assume que as características do classificador são independentes da classe. Embora essa suposição seja muitas vezes errada, este algoritmo consegue competir e estar à altura de outros (Rish, 2005);
- **K Vizinhos Mais Próximos (*K nearest neighbors* - KNN):** tenta classificar cada amostra de um conjunto de dados avaliando a sua distância em relação aos vizinhos mais próximos. No caso dos mesmos pertencerem maioritariamente a uma classe, a amostra em questão será classificada nessa categoria (M. L. Zhang & Zhou, 2007);
- **Árvores de Decisão (AD):** têm sido bastante aplicadas para modelos de classificação, uma vez que se assemelham bastante ao raciocínio humano e são de aprendizagem bastante fácil. Estas árvores são constituídas por nós internos, os quais representam as características de um conjunto de dados; ramos, os quais representam as regras de decisão; e nós folha, representando o resultado (Navada et al., 2011);
- ***Gradient Boosting Decision Tree* (GBDT):** é um algoritmo que executa várias árvores de decisão. Treina-as de uma forma sequencial corrigindo os erros das árvores anteriores, de modo a conseguir o resultado mais preciso possível. GBDT é popular devido à sua exatidão, eficiência e interoperabilidade (Z. Zhang & Jung, 2019).
- ***Extreme Gradient Boosting* (XGBoost):** é um algoritmo melhorado do GBDT, o qual consegue lidar com as árvores de decisão de maneira mais eficiente e com um conjunto de dados muito superior (T. Chen & Guestrin, 2016).
- ***Light Gradient Boosting Machine* (LightGBM):** é também uma versão melhorada do GBDT. É otimizado para trabalhar com um grande conjunto de dados e melhorar o

desempenho no treino dos mesmos (Ke et al., 2017). É similar ao algoritmo de XGBoost, no entanto tem a vantagem de ser mais rápido a treinar os dados, é mais eficiente no treino, utiliza menos recursos de memória, tem mais exatidão nos resultados e consegue treinar mais quantidades de dados(Kasturi, 2019).

- ***Random Forest* (RF):** é um algoritmo que combina, de forma aleatória, várias árvores de decisão e agrega as previsões através da média. Funciona particularmente bem em cenários onde o número de variáveis é muito maior do que o número de observações (Biau & Scornet, 2016).

2.3.2 Aprendizagem Não Supervisionada (*Unupervised Learning*)

Ao contrário do que acontece na aprendizagem supervisionada, a abordagem de aprendizagem não supervisionada consiste em reconhecer padrões existentes não identificados à priori nos dados fornecidos (Mahesh, 2019). Assim, os algoritmos são deixados à sua própria conceção para descobrir e apresentar uma possível estrutura interessante nos dados analisados (Su et al., 2018). Como tal, com esta técnica, o principal objetivo prende-se com a descoberta de estruturas desconhecidas, de semelhanças entre dados e de possíveis agrupamentos dos mesmos (Károly et al., 2018).

O agrupamento (*Clustering*) é considerado um dos principais tópicos na abordagem de aprendizagem não supervisionada. Este é um método responsável por encontrar uma determinada estrutura num conjunto de dados, a qual se se caracteriza pela maior semelhança dentro do mesmo agrupamento e pela maior disparidade entre diferentes agrupamentos (Figura 4) (Sinaga & Yang, 2020).

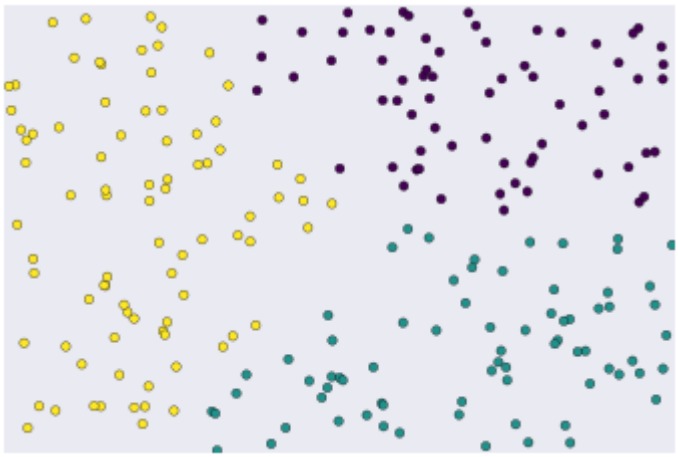


Figura 4 – Exemplo de *clustering* dividido em três agrupamentos (Google, 2022)

K-means, um dos principais algoritmos do clustering e da aprendizagem não supervisionada, tem como grande objetivo dividir 'n' observações em 'k' agrupamentos, nos quais cada observação pertence ao agrupamento com a média mais próxima. A média das observações de um determinado agrupamento define o centro do mesmo (Figura 5) (Su et al., 2018).

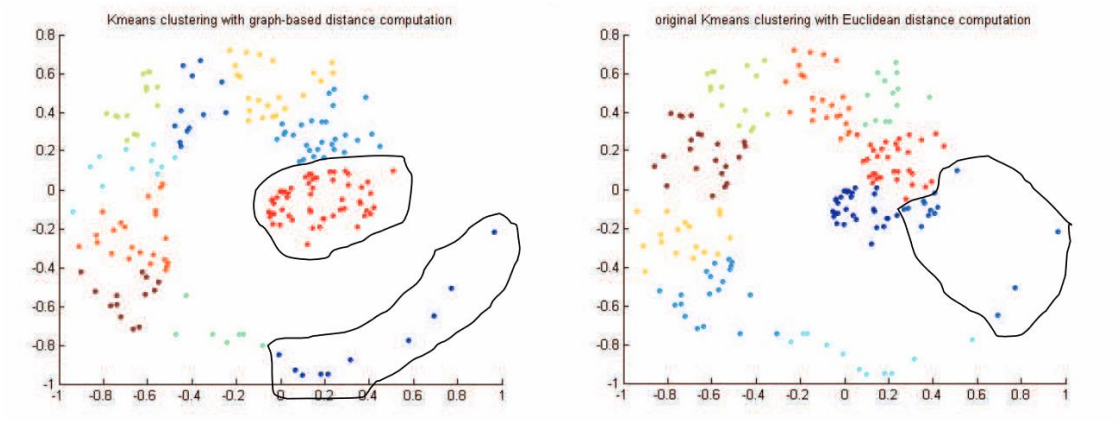


Figura 5 – Exemplos do algoritmo K-means (Nguyen et al., 2005)

Outro dos tópicos bastante utilizados na aprendizagem não supervisionada é a redução de dimensionalidade. Este é responsável por reduzir a redundância, o ruído dos dados, a complexidade dos algoritmos de aprendizagem e melhorar a precisão da classificação (Huang et al., 2019). É um processo que pode ser realizado de duas formas: através da redução do número de variáveis aleatórias do modelo de treino, mantendo apenas as mais relevantes do conjunto de dados inicial; ou então realizando um levantamento da redundância dos dados de entrada, podendo juntar várias variáveis numa só, ficando com menos variáveis à partida e contendo a mesma informação que teria inicialmente (Sorzano et al., 2014).

Para além dos acima referidos, existe também o tópico de regras de associação, as quais são bastante úteis quando um determinado utilizador deseja segmentar dados. De forma relativamente intuitiva, uma regra de associação identifica padrões de informação que ocorrem frequentemente num conjunto de dados e, consequentemente, cria associações com a informação obtida (Lent et al., 1997).

2.3.3 Aprendizagem Semi-Supervisionada (Semi-*supervised Learning*)

A aprendizagem semi-supervisionada é uma junção das técnicas de aprendizagens supervisionadas e não supervisionadas (Figura 6) (Mahesh, 2018). Como previamente já discutido, enquanto que na aprendizagem supervisionada o modelo de dados é treinado apenas com exemplos identificados/rotulados, na aprendizagem não supervisionada esses mesmos modelos de dados são treinados apenas com exemplos não identificados/não rotulados. Na semi-supervisionada o modelo é treinado com uma combinação de exemplos rotulados e exemplos não rotulados (Goldberg, 2009).

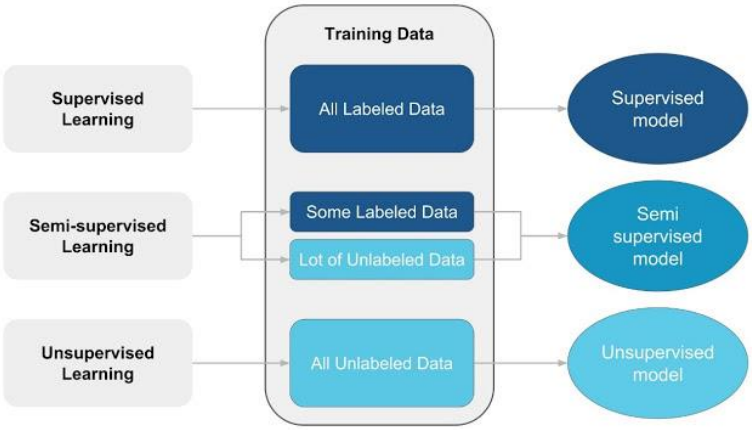


Figura 6 – Relação entre aprendizagem semi-supervisionada, supervisionada e não supervisionada (Ibañez, 2019)

A aprendizagem semi-supervisionada utiliza dados rotulados para classificar os dados não rotulados, de forma a desenvolver melhores classificadores. No geral, a aprendizagem semi-supervisionada pode fornecer resultados comparáveis aos da aprendizagem supervisionada, utilizando menos dados rotulados. No entanto, é importante ter em conta que o desempenho do modelo pode ficar limitado caso o número de exemplos rotulados disponíveis seja insuficiente (Prakash & Nithya, 2014).

2.3.4 Aprendizagem por Reforço (Reinforcement Learning)

Aprendizagem por reforço é um tipo de aprendizagem situado entre a aprendizagem supervisionada e a aprendizagem não supervisionada (van Otterlo & Wiering, 2012). Esta é a aprendizagem de um mapeamento de situações para ações, onde um determinado agente recebe uma recompensa por realizar essas mesmas ações. Tais ações levam o agente a aprender por tentativa erro, conduzindo-o ao resultado esperado (Figura 7). Em nenhuma instância é dito ao agente quais as ações a tomar, sendo que o mesmo deve descobrir quais produzem a maior recompensa, experimentando-as. Estas ações podem afetar não só a recompensa imediata, como também a situação seguinte e, através dela, todas as recompensas subsequentes (Sutton, 1992). Ou seja, o agente tem em consideração o *feedback* adquirido para melhorar a sua tomada de decisão ao longo do tempo.

O modelo de aprendizagem por reforço pode, assim, ser dividido em quatro componentes principais:

- **Agente:** É o agente que decide quais ações tomar, interagindo com o ambiente. Toma decisões com base nas informações recebidas do ambiente e é responsável pelas suas escolhas(Kim, 2022).
- **Ações:** O agente pode optar por tomas variadas ações, as quais podem ser ações simples, através de um único movimento, ou complexas, podendo fazer vários movimentos. Estas ações irão ter impacto numa futura recompensa (Kim, 2022).

- **Ambiente:** O ambiente é o mundo no qual os agentes operam. Este poderá ser real ou virtual (Sutton & Barto, 1999).
- **Recompensas:** As recompensas são variáveis associadas às ações tomadas pelos agentes. Estas definem o objetivo do agente numa determinada situação, sendo que podem ser positivas ou negativas. Dessa forma, o agente usa essa informação para aprender quais ações são melhores para atingir o resultado desejado (Sutton & Barto, 1999).

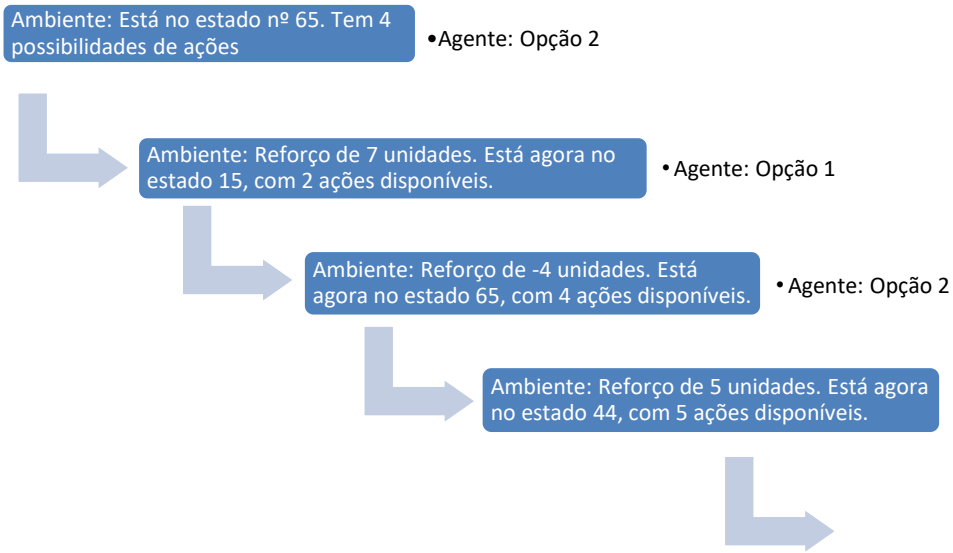


Figura 7 – Interação entre ambiente e agente num senário de aprendizagem por reforço (baseado em Kaelbling et al., 1996)

2.3.5 Redes Neuronais Artificiais

As Redes Neuronais Artificiais (RNA) podem ser utilizadas pelas quatro técnicas acima referidas, aprendizagem supervisionada, não supervisionada, semi-supervisionada e por reforço.

As RNA são redes formadas por nós que constituem unidades de processamento que tentam simular o funcionamento dos neurónios. As RNA obtêm conhecimento através da deteção de padrões e relações nos dados e, consequentemente, aprendem através da experiência que vão obtendo (Agatonovic-Kustrin & Beresford, 2000).

As RNA são uma categoria de Redes Neuronais onde, dentro dessa categoria, existem vários tipos de redes neuronais artificais, nomeadamente: Redes Neuronais Recorrentes (*Recurrent Neural Network*, RNN), Modelo Perceptrão Multicamada (*Multilayer Perceptron, MLP)*, Redes Neuronais Convolucionais (*Convolutional Neural Network*, CNN) e Redes Adversárias Generativas (*Generative Adversarial Networks, GAN).*

As RNN são redes que têm a capacidade de manter em memória dados antigos para processar dados novos. As RNN são especialmente usadas para tarefas de processamento de linguagem natural, tradução automática, reconhecimento de fala e previsões de séries temporais. Dois tipos mais frequentes RNN são as LSTM (Long Short-term Memory) e GRU (*Gated Recurrent Units*) (DiPietro & Hager, 2019).

As MLP são um tipo de rede neuronal artificial do tipo *feedforward*, sendo que a informação que recebe vai apenas numa direção, da camada de entrada, passando pelas camadas intermédias, até à camada de saída. Esta rede é principalmente usada em problemas de reconhecimento de padrões e interpolação (University & 2005, 2005).

As CNN são compostas por neurónios que se otimizam através da aprendizagem em semelhança com as restantes RNA. As CNN têm como principal diferença para algumas das RNA tradicionais um grande domínio no reconhecimento de padrões de imagem. Essa vantagem permite que consiga codificar características específicas e ainda reduza os parâmetros necessários para configurar o modelo (O'Shea & Nash, 2015).

As GAN são arquiteturas de redes neuronais profundas baseadas na teoria de jogos. O objetivo de um modelo generativo é estudar um conjunto de dados de treino e aprender a distribuição que os gerou. Desta forma, as GAN são capazes de gerar mais exemplos a partir da distribuição de probabilidade estimada (Metz et al., 2016).

Dependendo do grau de profundidade da rede neuronal, as RNA podem ser consideradas algoritmos de *machine learning* ou de *deep learning* (DL), consoante são menos ou mais profundas, respetivamente.

2.4 Métricas de Avaliação

A qualidade dos algoritmos como resposta a um determinado problema pode ser medida através de métricas de avaliação. Além disso, a utilização das mesmas métricas na comparação de resultados obtidos através de diferentes algoritmos é de extrema importância para uma análise coerente e correta.

Existem diversas métricas de avaliação de algoritmos em IA, sendo a utilização de métricas apropriadas para cada problema de extrema importância, uma vez que o valor das mesmas reflete a qualidade de um modelo. Algumas das métricas mais utilizadas em sistema de previsão são R2, RMSE, MAE e MSE.

De seguida apresenta-se uma breve descrição sobre cada uma das métricas supracitadas.

2.4.1 R^2

A métrica R^2, coeficiente de determinação, é uma métrica estatística comumente utilizada para avaliar o desempenho de modelos de regressão. Esta métrica pode ser calculada através da seguinte fórmula:

$$\boldsymbol{R2 = 1 - \frac{SSE}{SST}}$$

onde:

- SSE, *Sum of Squared Errors,* representa a soma dos quadrados dos resíduos, o qual simboliza a quantidade de variabilidade não explicada pelo modelo;
- SST, *Total Sum of Squares*, representa a soma total dos quadrados, a qual simboliza a variabilidade total da variável de resposta.

Esta métrica varia entre 0 e 1 e é normalmente expressa em termos percentuais, sendo que quanto maior o R^2, mais explicativo é o modelo linear, ou seja, melhor o mesmo se ajusta à amostra.

A métrica R^2 é simples de calcular e de interpretar, permitindo a comparação direta entre diferentes modelos de previsão. Contudo, esta métrica pode ser influenciada por *outilers*, especialmente quando os mesmos têm um impacto significativo na variabilidade dos dados. Além disso, esta métrica pode não ser adequada para avaliar modelos de previsão mais complexos, nomeadamente modelos não lineares. Como tal, é de salientar a importância da inclusão de outras métricas como a MSE, RMSE e MAE (Chicco et al., 2021).

2.4.2 MSE

A métrica MSE, *Mean Square Error*, é utilizada para quantificar a média dos quadrados dos erros entre as previsões do modelo e os valores reais. Esta métrica é calculada através da seguinte fórmula:

$$\boldsymbol{MSE = \sum_{i=1}^{n} (yi - pi)^2}$$

onde:

- n é o número de amostras;
- yi é o valor observado para cada amostra;
- pi é o valor previsto pelo modelo para cada amostra;

pelo que quanto menor o valor de MAE, melhor é o resultado do modelo.

Dado o exposto, esta métrica apresenta sensibilidade a *outliers* uma vez que, ao elevar ao quadrado os erros, o MSE penaliza erros grandes de forma mais significativa do que erros pequenos. Assim, um único *outlier* pode distorcer significativamente o valor desta métrica.

Uma das principais desvantagens desta métrica prede-se com a dificuldade de interpretação do seu resultado, devido à sua natureza quadrática. Para colmatar este inconveniente, a métrica RMSE é frequentemente utilizada (Chicco et al., 2021).

2.4.3 RMSE

A métrica RMSE, *Root Mean Square Error*, é utilizada para calcular a raiz quadrada da média dos quadrados dos erros entre as previsões do modelo e os valores reais. Esta métrica é calculada através da seguinte fórmula:

$$\boldsymbol{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{yi} - \boldsymbol{pi})^{\mathbf{2}}}$$

onde:

- n é o número de amostras;
- yi é o valor observado para cada amostra;
- pi é o valor previsto pelo modelo para cada amostra.

Também aqui é valido que um menor valor de RMSE significa um melhor ajuste do modelo e uma maior precisão nas previsões do mesmo.

Esta métrica é, assim, uma variação do MSE, uma vez que raiz quadrada é aplicada para obter uma medida de erro na mesma escala em que se encontram os valores originais, pelo que também ela é sensível a *outliers (Chai & Draxler, 2014)*.

2.4.4 MAE

A métrica MAE, *Mean Absolute Error*, é utilizada, como o próprio nome indica, para quantificar a média da diferença absoluta entre as previsões de um determinado modelo e os valores reais utilizados. Esta métrica é calculada através da seguinte fórmula:

$$\boldsymbol{MAE} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{n}|\boldsymbol{yi} - \boldsymbol{pi}|$$

onde:

- n é o número de amostras;

- yi é o valor observado para cada amostra;
- pi é o valor previsto pelo modelo para cada amostra;

pelo que quanto menor o valor de MAE, melhor é o resultado do algoritmo. Esta métrica apresenta diversas vantagens, nomeadamente o facto de ser de fácil cálculo e interpretação e de ser simples de otimizar.

A métrica MAE é menos sensível a *outliers* quando comparando com outras métricas, como o MSE, uma vez que pesa todos os erros de igual forma, independentemente da sua magnitude. Consequentemente, esta métrica não faz uma penalização diferenciada entre grandes ou pequenos erros. Assim, a MAE deve ser principalmente utilizada em casos em que os erros próximos da mediana são mais importantes do que os erros extremos (Chai & Draxler, 2014).

2.5 Sistemas de Previsão

É possível encontrar-se vários artigos na literatura acerca de sistemas de previsão de preços. Tais sistemas de previsão têm por base um conjunto de dados, com diversas características, os quais dependem da área de estudo em questão. Com estes dados e características é possível, através de algoritmos de ML, tentar determinar-se o preço mais justo e real para um determinado produto, bem ou objeto.

Alguns dos contextos para os quais se têm desenvolvido sistemas de previsão de preços prendem-se, por exemplo, com a compra e venda de ouro (Farahani & Mehralian, 2013), moedas digitais (McNally et al., 2018), ações (Kumar et al., 2018) e de casas (Banerjee & Dutta, 2018). Alguns dos algoritmos utilizados no desenvolvimento destes sistemas foram o RF, RNA, SVM, KNN, NB, LSTM, entre outros.

Farahani & Mehralian, 2013 utilizaram os algoritmos RNA e *Adaptive Neuro-Fuzzy Inference* Systems (ANFIS) para prever o preço do ouro, bem como um terceiro resultante de uma abordagem híbrida que realiza uma média ponderada entre os dois algoritmos. Para a avaliação dos algoritmos utilizaram como métricas o *Root Mean Square Error* (RMSE), *Mean Tendency Error* (MTE) e percentagem de erro e verificaram que, apesar de igualmente válidos para o estudo em questão, a performance do algoritmo de ANFIS foi ligeiramente superior ao de RNA. Já o algoritmo híbrido, para além de uma exatidão comparável com ANFIS e RNA, apresentou melhores resultados. Neste estudo foi ainda avaliado o efeito da aplicação do algoritmo *Wavelet Denoising* no conjunto de dados usados, tendo-se constatado que o mesmo melhorou, significativamente, a performance da RNA.

McNally et al., 2018 realizaram um estudo que verifica com que exatidão se pode prever o preço da *bitcoin* em dólares. Para tal implementaram os algoritmos de RNN, LSTM e ARIMA, os quais foram avaliados através de critérios de sensibilidade, especificidade, precisão, exatidão e RMSE. Após análise dos vários parâmetros de avaliação, os autores concluíram que o algoritmo de

LSTM foi o que apresentou melhores resultados, apresentando melhor capacidade de reconhecer tendências a longo prazo.

I. Kumar et al., 2018 testaram os algoritmos SVM, RF, KNN e NB para obter a previsão do preço de ações no mercado da bolsa e Banerjee & Dutta, 2018 testaram os algoritmos SVM, e RNA para determinar se o preço de uma determinada casa iria aumentar ou diminuir. No estudo conduzido por I. Kumar et al., 2018, os algoritmos foram medidos pela sua exatidão, enquanto que no estudo efetuado por Banerjee & Dutta, 2018, para além da exatidão, foram também medidas a precisão, a sensibilidade e a especificidade. Em ambos os casos, o algoritmo através do qual obtiveram melhores resultados foi o RF.

No entanto, no caso de Banerjee & Dutta, 2018, graças às medidas extra utilizadas para avaliação de algoritmos, nomeadamente a precisão, chegaram à conclusão de que o algoritmo RF demonstrou um sobreajuste (*over fitting*), pelo consideraram que o SVM foi o algoritmo mais fiável.

É também importante referir que, após a primeira abordagem de I. Kumar et al., 2018, o dataset utilizado foi reduzido para metade das ocorrências e, nesse caso, o melhor resultado foi obtido através do algoritmo NB, apesar dos autores terem também concluído que esta alteração diminuiu a exatidão de todos os algoritmos testados.

Com a análise dos estudos elaborados por I. Kumar et al., 2018 e Banerjee & Dutta, 2018 pode concluir-se, respetivamente, que a quantidade de dados utilizados no desenvolvimento e teste de algoritmos é preponderante para o desempenho dos mesmos e que, dependendo dos critérios de avaliação aplicados, podem tirar-se diferentes conclusões acerca do melhor algoritmo para uma determinada aplicação.

2.5.1 Sistemas de Previsão de Preços de Carros Usados

Para além da aplicação dos sistemas de previsão descrita no subcapítulo anterior, estes sistemas são também bastante usados no setor automóvel, nomeadamente no setor dos carros usados.

O presente subcapítulo visa analisar alguns dos trabalhos que já foram desenvolvidos no âmbito em causa, para os quais serão apresentados os diferentes algoritmos e respetivas métricas de avaliação, bem como os resultados obtidos. Além disso, serão também apontadas as principais limitações e possíveis melhorias encontradas em cada estudo, sempre que as mesmas existirem, de modo a melhor se compreender como é que tais limitações podem ser colmatadas e como é que tais melhorias podem ser aplicadas no âmbito da dissertação em questão.

Grande parte dos algoritmos utilizados nos estudos apurados utilizam algoritmos de aprendizagem supervisionada para prever eventos futuros a partir de dados anteriores e atuais. Os algoritmos que utilizam esta aprendizagem comparam os resultados obtidos com os resultados reais e esperados, identificando erros e alterando os modelos com base nesses mesmos resultados (Saravanan & Sujatha, 2019).

É possível encontrar-se vários artigos na literatura acerca de sistemas de previsão de preços de automóveis usados. Tais sistemas de previsão têm por base um conjunto de dados, com diversas características sobre os carros. Na Tabela 7 podem observar-se as diferentes características que foram usadas nos diferentes artigos estudados.

Tabela 7 – Características de carros usados nos modelos utilizados pelos artigos estudados

Característica	**Narayana et al., 2021**	**C. Chen et al., 2017**	**Longani et al., 2021**	**Gupta et al., 2022**	**Jin, 2021**	**Hankar et al., 2022**	**H. Zhang, 2022**	**Arora et al., 2022**	**Narayana et al., 2022**
Ano do Carro	X	X	X	X	X	X		X	X
Ano do Modelo		X							
Cilindrada		X	X	X	X				
Combustível	X	X	X	X		X	X	X	X
Compressão				X					
Comprimento				X					
Comprimento Cilindro				X					
Data Transação							X		
Distância Eixos		X							
Estado do Veículo		X					X		
ID Venda							X		
Imposto					X				
Km p/ Litro			X	X	X				
Largura				X					
Localização Motor				X					
Marca e Modelo	X	X		X		X	X	X	X
N Cilindros				X					
Nº de Assentos		X	X						X
Nº de Portas		X		X					
Nº Donos	X		X						
Nº Mudanças		X							
Nº Rodas		X		X					
Nome Transação							X		
Poder Fiscal						X			
Oferta/Requisição							X		
Pais da Marca		X							
Perda Energia Motor				X					
Peso				X					
Potência			X	X			X		
Preço de venda	X	X	X	X	X	X	X	X	X
Preço atual								X	
Quilometragem	X	X	X		X	X	X	X	X
Região		X					X		
Sistema de Combustível				X					
Tamanho Cilindros				X					
Tara				X					
Tempo de Venda							X		
Tipo de Caixa	X	X	X	X			X	X	X
Tipo de Comprador									X

Segmento		**X**							
Tipo de Veículo		**X**		**X**			**X**		
Tipo de Vendedor	**X**	**X**	**X**				X	**X**	**X**
Tipo Motor		**X**		**X**					
Transmissão				**X**					
Tipo de volante									**X**

Com acesso a várias das características descritas, os autores dos artigos estudados criaram modelos onde aplicaram alguns dos seguintes algoritmos de ML: RL, AD, RF, XGBoost, GBDT, LightGBM, SVM KNN e RNA. Para análise e avaliação dos algoritmos desenvolvidos, os diversos autores utilizaram algumas das seguintes métricas: *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), RMSE, *Mean Absolute Percentage Error* (MAPE) e R2, de modo a concluírem qual dos algoritmos obteve melhor performance na previsão de preços de carros usados.

Tome-se como exemplos Narayana et al., 2021, os quais testaram os algoritmos RL, AD e RF; C. Chen et al., 2017, que testaram os algoritmos RL e RF; Gupta et al., 2022, os quais testaram os algoritmos RL, AD, RF, SVM e Elastic Net; Jin, 2021 que testou os algoritmos RL, AD, RF, SVM e Regressão Polinomial (RP); e Narayana et al., 2022), os quais testaram os algoritmos RL e RF. Todos estes estudos têm em comum o facto de o algoritmo de RF ter sido considerado o melhor, à exceção de Gupta et al., 2022, independentemente do número de carros e das características usadas terem sido diferentes.

Longani et al., 2021, Hankar et al., 2022 e Arora et al., 2022 também testaram o algoritmo de RF. No entanto, para além deste, Longani et al., 2021 testaram ainda o algoritmo XGBoost, Hankar et al., 2022 os algoritmos RL, XGBoost, KNN e RNA e Arora et al., 2022 os algoritmos RL, XGBoost e KNN. Nos três casos, o algoritmo de XGBoost superou os seus adversários em todas as métricas abordadas por cada autor.

H. Zhang, 2022 foi, dos artigos estudados, o que testou mais algoritmos, nomeadamente: RL, RF, XGBoost, GBDT, LightGBM e SVM. Neste estudo, o algoritmo que obteve melhores resultados foi o LightGBM.

A Tabela 8 resume os algoritmos estudados em cada um dos artigos, bem como as métricas utilizadas para a avaliação dos respetivos algoritmos, destacando a negrito os algoritmos que foram mais eficazes para cada um dos artigos. É de salientar que algoritmos: GBDT, Elastic Net, Regressão Polinomial e RNA não foram incluídos na tabela, devido aos seguintes factos:

- Foram testados apenas uma vez e por um único artigo;
- Em nenhuma das métricas foram superiores aos algoritmos estudados.

Tabela 8 - Principais algoritmos utilizados nos artigos estudados

Artigos	RL	AD	RF	XGBoost	LightGBM	SVM	KNN
Narayana et al., 2021	MAE: 0.37998 MSE: 0.36805 RMSE: 0.60667	MAE: 0.23565 MSE: 0.21704 RMSE: 0.46587	**MAE: 0.19780 MSE: 0.10122 RMSE: 0.31816**				
C. Chen et al., 2017	RMSE: 0.26000		**RMSE: 0.05200**				
Longani et al., 2021			MAE: 1.13000 MSE: 11.89000 RMSE: 3.44000	**MAE: 0.17000 MSE: 0.28000 RMSE: 0.53000**			
Gupta et al., 2022	MAE: 1100.7714 RMSE: 1442.1927 R2: 0.9652	**MAE: 40.8823 RMSE: 216.4971 R2: 0.9992**	MAE: 644.8759 RMSE: 1131.8468 R2: 0.9785			MAE: 2389.5503 RMSE: 1055.1198 R2: 0.7595	
Jin, 2021	R2: 0.72354	R2: 0.85140	**R2: 0.90416**			R2: 0.83545	
Hankar et al., 2022	RMSE: 63933.52 R2: 0.57000		RMSE: 44939.79 R2: 0.74000	**RMSE: 44516.20 R2: 0.80**			RMSE: 51224.96 R2: 0.70000
H. Zhang, 2022	MAE: 2.5552 MAPE: 0.1652 RMSE: 1.9126		MAE: 1.5769 MAPE: 0.1669 RMSE: 1.8982	MAE: 1.5247 MAPE: 0.1602 RMSE: 1.8103	**MAE: 1.4903 MAPE: 0.1591 RMSE: 1.7739**	MAE: 1.6237 MAPE: 0.1773 RMSE: 1.9132	
Arora et al., 2022	Precisão: 79.86%		Precisão: 90.08%	**Precisão: 91.58%**			Precisão: 37.85%
Narayana et al., 2022	MAE: 9.5351 MSE: 3.1199 RMSE: 6.4762		**MAE: 4.1918 MSE: 0.4069 RMSE: 2.4082**				

Dado o exposto, dos artigos estudados, pode concluir-se que:

- O algoritmo RL foi, na grande maior parte dos algoritmos estudados, o que apresenta piores resultados;
- O algoritmo RF foi considerado como o melhor, sempre que os algoritmos XGBoost e LightGBM não foram testados, com exceção do artigo Gupta et al., 2022, onde o algoritmo AD foi o melhor;
- No caso em que foram testados, simultaneamente, os algoritmos RF e XGBoost, mas em que não foi testado o algoritmo LightGBM, o algoritmo que apresentou melhor performance foi o XGBoost;
- No caso em que foram testados, simultaneamente, os algoritmos XGBoost e LightGBM, o algoritmo que apresentou melhor performance foi o LightGBM.

No que respeita a melhorias de trabalho futuro, existem três que se destacam por terem sido referidas mais do que uma vez pelos autores estudados, nomeadamente:

- Aumento do *dataset* usado e, consequentemente, do número de carros analisados: Narayana et al., 2021, C. Chen et al., 2017, Longani et al., 2021, Jin, 2021 e Hankar et al., 2022;

- Aumento do número de características dos carros: Narayana et al., 2021, Jin, 2021 e Hankar et al., 2022;

- Teste de um maior número de algoritmos:

    - Gupta et al., 2022 sugeriu o teste de algoritmos como KNN e algoritmos genéticos;

    - Jin, 2021 sugeriu o teste de algoritmos como NB, LSTM e XGBoost;

    - H. Zhang, 2022 sugeriu a combinação do algoritmo LightGBM com outros, visto que foi o algoritmo que apresentou melhores resultados.

No que concerne a limitações, e de um modo geral, a principal descrita pelos autores prende-se com o conjunto de dados utilizado visto que, muitas vezes, o mesmo estava incompleto e/ou com dados corrompidos. Tal obrigou os autores a eliminar muitos dos seus dados, comprometendo o tamanho e a qualidade final dos mesmos e, consequentemente, a performance dos algoritmos testados. C. Chen et al., 2017 acrescentaram também que a construção do modelo com o algoritmo de RF foi bastante complicada, devido ao elevado tempo que o mesmo demorou a ser treinado.

2.6 Conclusão

A IA é uma área cada vez mais explorada no mundo atual, tendo sido já desenvolvidas diversas subáreas, nomeadamente reconhecimento de fala, sistemas multiagente, visão por computador, sistemas periciais, PLN, robótica e ML. Relativamente a esta última, existem várias técnicas que podem ser utilizadas dependendo do contexto da sua aplicação, como por exemplo aprendizagem supervisionada (*supervised learning*), não-supervisionada (*unsupervised learning*), semi-supervisionada (*semi supervised learning*) e aprendizagem por reforço (*reinforcement learning*).

Dentro das técnicas referidas, a aprendizagem supervisionada tem sido cada vez mais utilizada na área de sistema de previsão de preços, nomeadamente de carros usados. Neste contexto, já foram testdados diversos algoritmos, sendo os mais relevantes o SVM, LSTM, RL, RP, NB, KNN, AD, GBDT, XGBoost, LightGBM e RF.

Para além destes, as RNA também fazem parte desta técnica, apesar de também poderem ser utilizadas por todas as técnicas acima referidas.

De forma a melhor se perceber que trabalhos já foram desenvolvidos neste âmbito, recorreu-se à metodologia PRISMA.

Dos artigos estudados conclui-se que já foram analisados diversos sistemas de recomendação de preços de carros usados, maioritariamente no mercado estrangeiro. Desses artigos, pôde aferir-se que os algoritmos que obtiveram os melhores resultados foram o RF, XGBoost e LightGBM.

3 Dados e Metodologia

Neste capítulo será apresentada a forma como o conjunto de dados utilizado para a elaboração do modelo desenvolvido foi obtido, bem como as considerações tidas em conta no que respeita à proteção e segurança dos mesmos. Posteriormente, é explicitada a análise e o tratamento de dados efetuados para otimizar os resultados obtidos com o modelo construído.

A seleção dos algoritmos e as ferramentas que serão implementados e usadas, bem como a justificação da escolha dos mesmos, também será apresentada no presente capítulo.

De seguida, será apresentada a metodologia de teste e as métricas utilizadas para avaliar a precisão e a fiabilidade dos resultados do modelo desenvolvido.

Por fim, o capítulo termina com uma breve conclusão das principais ideias passíveis de serem retiradas no seguimento do mesmo.

3.1 Conjunto de Dados

A previsão do custo de carros usados através de Machine Learning (ML) é um tema que já foi desenvolvido por diversos autores, veja-se o capítulo 2.5.1. No entanto, a maior parte dos artigos encontrados na literatura e estudados no âmbito da presente dissertação são relativos ao mercado estrangeiro. De facto, e após uma vasta pesquisa sobre o tema no mercado português, chegou-se à conclusão de que a quantidade de informação publicada neste sentido é inexistente.

Dado o exposto, e de forma a colmatar tal lacuna, procedeu-se à aproximação a várias empresas do ramo automóvel que pudessem fornecer dados para o desenvolvimento do modelo pretendido. Neste sentido, a empresa Standvirtual prontificou-se a fornecer os dados necessários ao desenvolvimento de um modelo de previsão de preço de carros usados aplicados ao contexto do mercado nacional.

Numa primeira fase foram facultados os dados relativos a todas as publicações de anúncios divulgados no Standvirtual, divididos em três ficheiros distintos, referentes aos meses de Outubro, Novembro e Dezembro de 2022. No entanto, e sendo a escassez de dados uma das

principais limitações apontadas pelos autores dos artigos estudados no capítulo 2.5.1, a empresa Standvirtual prontificou-se a disponibilizar mais dados, fornecendo mais cinco ficheiros. Tais ficheiros correspondem aos meses de Julho, Agosto e Setembro de 2022 e a Janeiro e Fevereiro de 2023, aumentando em mais do dobro os dados a analisar.

3.1.1 Proteção de Dados, Segurança e Aspetos Éticos

A área da ML está em constante desenvolvimento e em crescente expansão, transformando e impactando diversos setores. Tal desenvolvimento e crescimento depara-se, indiscutivelmente e cada vez mais, com o desafio da regulamentação do uso de dados e da proteção dos mesmos, uma vez que estes são o principal recurso dos sistemas de ML. Assim, e para garantir a segurança dos dados usados em algoritmos de ML, é importante adotar-se um conjunto de medidas que impeçam o uso abusivo e incorreto dos dados em questão.

Dado o exposto, o presente subcapítulo pretende descrever as medidas de segurança e os aspetos éticos que foram tidos em conta e respeitados durante a análise, tratamento e utilização dos dados fornecidos pela empresa Standvirtual.

Neste contexto, a anonimização dos dados fornecidos pelo Standvirtual foi uma das principais preocupações tidas em consideração, de forma a nunca comprometer a identidade dos utilizadores desta plataforma. Para tal, foram eliminados todos e quaisquer dados pessoais contidos no conjunto de dados fornecidos, primeiramente pela própria empresa e, numa segunda fase, aquando da receção dos dados. Entenda-se, com dado pessoal, toda a "informação relativa a uma pessoa singular identificada ou identificável («titular dos dados»); é considerada identificável uma pessoa singular que possa ser identificada, direta ou indiretamente, em especial por referência a um identificador, como por exemplo um nome, um número de identificação, dados de localização, identificadores por via eletrónica ou a um ou mais elementos específicos da identidade física, fisiológica, genética, mental, económica, cultural ou social dessa pessoa singular" (*Proteção de Dados Pessoais | UCP*, 2016).

Assim, e de modo a não quebrar quaisquer aspetos éticos ou de segurança nos dados fornecidos, foram apenas utilizados nos modelos de dados o *id* dos carros, as características dos mesmos e as suas descrições.

3.1.2 Análise e Tratamento de Dados

É senso comum que a qualidade do resultado dos modelos desenvolvidos começa com a qualidade dos dados que são fornecidos como *input* na etapa de treino dos algoritmos. Assim, e após a receção dos dados fornecidos por parte do Standvirtual, foi realizada uma análise aos mesmos, de modo a se perceber qual a melhor abordagem a aplicar relativamente ao tratamento do conjunto de dados em causa.

Os oito ficheiros fornecidos continham, no total, mais de 58 milhões de linhas de dados com o *id* do veículo, o nome da característica e a descrição da mesma, de uma forma totalmente aleatória e sem qualquer padrão (Figura 8).

```
1   listing_nk  feature value
2   8080156212  pre_crash_system    0
3   8080156212  rear_transversal_curtain_airbag 0
4   8080615330  leather_steering_wheel  1
5   8080615330  side_pre_crash_system   0
6   8082630641  soundsystem 0
7   8082630641  brake_assist    0
8   8082630641  electric_parking_brake  0
9   8083123433  hill_descent_control    0
10  8085084973  armrest_front   1
11  8085519752  origin  national
12  8085764990  mileage 60000
13  8085764990  active_lane_change_assistant    0
14  8085806520  internet_access 0
15  8085806520  passenger_airbag    1
16  8085806520  central_airbag_driver_and_passenger 0
17  8085806520  side_airbag_driver  0
18  8085859101  city_emergency_brake_assist 0
19  8085859101  active_driver_conditioning_monitoring   0
20  8085859552  seat_belt_airbag_rear   0
21  8086007507  digital_key 0
22  8086495368  head_up_display 1
23  8086495368  rear_view_camera    1
24  8086495368  front_cowbar    0
25  8086495368  authorized_dealer   0
26  8086597196  keyless_go  0
27  8086597196  lane_control_assistant  0
28  8086664962  vehicle_class   class_2
```

Figura 8 – Excerto do ficheiro fornecido pelo Standvirtual relativamente aos dados de anúncios publicados durante o mês de Fevereiro de 2023

Dado o exposto, houve necessidade de se organizar os dados fornecidos numa única tabela, de modo a que todas as descrições das características de um carro associadas a um mesmo *id* fossem colocadas na mesma linha, separadas por colunas (Tabela 9). Para tal, foram realizados os seguintes passos:

- Substituição de espaços por vírgulas e transformação dos ficheiros *tsv* para *csv*;
- Junção dos ficheiros relativos a Julho, Agosto, Setembro, Outubro, Novembro e Dezembro de 2022 e Janeiro e Fevereiro de 2023 num único ficheiro;
- Eliminação de dados duplicados;
- Iniciação dos *ID's* previamente facultados a 0;
- Alteração do nome das características fornecidas para português.

Após esta organização pôde-se constatar que foram obtidos dados de 198107 carros com 230 características.

Tabela 9 – Exemplo da tabela elaborada após organização dos dados fornecidos

Id	Ano	Marca	Modelo	…
0	2001	Fiat	Punto	…
1	2007	Renault	Clio	…
2	2013	BMW	320d	…
3	2018	AUDI	A5	…
…	…	…	…	…

Seguidamente, procedeu-se à divisão dos dados em três categorias diferentes:

- ***Conjunto de dados A*:**
    - Para o *Conjunto de dados A* teve-se em consideração as 50 características mais comuns a todos os carros.
- ***Conjunto de dados B*:**
    - Para o *Conjunto de dados B* teve-se em consideração 30 das 50 características mais comuns a todos os carros.
- ***Conjunto de dados C*:**
    - O *Conjunto de dados C* contém apenas 10 das 30 características acima referidas. Estas 10 características foram selecionadas cruzando a informação fornecida pelo Standvirtual com a informação retirada do estudo previamente realizado em 2.5.1. De facto, as características mais utilizadas pelos vários artigos mencionados são 12: Ano do Carro; Cilindrada; Tipo de Combustível; Consumo; Marca; Modelo; Potência; Preço; Quilometragem; Tipo de Caixa; Tipo de Veículo e Tipo de Vendedor. No entanto, as características Consumo e Tipo de Vendedor, apesar de fazerem parte das 238 características iniciais, não apresentam informação suficiente para serem utilizadas, pelo que foram excluídas.

Após a divisão do *dataset* nos conjuntos de dados supracitados, procedeu-se à eliminação de todos os carros em que pelo menos uma das características não continha qualquer tipo de descrição ou continha dados corrompidos, de modo a evitar possíveis erros por parte dos algoritmos, devido a dados incompletos ou ilegíveis. Após o referido tratamento, o conjunto de dados A ficou reduzido a 57038 carros, o conjunto de dados B a 70253 carros e o conjunto de dados C a 192799 carros.

A Tabela 10 compila as diversas características utilizadas nos vários conjuntos de dados A, B e C. As características do conjunto C pertencem ao conjunto B, que por sua vez pertence ao conjunto A.

Tabela 10 – Características dos Conjuntos A, B e C.

Conjunto A	**Conjunto B**	**Conjunto C**	Tipo de Carro	Cilindrada	Versão
			Combustível	Quilometragem	Airbag Condutor
			Potência	Ano	Vidros Elétricos
			Modelo	Preço	Detalhe Versão
			Tipo de Caixa	Marca	Volante Multi Funções
		Cor	Mês	ESP	Número portas
		Negociável	Garantia	Capacidade	Sistema de Travagem Assistida
		Airbag Passageiros	Direção Assistida	Retoma	Financiamento
		Fixação Cadeira Crianças	Volante em Pele	A/C	
	Luzes Diurnas	Bluetooth	Portas USB	Segunda Chave	Radio
	Faróis de Nevoeiro	Fumadores	Navegação	Valor sem IUC	Espelhos Elétricos
	Airbag Lateral	IVA Dedutível	Descanso de Braço	Sensores de Estacionamento	Chave Inteligente
	Sensor de Chuva	Sensor de Pressão dos Pneus	Funções de Radio no Volante		
	Valor sem ISV	Start and Stop			

3.1.2.1 Análise Exploratória dos Dados

O presente subcapítulo visa apresentar os resultados obtidos após uma análise exploratória dos dados tratados no capítulo 3.1.2. O objetivo de tal análise prende-se com a descoberta de padrões ou tendências, a identificação de anomalias, a constatação de possíveis relações entre variáveis e a extração de informação que possa vir a ser útil aquando da aplicação do *dataset* na construção de modelos de previsão e respetiva análise dos seus resultados. Dado o exposto, apresentar-se-ão, de seguida, alguns gráficos elaborados aquando da análise exploratória aplicada aos conjuntos de dados A, B e C.

Através da análise da Figura 9, é possível verificar-se que os *SUV* as carrinhas são o tipo de carro com maior expressão no *dataset* do conjunto de dados A. Por oposição, os cabrio, coupé e mini citadinos são o tipo de carro com menor expressão, representando, no total, menos de 15% dos tipos de carro. Os mesmos resultados também se obtiveram para o conjuntos de dados B e C, com ligeiras diferenças nas percentagens, sendo que no conjunto de dados B, as carrinhas apresentavam uma maior percentagem do que os SUV's. Os respetivos gráficos dos conjuntos B e C encontram-se no Anexo A.

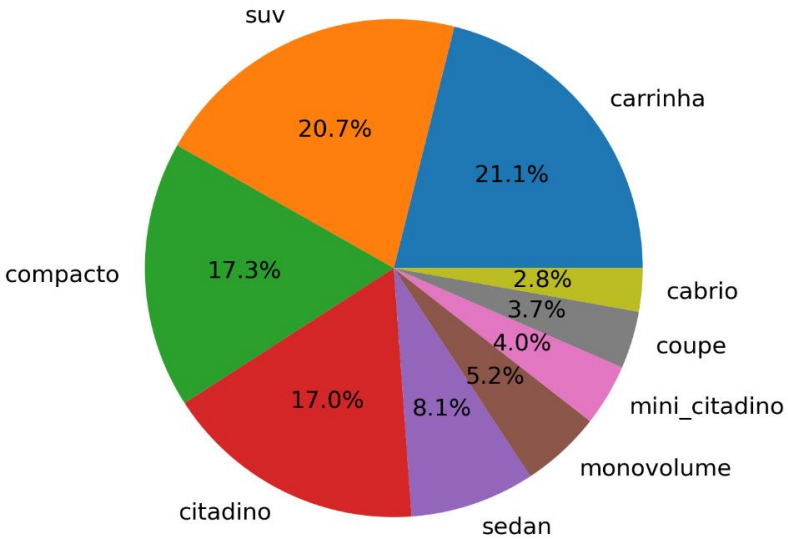


Figura 9 - Percentagem de cada tipo de carro existente no conjunto de dados A

A Figura 10 permite constatar que quase 70% dos carros possuem caixa manual e que cerca de 30% possuem caixa automática no conjunto de dados A. O mesmo resultado também se obteve para os conjuntos de dados B (69.4% manual e 30.6% automática) e C (68.5% manual e 31.5% automática). Os respetivos gráficos dos conjuntos B e C encontram-se no Anexo A.

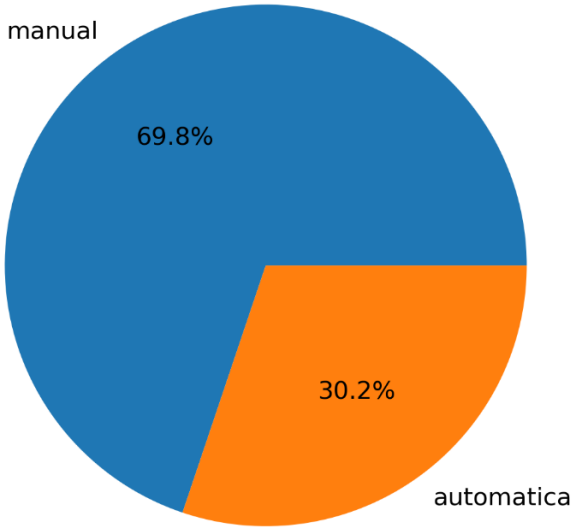


Figura 10 - Percentagem do número de carros com caixa manual e automática existente no conjunto de dados A

A Figura 11 permite verificar que o gasóleo é o tipo de combustível mais utilizado como fonte de energia nos conjuntos de dados A, B e C, uma vez que é usado por 38088 em 57038 carros no conjunto A, por 47689 em 70253 no conjunto de dados B e por 123736 em 192799 carros no conjunto de dados C. Os respetivos gráficos dos conjuntos B e C encontram-se no Anexo A.

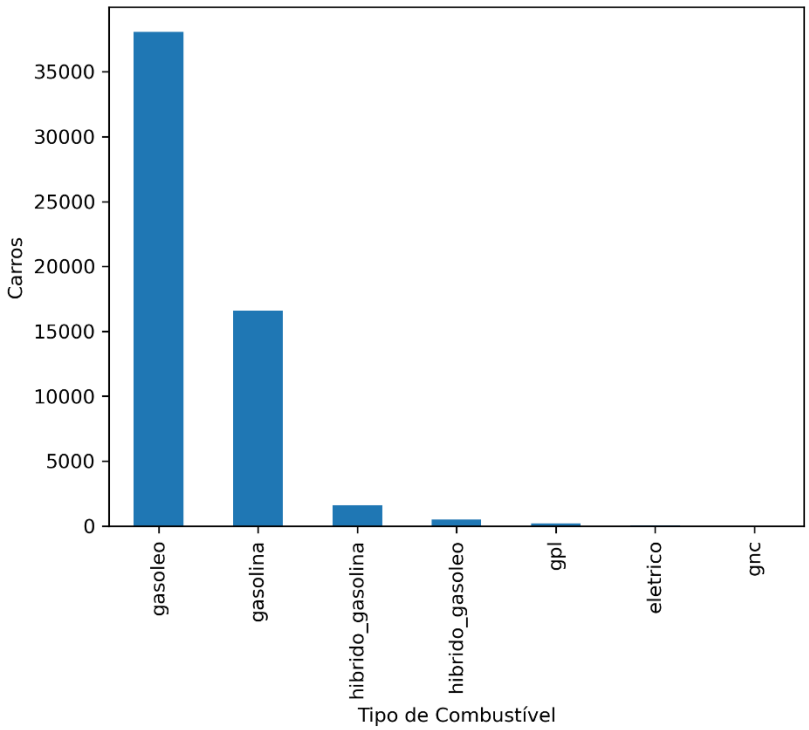


Figura 11 - Número de carros de cada tipo de combustível existente no conjunto de dados A

No que respeita às marcas dos automóveis pertencentes ao conjunto de dados A, a Figura 12 demonstra que a Mercedes-Benz, a Renault, BMW e a Peugeot são as mais populares entre os carros publicados no Standvirtual, constituindo 11,96%, 11,44%, 11,05% e 10,59% do total de carros, respetivamente. Das totais 55 marcas existentes, foram agrupadas 30 na categoria "outras" que equivalem a 3,51% do total de carros. Os resultados obtidos para os conjuntos de dados B e C são muito idênticos aos obtidos com o conjunto de dados A, pelo que os respetivos gráficos não se encontram abaixo espelhados, mas sim no Anexo A.

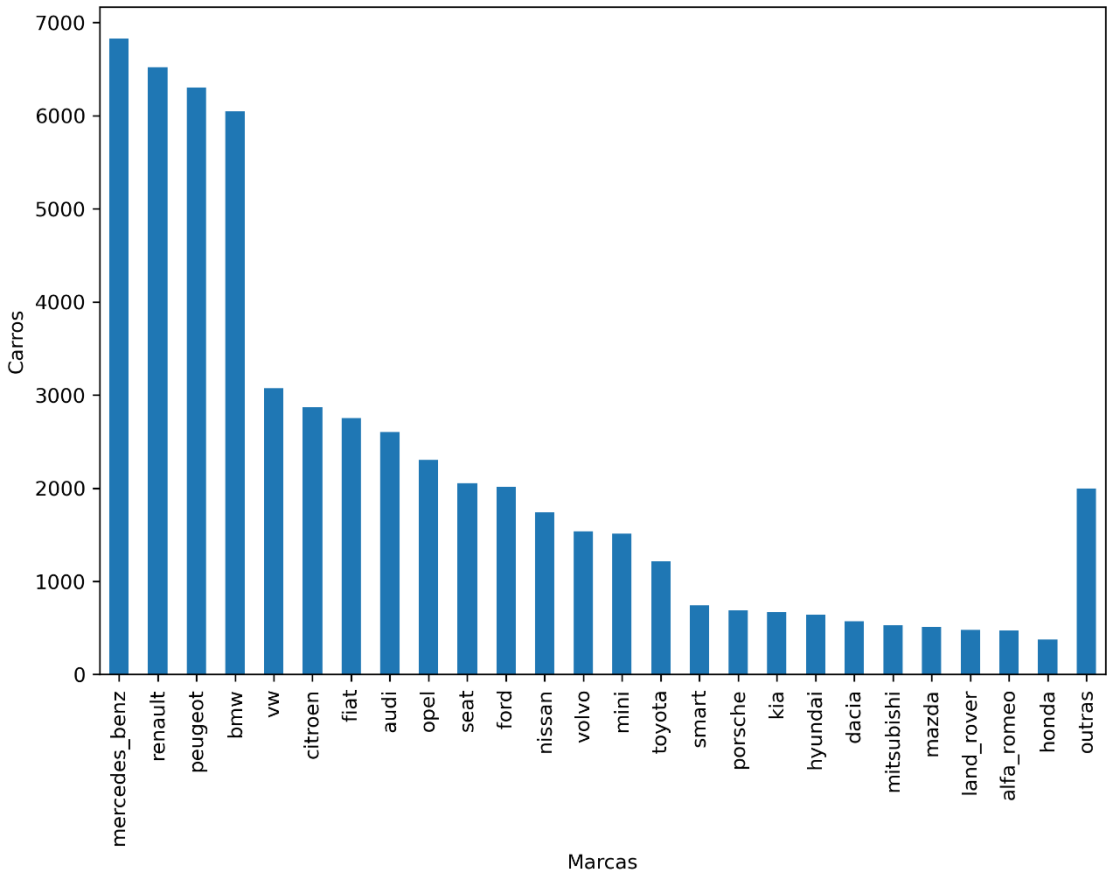


Figura 12 – Número de carros de cada marca existente no conjunto de dados A

3.1.2.2 Codificação de Características Categóricas

A codificação de características categóricas é o processo pelo qual variáveis que se encontram em formato de linguagem alfabética são transformadas em linguagem numérica. Tal codificação é de extrema importância em algoritmos de *machine learning,* já que a grande maioria deles não permite que sejam utilizadas variáveis categóricas na sua execução, uma vez que não é possível efetuar cálculos com variáveis em linguagem alfabética(Potdar et al., 2017).

O método utilizado para a codificação categórica nos conjuntos de dados A, B e C foi o método de *Ordinal Encoder,* uma vez que é dos métodos mais utilizados nos sistemas de previsão estudados. Neste método é atribuído um número inteiro único a cada categoria existente na variável em questão. Ao contrário de outros *encoders*, como o *One-hot encoder*, o *ordinal encoder* não adiciona mais colunas no conjunto de dados, mantendo a sua dimensão original e permitindo uma melhor performance computacional. (Eye & Clogg, 1996). O método *Ordinal Encoder* tem ainda a vantagem de organizar as variáveis consoante o seu peso. No entanto, esta característica não teve qualquer impacto no *dataset* em questão, uma vez que não existe uma relação de hierarquia entre as variáveis do mesmo.

Dado o exposto, foi necessário a aplicação da codificação *ordinal encoder* nas características apresentadas em linguagem alfabética. A Tabela 11 apresenta as características dos conjuntos de dados A, B e C que foram alvo desta codificação.

Tabela 11 – Características alvo de codificação nominal nos conjuntos de dados A, B e C

	Conjunto de dados A	Conjunto de dados B	Conjunto de dados C
Tipo de Carro	X	X	X
Cor	X	X	
Combustível	X	X	X
Marca	X	X	X
Modelo	X	X	X
Moeda	X		
Tipo de caixa	X	X	X
Versão	X	X	
Versão detalhada	X	X	
Tipo de ar condicionado	X	X	

A Tabela 12 apresenta um excerto do conjunto de dados C depois de aplicada a codificação nominal.

Tabela 12 – Excerto do conjunto de dados C depois de aplicada a codificação nominal

ID	Tipo de carro	Cilindrada	Potência	Ano	Combustível	Tipo de caixa	Marca	Quilometragem	Modelo	Preço
0	8	3000	306	2010	0	0	14	82000	1574	40900
1	2	1600	92	2011	0	1	70	90000	120	11990
2	4	5400	625	2009	2	1	34	27800	1125	69900
3	8	6000	325	2006	4	0	37	84500	882	54900
4	3	1995	204	2009	0	0	14	219000	18	14999
5	4	5474	442	2001	2	0	31	18895	180	95000

3.1.2.3 Análise de *Outliers*

Outliers são valores extremos, os quais se encontram muito distantes dos demais valores num conjunto de dados. Tais valores podem resultar de erros humanos ou computacionais, por exemplo aquando da coleção e/ou registo de dados. Os *outliers* podem ter um impacto significativo na análise de dados e em modelos de *machine learning*, uma vez que podem comprometer a precisão dos modelos e, consequentemente, a qualidade e fiabilidade dos resultados obtidos. Assim, torna-se imprescindível a remoção de *outliers*, de forma a garantir que os modelos de *machine learning* são treinados com base em dados fidedignos e representativos (Chaudhary & Lee, 2016).

Dado o exposto, foram eliminados todos os automóveis cujas características cilindrada, potência, quilometragem, ano ou preço apresentavam um valor de *Z-score* superior a ± 3.

As Figuras 13, 14, 15 e 16 apresentam a relação entre o preço e a quilometragem, ano, potência e cilindrada, respetivamente, antes e depois da remoção de *outliers* relativamente ao conjunto de dados A. Os respetivos gráficos relativos aos conjuntos de dados B e C encontram-se no Anexo B.

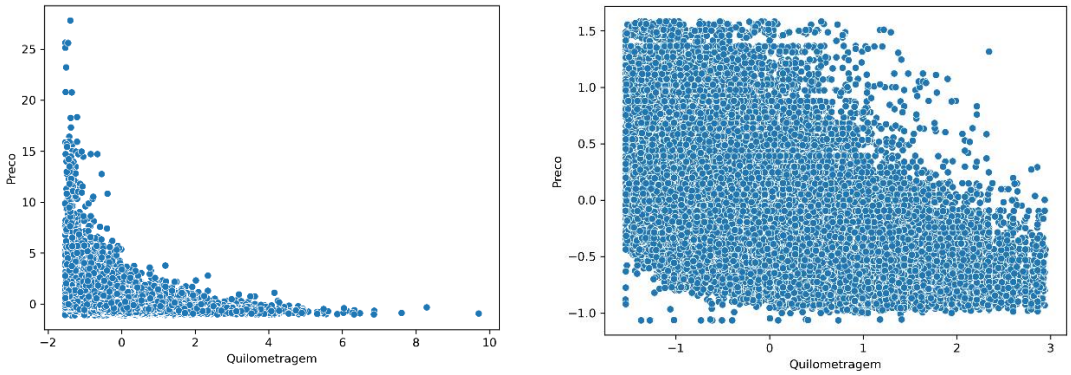


Figura 13 – Relação preço/quilometragem antes (esquerda) e depois (direita) da deteção e remoção de outliers do conjunto de dados A

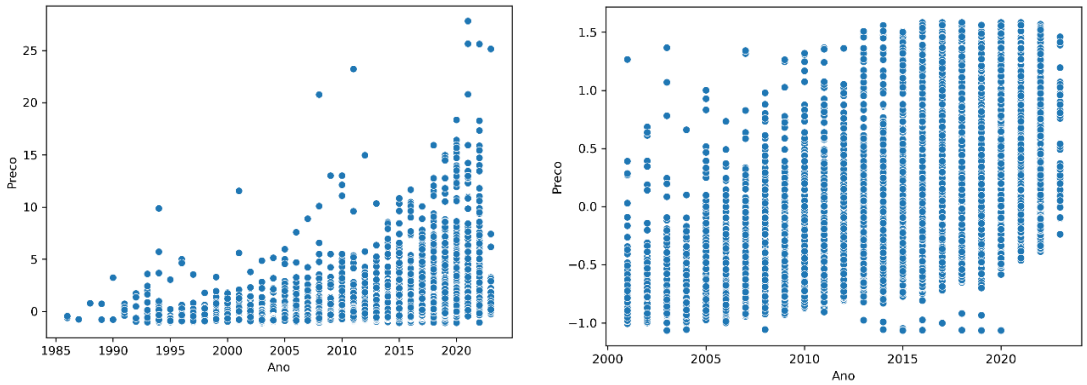


Figura 14 - Relação preço/ano antes (esquerda) e depois (direita) da deteção e remoção de outliers do conjunto de dados A

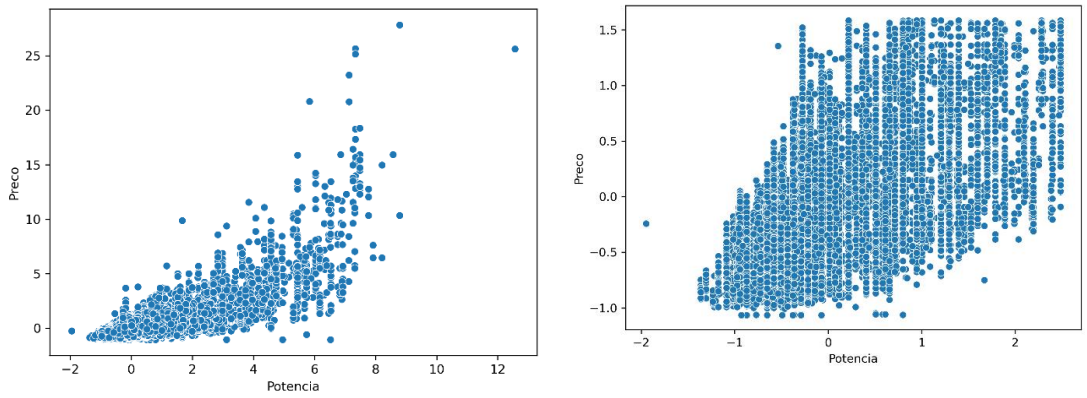


Figura 15 - Relação preço/potência antes (esquerda) e depois (direita) da deteção e remoção de outliers do conjunto de dados A

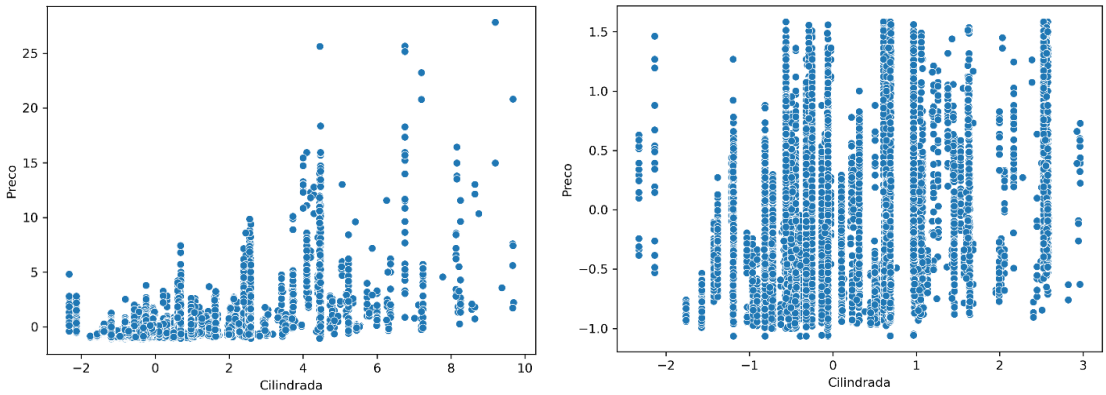


Figura 16 - Relação preço/cilindrada antes (esquerda) e depois (direita) da deteção e remoção de outliers do conjunto de dados A

Antes da remoção dos *outliers* existiam, no conjunto A 57038 carros, no conjunto de dados B 70253 carros e no conjunto de dados C 192799 carros. Após a aplicação do método acima referido, o conjunto de dados A passou a ter um total de 52995 carros, conjunto de dados B 65050 carros e o conjunto de dados C 182119 carros.

3.1.2.4 Normalização de Variáveis

A normalização de variáveis é um processo que visa transformar variáveis numa escala comum e comparável. Como tal, esta normalização é utilizada quando existe uma grande variação nas escalas das variáveis ou quando se pretende comparar diretamente a magnitude das mesmas. A normalização de variáveis é importante, uma vez que permite a comparação de variáveis, melhora o desempenho e a precisão de algoritmos de *machine learning* e facilita a interpretação dos resultados obtidos, garantindo que cada campo contribui igualmente para a construção do modelo (Nkikabahizi et al., 2022).

Existem diversos métodos de normalização de variáveis, sendo que um dos mais comuns é o método da normalização *Z-score*. Tal método baseia-se na diferença entre o valor original da variável e a média dos valores da variável dividido pelo desvio padrão (Nkikabahizi et al., 2022).

Dado o exposto, a normalização de variáveis foi aplicada nos conjuntos de dados A, B e C, nas seguintes características: cilindrada, potência, quilometragem e preço. As restantes características não foram alvo do processo de normalização, uma vez que não existia uma discrepância significativa entre os seus valores ou porque não existia qualquer relação entre os mesmos (por exemplo, valores da marca do automóvel).

A Tabela 13 apresenta um exemplo dos valores das características depois de codificadas e normalizadas, relativas ao conjunto de dados C.

Tabela 13 – Exemplo dos valores das características depois de codificadas e normalizadas relativas ao conjunto de dados C

ID	Tipo de carro	Cilindrada	Potência	Ano	Combustível	Tipo de caixa	Marca	Quilometragem	Modelo	Preço
0	8	2,31646	2,31618	2010	0	0	14	-0,43346	1574	0,61372
1	2	-0,07706	-0,61456	2011	0	1	70	-0,33580	120	-0,33578
2	4	6,41964	6,68490	2009	2	1	34	-1,09504	1125	1,56619
3	8	7,44544	2,57639	2006	4	0	37	-0,40294	882	1,07353
4	3	0,59825	0,91928	2009	0	0	14	1,23883	18	-0,23696
5	4	6,54616	4,17871	2001	2	0	31	-1,20374	180	2,39056
6	8	2,18995	0,54952	2003	0	0	48	1,83958	1241	-0,07271
7	4	3,68419	5,24692	2015	2	0	73	-1,16584	285	4,19692
8	7	7,44202	5,79472	2005	2	0	13	0,25865	586	1,40525

3.1.2.5 Relação entre Características

De forma a melhor se compreender a relação entre características é possível recorrer-se à criação de uma matriz de correlação. Tal matriz é utilizada para identificar padrões entre as diversas variáveis existentes nos conjuntos de dados, mostrando, através de valores numéricos, quais as que têm maior proximidade entre si. A matriz de correlação é composta por valores entre -1 e 1, sendo que o valor 1 indica que as variáveis são totalmente diretamente proporcionais, enquanto que o valor -1 significa que as variáveis são totalmente inversamente proporcionais.

A Figura 17 representa a matriz de correlação aplicada ao conjunto de dados C (a matriz de correlação relativa aos conjuntos de dados A e B encontram-se no Anexo C. – o conjunto de dados A não foi tomado como exemplo, como nos casos anteriores, devido à grande dimensão da respetiva matriz de correlação, a qual impede uma leitura clara dos dados nela contida). Através da análise da mesma pode concluir-se que as variáveis potência e cilindrada, potência e preço e ano do carro e preço apresentam uma maior regressão linear positiva, com valores de 0.71, 0.63 e 0.57, respetivamente. Por oposição, as variáveis ano e quilometragem, peço e tipo de caixa e tipo de caixa e potência apresentam uma maior regressão linear negativa, sendo que quando o valor de uma variável aumenta, o da outra diminui.

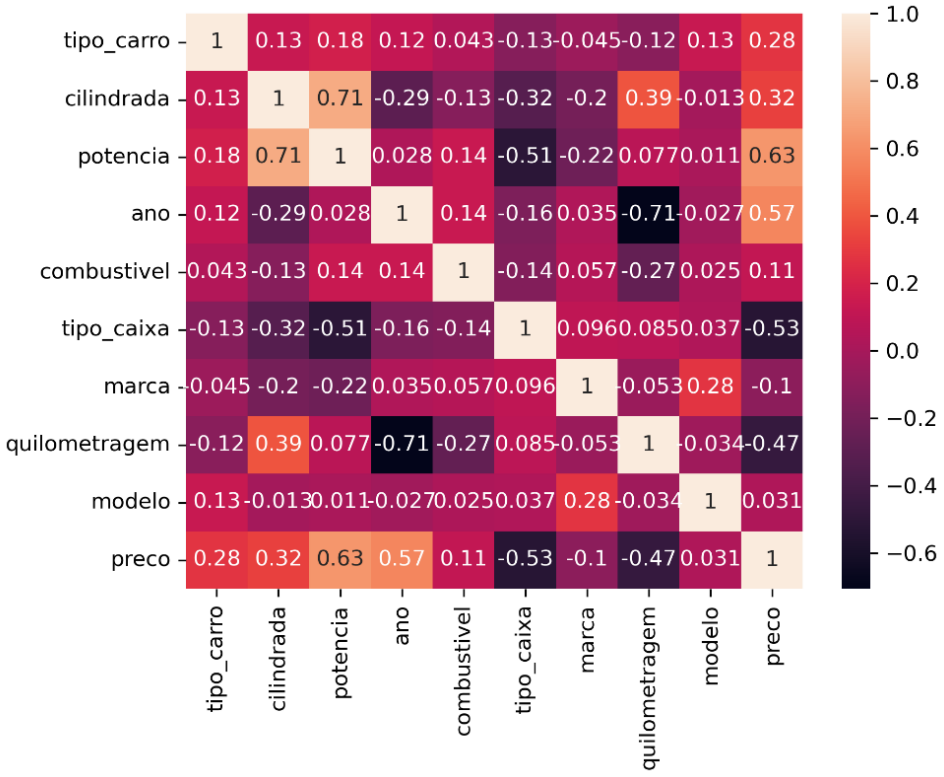


Figura 17 – Matriz de correlação do conjunto de dados C

3.2 Metodologia de Treino, Validação e Teste

No processo de criação do modelo de previsão de preço de carros usados irão ser testados os algoritmos RF, XGBoost, LightGBM e RL. A escolha destes algoritmos deveu-se ao facto de, nos artigos estudados no subcapítulo 2.5.1, os primeiros três terem sido os que obtiveram melhores resultados e o RL o mais comumente usado. Para além destes, procedeu-se ainda aos testes das redes neuronais MLP e CNN, visto que são também duas técnicas utilizadas em problemas de modelos de previsão.

De forma a se poder avaliar e verificar o desempenho, a eficácia e a robustez de um determinado modelo, aplicar-se-á uma metodologia de treino, validação e testes.

Todos os algoritmos irão ser testados através de técnicas de treino, validação e teste, com uma percentagem de 80% para treino e 20% para teste, sendo que a percentagem de validação se encontra englobada na percentagem correspondente à fase treino. O conjunto de treino é, como o próprio nome indica, utilizado para treinar o modelo, enquanto que o conjunto de testes é usado para avaliar o desempenho final do modelo, simulando a sua aplicação em dados que não foram utilizados na fase de treino. Na fase de validação são ajustados os hiperparâmetros, os quais variam consoante o algoritmo que se está a estudar, e os quais são utilizados para melhorar o desempenho do modelo.

Cada um dos algoritmos dos três conjuntos de dados A, B e C será treinado, validado e testado duas vezes: com aplicação de hiperparâmetros padrão (mais comumente usados) e com aplicação de hiperparâmetros introduzidos manualmente antes da fase de treino. Os hiperparâmetros utilizados nos diversos algoritmos estudados foram os seguintes:

- ***n_estimator*:** representa o número de árvores de decisão que são criadas pelo algoritmo;
- ***max_depth*:** controla a profundidade máxima das árvores de decisão;
- ***min_samples_split*:** define o número mínimo de amostras necessárias num nó para que a divisão seja considerada;
- ***colsample_bytree:*** controla a fração de variáveis que serão amostradas aleatoriamente ao construir cada árvore durante o processo de treino do modelo;
- ***min_child_weight*:** controla a profundidade mínima de um nó filho com base no peso total dos exemplos;
- ***min_chil_sample:*** defini o peso mínimo necessário para criar uma nova partição num nó durante o treino da árvore;
- ***min_sample_leaf*:** especifica o número mínimo de amostras necessárias numa folha de uma árvore de decisão;
- ***learning_rate*:** controla a velocidade com que o modelo aprende a partir dos erros cometidos durante a fase de treino;
- ***subsample*:** representa a fração de amostras que é usada para treinar cada árvore de decisão.
- ***epoch*:** determina quantas vezes o conjunto completo de dados será utilizado para treinar o modelo.
- ***optimizer*:** ajusta os pesos do modelo durante o treino, de forma a minimizar as perdas o máximo possível;
- ***alpha*:** reduz a complexidade do modelo, retirando peso às variáveis que têm menos relevância**;**
- ***max_iter*:** controla o número máximo de iterações durante a fase de treino do modelo;
- ***reg_alpha*:** adiciona uma penalidade proporcional à soma dos valores absolutos dos coeficientes do modelo, potenciando a seleção das características mais relevantes para a tarefa em causa;
- ***reg_lambda:*** adiciona uma penalidade proporcional à soma dos quadrados dos coeficientes do modelo, de forma a evitar o *overfitting*;
- ***loss function*:** determina a função que calcula a diferença entre as previsões feitas pela rede neural e os valores reais dos dados durante a fase de treino;
- ***activation function*:** determina a função de ativação que será aplicada aos resultados intermediários de cada neurónio na rede neural;
- ***filters*:** extrai características relevantes dos dados;
- ***filter_size*:** define a quantidade de pontos que os filtros consideram para a CNN;
- ***pool_size*:** reduz a dimensionalidade das características, mantendo as mais relevantes para a CNN;
- ***hidden layers size*:** indica o número de neurónios na camada oculta da CNN.

3.3 Conclusão

O conjunto de dados utilizados para o desenvolvimento do modelo estudado no âmbito da presente dissertação foi fornecido pela empresa Standvirtual, assegurando a proteção e segurança dos respetivos dados. Tais dados continham um conjunto de características e respetivas descrições de mais de 198.000 carros.

De modo a se assegurar o melhor resultado possível do modelo desenvolvido no contexto do sistema de previsão de carros usados, foi realizada uma análise e aplicado um tratamento ao conjunto de dados recebidos. Tal tratamento passou, essencialmente, pela organização dos dados numa tabela, pela eliminação de carros cujas descrições das características estavam ausentes ou corrompidas, pela conversão em linguagem numérica da descrição das características previamente apresentadas em linguagem alfabética, pela remoção de *outliers e* pela normalização dos dados de determinadas características.

O tratamento de dados aplicados aos dados iniciais foi utilizado para maximizar a performance dos algoritmos a ser testados, os quais são RF, XGBoost, LightGBM, RL, MLP e CNN. Todos estes algoritmos foram testados através de técnicas de treino, validação e teste, com hiperparâmetros padrão e introduzidos manualmente antes da fase de treino, tendo sido atribuídos 80% do conjunto inicial de dados à fase de treino e 20% à fase de teste. A fase de validação encontra-se contemplada na fase de treino.

A metodologia de treino, validação e testes em algoritmos de ML e DL é fundamental para garantir que o modelo é capaz de prever corretamente as previsões pretendidas para novos dados, bem como para fornecer uma avaliação objetiva do seu desempenho, ajudando na tomada de decisões acerca da viabilidade de implantação do modelo em cenários reais.

4 Implementação, Análise e Discussão de Resultados

Após a análise e tratamento dos dados fornecidos, procedeu-se ao treino, validação e testes dos algoritmos RF, XGBoost, LightGBM, RL e das redes neuronais MLP e CNN nos conjuntos A, B, C, primeiramente através da utilização de hiperparâmetros padrão e, posteriormente, através hiperparâmetros modificados, com o intuito de se melhorar a performance dos algoritmos mencionados. A implementação dos algoritmos mencionados foi feita no computador HP EliteBook 840 G6 (Processador: intel i7 8565U, RAM: 16GB, Gráfica: intel UHD Graphics 620).

Os resultados obtidos foram comparados entre si e com os resultados apresentados nos artigos previamente estudados, detalhados no subcapítulo Sistemas de Previsão de Preços de Carros Usados.

A performance dos algoritmos foi avaliada através das métricas R2, RMSE, MAE e MSE, uma vez que também elas foram as mais utilizadas pelos autores estudados no subcapítulo Sistemas de Previsão de Preços de Carros Usados.

Os subcapítulos seguintes pretendem detalhar os resultados obtidos, bem como apresentar a análise dos mesmos e respetiva discussão.

4.1 Hiperparâmetros Padrão

Os algoritmos RF, XGBoost, LightGBM e as redes neuronais MLP e CNN foram primeiramente testados com hiperparâmetros padrão, ou seja, comumente usados nos respetivos algoritmos. Os hiperparâmetros padrão utilizados nos algoritmos RF, XGBoost, LightGBM, MLP e CNN encontram-se detalhados na Tabela 14. O algoritmo RL não possui hiperparâmetros padrão.

Tabela 14 - Hiperparâmetros padrão usados em cada um dos algoritmos RF, XGBoost, LightGBM e MLP, para o conjunto de dados A, B e C

	RF	XGBoost	LightGBM	MLP	CNN
n_estimator	100	100	100		
max_depth	None	3	None		
min_sample_split	2				
min_child_sample			20		
min_chil_weight		1			
min_sample_leaf	1				
colsample_bytree		1.0	1.0		
learning_rate		0.1	0.1		
subsample		1.0	1.0		
epoch				10	100
optimizer				adam	adam
reg_alpha		0.0	0.0		
reg_lambda		1.0	0.0		
loss function				MSE	MSE
activation function				Rectified Linear Unit	Rectified Linear Unit
filters					32
filter_size					3
pool_size					1
hidden layers size					100

Os resultados dos modelos obtidos após a execução dos algoritmo com os hiperparâmetros padrão acima descritos estão representados nas Tabelas 15, 16 e 17.

Tabela 15 – Resultados obtidos para os algoritmos (hiperparâmetros padrão) RF, XGBoost, LightGMB, RL, MLP e CNN para o conjunto de dados A

	MAE	RMSE	MSE	R2
RF	**0.14527**	0.21827	0.04764	0.95264
XGBoost	0.14820	**0.20979**	**0.04401**	**0.95625**
LightGBM	0.15990	0.22537	0.05079	0.94952
RL	0.29203	0.39574	0.15661	0.84434
MLP	0.18573	0.26325	0.06930	0.93112
CNN	0.18789	0.77191	0.59584	0.40864

Tabela 16 - Resultados obtidos para os algoritmos (hiperparâmetros padrão) RF, XGBoost, LightGMB, RL, MLP e CNN para o conjunto de dados B

	MAE	RMSE	MSE	R2
RF	**0.14306**	0.22742	0.05172	0.94741
XGBoost	0.15267	**0.22385**	**0.05011**	**0.94905**
LightGBM	0.16758	0.24155	0.05834	0.94067
RL	0.30468	0.41454	0.17185	0.82527
MLP	0.19475	0.28929	0.08369	0.91491

CNN	0.17887	0.82033	0.67294	0.31297

Tabela 17 - Resultados obtidos para os algoritmos (hiperparâmetros padrão) RF, XGBoost, LightGMB, RL, MLP e CNN para o conjunto de dados C

	MAE	**RMSE**	**MSE**	**R2**
RF	**0.12900**	**0.21190**	**0.04490**	**0.95492**
XGBoost	0.14210	0.22154	0.04907	0.95073
LightGBM	0.16003	0.24517	0.06011	0.93964
RL	0.31965	0.46686	0.21796	0.78118
MLP	0.16082	0.25463	0.06484	0.93491
CNN	0.14852	0.23309	0.05433	0.94540

A análise das Tabelas 15, 16 e 17 permite constatar que:

- O melhor resultado do algoritmo RF, MLP e CNN foi obtido no conjunto de dados C, seguindo-se do conjunto de dados A e por fim B;
- O melhor resultado do algoritmo XGBoost foi obtido no conjunto de dados A, seguindo-se do conjunto de dados C e por fim B;
- O melhor resultado dos algoritmos LightGBM e RL foram obtidos no conjunto de dados A, seguindo-se do conjunto de dados B e por fim C;
- Os algoritmos RF, XGBoost e LightGBM apresentam resultados muito próximos nos conjuntos da dados A, B e C;
- O XGBoost é o que apresenta, em três das quatro métrica utilizadas, melhores resultados no conjunto de dados A e B;
- No conjunto de dados C, o que apresenta melhores resultados em todas as métricas estuadas é o RF;
- O algoritmo XGBoost aplicado ao conjunto A é o que apresenta o melhor resultado, quando comparado com os restantes algoritmos dos conjuntos A, B e C;
- O algoritmo RL é o que apresenta piores resultados no conjunto de dados C;
- O algoritmo CNN é o que apresenta piores resultados no conjunto de dados A e B;
- Os resultados do algoritmos CNN melhoram no conjunto de dados C, quando comparados com os resultados obtidos nos conjuntos de dados A e B.

4.2 Hiperparâmetros Modificados

De forma a se melhorar a performance dos algoritmos estudados, procedeu-se à modificação de alguns hiperparâmetros nos mesmos, nos conjuntos de dados A, B e C, através do uso da função *GridSearchCV*, com um valor de *cross validation* de 5.

Os valores dos hiperparâmetros alterados previamente à fase de treino nos diversos algoritmos encontram-se detalhados na Tabela 18. Para os restantes hiperparâmetros utilizaram-se os valores padrão descritos na Tabela 14.

Tabela 18 – Hiperparâmetros introduzidos em cada um dos algoritmos, para o conjunto de dados A, B e C

	RF	XGBoost	LightGBM	RL	MLP	CNN
n_estimators	100	100	100			
	500	500	500			
	1000	1000	1000			
max_depth	0	3	3			
	5	5	5			
	10	7	7			
min_sample_split	2					
	5					
	10					
min_sample_leaf	1					
	2					
	4					
learning_rate		0.1	0.1			
		0.01	0.01			
		0.001	0.001			
subsample		0.8	0.8			
		1	1			
epoch					50	100
					100	300
					300	500
optimizer					Rmsprop	Rmsprop
					Adam	Adam
alpha				0.01		
				10		
				100		
max_inter				100		
				500		
				1000		
filters						64
filter_size						2
						3

Os resultados dos modelos obtidos após a execução dos algoritmos com os hiperparâmetros modificados acima descritos estão representados nas Tabelas 19, 20 e 21.

Tabela 19 - Resultados obtidos para os algoritmos (com hiperparâmetros modificados) RF, XGBoost, LightGMB, RL, MLP e CNN para o conjunto de dados A

	MAE	RMSE	MSE	R2
RF	0.14407	0.21701	0.04709	0.95319
XGBoost	**0.12892**	**0.18947**	**0.03590**	**0.96432**
LightGBM	0.13270	0.19168	0.03674	0.96348
RL	0.29861	0.40594	0.16478	0.83621
MLP	0.18674	0.27406	0.07511	0.92535
CNN	0.19355	0.62368	0.38898	0.61395

Tabela 20 - Resultados obtidos para os algoritmos (com hiperparâmetros modificados) RF, XGBoost, LightGMB, RL, MLP e CNN para o conjunto de dados B

	MAE	RMSE	MSE	R2
RF	0.14095	0.22498	0.05062	0.94854
XGBoost	**0.12901**	**0.19792**	**0.03917**	**0.96017**
LightGBM	0.13514	0.20249	0.04100	0.95831
RL	0.30771	0.41977	0.17620	0.82084
MLP	0.17752	0.26420	0.06980	0.92903
CNN	0.16983	0.48837	0.23850	0.75650

Tabela 21 - Resultados obtidos para os algoritmos (com hiperparâmetros modificados) RF, XGBoost, LightGMB, RL, MLP e CNN para o conjunto de dados C

	MAE	RMSE	MSE	R2
RF	0.12778	0.20824	0.04336	0.95646
XGBoost	**0.12585**	**0.19868**	**0.03947**	**0.96036**
LightGBM	0.13117	0.20726	0.04296	0.95687
RL	0.31915	0.46786	0.21889	0.78024
MLP	0.14475	0.22834	0.05214	0.94766
CNN	0.14459	0.22753	0.05177	0.94798

Os hiperparâmetros que conduziram à melhor performance dos diversos modelos encontram-se explicitados na Tabela 22. Na coluna *Resultados* desta tabela, entenda-se por:

- RA, RB, RC: Algoritmo *Random Forest* aplicado ao conjunto de dados A, B e C, respetivamente;
- XA, XB, XC: Algoritmo *XGBoost* aplicado ao conjunto de dados A, B e C, respetivamente;
- LA, LB, LC: Algoritmo *LightGBM* aplicado ao conjunto de dados A, B e C, respetivamente;

- RLA, RLB, RLC: Algoritmo *Regressão Linear* aplicado ao conjunto de dados A, B e C, respetivamente;
- MLPA, MLPB, MLPC: Algoritmo *Multilayer Perceptron* aplicado ao conjunto de dados A, B e C, respetivamente.
- CNNA, CNNB, CNNC: Algoritmo *Convolutional Neural Networks* aplicado ao conjunto de dados A, B e C, respetivamente.

Tabela 22 – Hiperparâmetros modificados que conduziram aos melhores resultados dos vários algoritmos para cada conjunto de dados A, B e C.

	RF (R)	**XGBoost (X)**	**LightGBM (L)**	**RL**	**MLP**	**CNN**	**Resultados**
n_estimators	500	500	500				RC
	1000	1000	1000				RA/RB/XA/XB/XC/LA/LB/LC
max_depth	0						RA/RB/RC
		7	7				XA/XB/XC/LA/LB/LC
min_sample_split	2						RA/RB
	5						RC
min_sample_leaf	1						RA/RB/RC
subsample		0.8	0.8				XA/XB/LA/LB/LC
		1					XC
epoch					100	100	MLPA/MLPB/CNNA/CNNB
					300	300	MLPC
						500	CNNC
optimizer					Rmsprop	Rmsprop	MLPA/MLPB/CNNA/CNNB
					Adam	Adam	MLPC/CNNC
alpha				0.01			RLA/RLB/RLC
max_inter				100			RLC
				500			RLA/RLB
filters						64	CNNA/CNNB/CNNC
filter_size						2	CNNA/CNNB/CNNC

Através das Tabelas 19, 20, 21 e 22 é possível constatar-se que:

- Relativamente ao algoritmo RF:

- O melhor resultado do algoritmo foi obtido no conjunto de dados C, seguindo-se do conjunto de dados A e por fim B;
- Os hiperparâmetros que conduziram aos resultados apresentados foram:
  - *n_estimators*=1000 e *min_sample_split*=2 para os conjuntos A e B;
  - *n_estimators*=500 e *min_sample_split*=5 para o conjunto C;
  - *max_depth*=0 e *min_sample_leaf*=1 para os conjuntos A, B e C;

- Relativamente ao algoritmo XGBoost:
  - O melhor resultado do algoritmo foi obtido no conjunto de dados A, seguindo-se do conjunto de dados C e por fim B;
  - Os hiperparâmetros que conduziram aos resultados apresentados foram:
    - *n_estimators*=1000, *max_depth*=7 e *learning_rate*=0.1 para os conjuntos A, B e C;
    - *subsample*=0.8 para os conjuntos A e B;
    - *subsample*=1 para o conjunto C;

- Relativamente ao algoritmo LightGBM:
  - O melhor resultado do algoritmo foi obtido no conjunto de dados A, seguindo-se do conjunto de dados B e por fim C;
  - Os hiperparâmetros que conduziram aos resultados apresentados foram:
    - *n_estimators*=1000, *max_depth*=7, *learning_rate*=0.1 e *subsample*=0.8 para os conjuntos A, B e C;

- Relativamente ao algoritmo RL:
  - O melhor resultado do algoritmo foi obtido no conjunto de dados A, seguindo-se do conjunto de dados B e por fim C;
  - Os hiperparâmetros que conduziram aos resultados apresentados foram:
    - *alpha*=0.01 para os conjuntos A, B e C;
    - *max_inter*=500 para os conjuntos A e B;
    - *max_inter*=500 para o conjunto C;

- Relativamente ao algoritmo MLP:

    - O melhor resultado do algoritmo foi obtido no conjunto de dados C, seguindo-se do conjunto de dados B e por fim A;
    - Os hiperparâmetros que conduziram aos resultados apresentados foram:
        - *epoch*=100 para os conjuntos A e B;
        - *epoch*=300 para o conjunto C;
        - *optimizer*=Rmsprop para os conjuntos A e B;
        - *optimizer*=Adam para o conjunto C;
- Relativamente ao algoritmo CNN:
    - O melhor resultado do algoritmo foi obtido no conjunto de dados C, seguindo-se do conjunto de dados B e por fim A;
    - Os hiperparâmetros que conduziram aos resultados apresentados foram:
        - *epoch*=100 para os conjuntos A e B;
        - *epoch*=500 para o conjunto C;
        - *optimizer*=Rmsprop para os conjuntos A e B;
        - *optimizer*=Adam para o conjunto C;
        - *filters*=64 e *filter_size*=2 para os conjuntos A, B e C
- Os algoritmos RF, XGBoost e LightGBM apresentam resultados muito próximos nos conjuntos de dados A, B e C. No entanto, a velocidade de execução do modelo com o algoritmo LighGBM é mais rápida do que a do algoritmo XGBoost o qual, por sua vez, é bastante mais rápida do que a velocidade do algoritmo RF;
- O algoritmo CNN é o que apresenta piores resultados nos conjuntos de dados A e B, melhorando novamente no conjunto de dados C;
- O algoritmo RL é o que apresenta piore resultado no conjunto de dados C;
- O XGBoost é o que apresenta, em todas as métricas, melhores resultados nos conjuntos de dados A, B e C;
- O algoritmo XGBoost aplicado ao conjunto A é o que apresenta o melhor resultado, quando comparado com os restantes algoritmos dos conjuntos A, B e C.

Após a análise realizada é possível aferir-se que o algoritmo XGBoost com hiperparâmetros modificados face aos hiperparâmetros padrão é o algoritmo que apresenta melhor performance no três conjuntos de dados A, B e C. Para além disso, é possível comprovar-se que o conjunto de dados A é o que apresenta melhores resultados, seguindo-se do conjunto de

dados C e, por fim, o conjunto de dados B, tal como se pode verificar através da Tabela 23. Se se analisar o número de carros e o número de características de cada conjunto (o conjunto de dados A contém 50 características e 57038 carros, o conjunto de dados B contém 30 características e 70253 carros e o conjunto C contém 10 características e 192799 carros) pode especular-se que o algoritmo XGBoost privilegia um maior número de características face a um maior número de carros mas que, quando o número de características diminui para um determinado valor, o algoritmo passa a privilegiar o número de carros de que dispõe.

Tabela 23 – Resultados do algoritmo XGBoost com hiperparâmetros modificados, para cada um dos conjuntos de dados A, B e C

	MAE	RMSE	MSE	R2
Conj. Dados A	0.12892	0.18947	0.03590	0.96432
Conj. Dados B	0.12901	0.19792	0.03917	0.96017
Conj. Dados C	0.12585	0.19868	0.03947	0.96036

4.3 Comparação e Discussão de Resultados

A alteração de determinados hiperparâmetros padrão conduziu a uma melhoria dos resultados dos algoritmos RF, XGBoost, LightGBM e CNN dos conjuntos de dados A, B e C, bem como do algoritmo MLP dos conjuntos de dados B e C. Apenas o algoritmo MLP do conjunto de dados A e o algoritmo RL dos conjuntos A, B e C obtiveram piores resultados com a alteração de determinados hiperparâmetros. Tabela 24 representa a taxa de variação dos resultados obtidos na métrica R2 com a utilização de hiperparâmetros padrão e hiperparâmetros modificados.

Tabela 24 – Taxa de variação dos resultados obtidos na métrica R2 nos algoritmos testados com hiperparâmetros padrão e hiperparâmetros modificados

	RF	XGBoost	LightGBM	RL	MLP	CNN
Conjunto A	0.05500%	0.80700%	1.39600%	-0.81300%	-0.57700%	20.53100%
Conjunto B	0.11300%	1.11200%	1.76400%	-0.44300%	1.41200%	44.35300%
Conjunto C	0.15400%	0.96300%	1.72300%	-0.09400%	1.37500%	0.25800%

A Tabela 25 apresenta os hiperparâmetros que foram alterados face aos hiperparâmetros padrão e que resultaram numa melhor (sublinhado a verde) e pior (sublinhado a vermelho) performance dos algoritmos executados nos conjuntos de dados A, B e C. Os hiperparâmetros que não se encontram explicitados na referida tabela obtiveram a melhor performance com os valores padrão.

Tabela 25 - Alteração dos hiperparâmetros que conduziram aos melhores resultados dos algoritmos RF, XGBoost, LightGBM e MLP nos conjuntos de dados A, B e C

			Nr estimator	Max_depth	Min_sample_split	Subsample	Epoch	Optimizer	Alpha	Max_inter	Filters	Filter_size
Hiperparâmetros Padrão	RF		100	0	2	1.0						
	XG Boost			3								
	Light GBM			0								
	MLP						10	Adam				
	CNN						100				32	3
Hiperparâmetros Modificados	RF	A	1000	0	2							
		B										
		C	500		5							
	XG Boost	A	1000	7		0.8						
		B										
		C				1.0						
	Light GBM	A	1000	7		0.8						
		B										
		C										
	RL	A							0.01	500		
		B										
		C								100		
	MPL	A					100	Rmsprop				
		B					100	Rmsprop				
		C					300	Adam				
	CNN	A					100	Rmsprop			64	2
		B										
		C					500	Adam				

Através da análise desta tabela é possível especular-se que:

- O aumento do hiper parâmetro *n_estimator* conduziu a melhores resultados nos algoritmos RF, XGBoost e LightGBM nos três conjuntos de dados A, B e C;
- O aumento do hiper parâmetro *max_depth* conduziu a melhores resultados nos algoritmos XGBoost e LightGBM nos três conjuntos de dados A, B e C;
- O aumento do hiper parâmetro *min_sample_split* conduziu a melhores resultados no algoritmo RF no conjunto de dados C;
- A diminuição do hiper parâmetro *subsample* conduziu a melhores resultados nos algoritmos XGBoost no conjunto de dados A e B e LightGBM nos três conjuntos de dados A, B e C;
- O aumento do hiper parâmetro *epoch* conduziu a melhores resultados no algoritmo MLP nos conjuntos de dados B e C, e a piores resultados no conjunto de dados A;

- O aumento do hiper parâmetro *epoch* conduziu a melhores resultados no algoritmo CNN no conjunto de dados C;
- A alteração do hiper parâmetro *optimizer* conduziu a melhores resultados no algoritmo MLP no conjunto de dados B e no algoritmo CNN para os conjuntos de dados A e B;
- A introdução dos hiperparâmetros *alpha* e *max_inter* conduziu a piores resultados no algoritmo RL nos conjuntos de dados A, B e C.

4.4 Características Mais Preponderantes

Uma vez que o algoritmo XGBoost foi o algoritmo que apresentou melhor resultados, decidiu-se executá-lo para um quarto conjunto de dados, o **conjunto de dados D**. O conjunto de dados D é constituído por 13 características, as quais, para além do preço, foram obtidas através do método *Shapely Additive exPlanations* (SHAP), aplicado ao algoritmo XGBoost do conjunto de dados A (Figura 18). Este método permite a um determinado modelo de ML perceber quais as variáveis que têm mais influência na obtenção da estimativa do preço do mesmo (Van den Broeck et al., 2022).

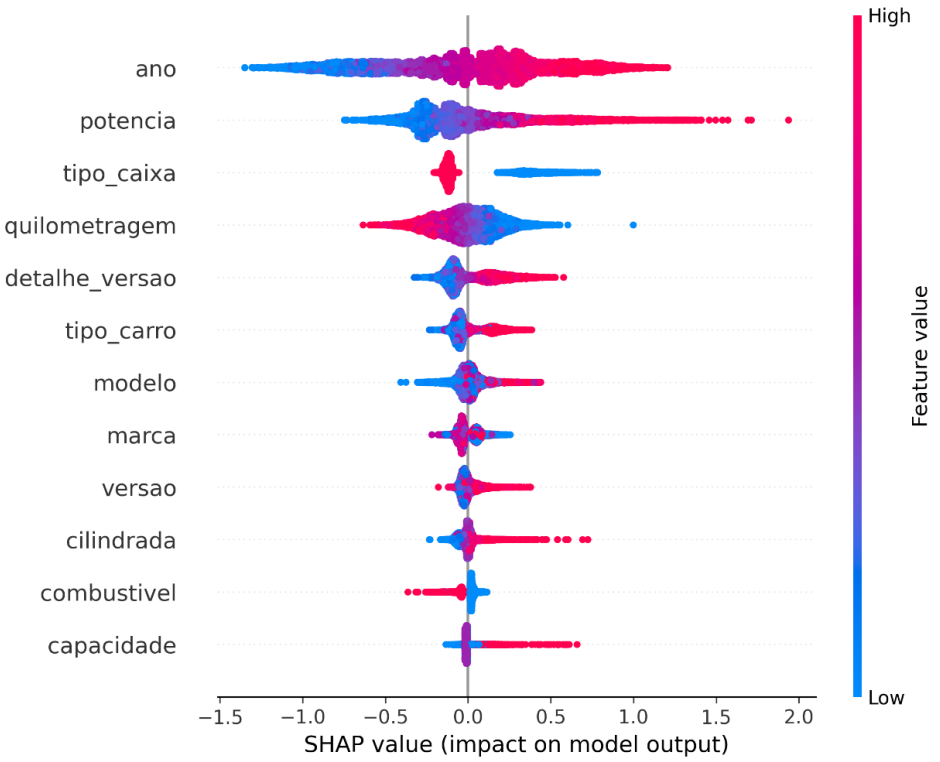


Figura 18 – Gráfico SHAP, o qual representa as características mais revelantes no modelo XGBoost do conjunto de dados A

Dado o exposto, o conjunto de dados D, para além de conter todas as características contidas no conjunto de dados C, contém ainda as características versão, tipo de versão e capacidade.

Após a aplicação do mesmo tratamento de dados efetuado aos conjuntos A, B e C, o conjunto de dados D ficou reduzido a 135701 carros.

O algortimo XGBoost aplicado ao conjunto de dados D foi testado com os mesmos hiperparâmetros descritos na Tabela 18, tendo a melhor performace do algortimo sido obtida com os mesmos hiperparâmetros do algortimo XGBoost aplicado ao conjunto de dado A e B, descritos na Tabela 22. O resultado obtido pelo algortimo XGBoost no conjunto de dados D, bem como a sua comparação com os resultados obtidos no conjunto de dados A, B e C, o número de carros em cada conjunto e a percentagem de variação da métrica R2 face ao conjunto com o melhor resultado nesta métrica são apresentados na Tabela 26.

Tabela 26 - Resultados do algoritmo XGBoost com hiperparâmetros modificados para cada um dos conjuntos de dados A, B, C e D

	Nº Auto	MAE	RMSE	MSE	R2	% variação
XGBoost A	52995	0.12892	0.18947	0.03590	0.96432	-
XGBoost B	65050	0.12901	0.19792	0.03917	0.96017	0.41500%
XGBoost C	182119	0.12585	0.19868	0.03947	0.96036	0.39600%
XGBoost D	135701	0.12389	0.18913	0.03577	0.96404	0.02800%

Através da análise da Tabela 26 é possível verificar-se que o resultado do algortimo XGBoost nos quatro conjuntos de dados é extremamente semelhante nas quatro métricas estudadas. De facto, a diferença entre o melhor resultado da métrica R2 entre o conjunto de dados com pior resultado (B) e com o melhor resultado (A) é apenas de 0.41500%. Ainda assim, os conjuntos A e D são os que possuem uma melhor performace, com um valor de R2 de 0.96432 e 0.96404, respetivamente.

Para além do exposto procedeu-se, ainda, ao teste dos algortimos RF, LightGBM, RL, MLP e CNN para o conjunto de dados D. Tais resultados, bem como os hiper parametros que conduziram à melhor performance dos algortimos, encontram-se detalhados na Tabela 27.

Tabela 27 - Resultados obtidos para os algoritmos (com hiperparâmetros modificados) RF, LightGMB, RL, MLP e CNN para o conjunto de dados D

	MAE	RMSE	MSE	R2	Hiperparâmetros
RF	0.12619	0.19860	0.03944	0.96035	n_estimators=1000 max_depth=0 min_samples_split=5 min_samples_leaf=1
LightGBM	0.13039	0.19804	0.03922	0.96057	learning_rate=0.1 max_depth=7 n_estimators=1000 subsample=0.8
RL	0.30001	0.41869	0.17530	0.82378	alpha=0.01 max_iter=100
MLP	0.14779	0.22473	0.05050	0.94923	epoch=100 optimizer=adam

CNN	0.14747	0.21856	0.04777	0.95192	epoch=500 optimizer=adam filters=64 filters_size=2

À semelhança do que ocorreu para os conjuntos de dados A, B e C, também no conjunto de dados D o algortimo que obteve melhores resultados foi o XGBoost, enquanto que o pior foi o RL.

As Tabelas 28, 29, 30, 31 e 32 apresentam os resultados obtido pelos algortimos RF, LightGBM, RL, MLP e CNN, respetivamente, no conjunto de dados D, bem como a sua comparação com os resultados obtidos no conjunto de dados A, B e C e a percentagem de variação da métrica R2 face ao conjunto com o melhor resultado nesta métrica. Estas tabelas permitem constatar, respetivamente:

- O algoritmo RF apresenta uma melhor performance quando testado com o conjunto de dados D;
- O algoritmo LightGBM apresenta uma performance muito semelhante quando testado com o conjunto de dados A e D sendo, no entanto, melhor em 3 das 4 métricas analisadas melhor com o conjunto de dados A;
- O algoritmo RL apresenta uma melhor performance quando testado com o conjunto de dados A;
- O algoritmo MLP apresenta uma performance muito semelhante quando testado com o conjunto de dados C e D sendo, no entanto, melhor em 3 das 4 métricas analisadas melhor com o conjunto de dados D;
- O algoritmo CNN apresenta uma melhor performance semelhante quando testado com o conjunto de dados C e D sendo, no entanto, melhor em 3 das 4 métricas analisadas com o conjunto de dados D;
- O algoritmo CNN é o que apresenta maior variação entre a performance dos diferentes conjuntos de dados.

Tabela 28 - Resultados do algoritmo RF com hiperparâmetros modificados para cada um dos conjuntos de dados A, B, C e D

	MAE	**RMSE**	**MSE**	**R2**	**% variação**
RF A	0.14407	0.21701	0.04709	0.95319	0.71600%
RF B	0.14095	0.22498	0.05062	0.94854	1.18100%
RF C	0.12778	0.20824	0.04336	0.95646	0.38900%
RF D	0.12619	0.19860	0.03944	0.96035	-

Tabela 29 - Resultados do algoritmo LightGBM com hiperparâmetros modificados para cada um dos conjuntos de dados A, B, C e D

	MAE	RMSE	MSE	R2	% variação
LightGBM A	0.13270	0.19168	0.03674	0.96348	-
LightGBM B	0.13514	0.20249	0.04100	0.95831	0.51700%
LightGBM C	0.13117	0.20726	0.04296	0.95687	0.66100%
LightGBM D	0.13039	0.19804	0.03922	0.96057	0.29100%

Tabela 30 - Resultados do algoritmo RL com hiperparâmetros modificados para cada um dos conjuntos de dados A, B, C e D

	MAE	RMSE	MSE	R2	% variação
RL A	0.29861	0.40594	0.16478	0.83621	-
RL B	0.30771	0.41977	0.17620	0.82084	1.53700%
RL C	0.31915	0.46786	0.21889	0.78024	5.59700%
RL D	0.30001	0.41869	0.17530	0.82378	1.24300%

Tabela 31 - Resultados do algoritmo MLP com hiperparâmetros modificados para cada um dos conjuntos de dados A, B, C e D

	MAE	RMSE	MSE	R2	% variação
MLP A	0.18674	0.27406	0.07511	0.92535	2.38800%
MLP B	0.17752	0.26420	0.06980	0.92903	2.02000%
MLP C	0.14475	0.22834	0.05214	0.94766	0.15700%
MLP D	0.14779	0.22473	0.05050	0.94923	-

Tabela 32 - Resultados do algoritmo CNN com hiperparâmetros modificados para cada um dos conjuntos de dados A, B, C e D

	MAE	RMSE	MSE	R2	% variação
CNN A	0.19355	0.62368	0.38898	0.61395	33.7970%
CNN B	0.16983	0.48837	0.23850	0.75650	19.5420%
CNN C	0.14459	0.22753	0.05177	0.94798	0.39400%
CNN D	0.14747	0.21855	0.04777	0.95192	-

4.5 Comparação e Discussão de Resultados com Literatura

Após uma análise comparativa entre os resultados obtidos nos conjuntos de dados A, B e C e D para os algortimos RF, XGBoost, LightGBM, RL, MLP e CNN com hiperparâmetros padrão e hiperparâmetros modificados pretende-se, neste subcapítulo, apresentar uma comparação entre os resultados mencionados e os obtidos pelos autores estudados em 2.5.1.

Dado o exposto, sumariou-se, na Tabela 33, os resultados de todos os algortimos testados para os conjuntos de dados A, B, C e D, bem como os resultados dos algortimos estudados no subcapítulo 2.5.1.

Tabela 33 - Comparação de resultados obtidos vs. artigos analisados

Conj. Dados vs. Artigos	**RL**	**RF**	**XGBoost**	**LightGBM**	**MLP**	**CNN**
Conj. Dados A (HP Padrão)	MAE: 0.29203 MSE: 0.15661 RMSE: 0.39574 R2: 0.84434	MAE: 0.14527 MSE: 0.04764 RMSE: 0.21827 R2: 0.95264	MAE: 0.14820 MSE: 0.04401 RMSE: 0.20979 R2: 0.95625	MAE: 0.15990 MSE: 0.05079 RMSE: 0.22527 R2: 0.94952	MAE: 0.18573 MSE: 0.06930 RMSE: 0.26325 R2: 0.93112	MAE: 0.18789 MSE: 0.59584 RMSE: 0.77191 R2: 0.40864
Conj. Dados B (HP Padrão)	MAE: 0.30468 MSE: 0.17185 RMSE: 0.41454 R2: 0.82527	MAE: 0.14306 MSE: 0.05172 RMSE: 0.22742 R2: 0.94741	MAE: 0.15267 MSE: 0.05011 RMSE: 0.22385 R2: 0.94905	MAE: 0.16758 MSE: 0.05834 RMSE: 0.24155 R2: 0.94067	MAE: 0.19475 MSE: 0.08369 RMSE: 0.28929 R2: 0.91491	MAE: 0.17887 MSE: 0.67294 RMSE: 0.82033 R2:0.31297
Conj. Dados C (HP Padrão)	MAE: 0.31965 MSE: 0.21796 RMSE: 0.46686 R2: 0.78118	MAE: 0.12900 MSE: 0.04490 RMSE: 0.21190 R2: 0.95492	MAE: 0.14210 MSE: 0.04907 RMSE: 0.22154 R2: 0.95073	MAE: 0.16003 MSE: 0.06011 RMSE: 0.24517 R2: 0.93964	MAE: 0.16082 MSE: 0.06484 RMSE: 0.25463 R2: 0.93491	MAE: 0.14852 MSE: 0.23309 RMSE: 0.05433 R2:0.94540
Conj. Dados A (HP Modific.)	MAE: 0.29861 MSE: 0.16478 RMSE: 0.40594 R2: 0.83621	MAE: 0.14407 MSE: 0.04709 RMSE: 0.21701 R2: 0.95319	MAE: 0.12892 MSE: 0.03590 RMSE: 0.18947 **R2: 0.96432**	MAE: 0.13270 MSE: 0.03674 RMSE: 0.19168 R2: 0.96348	MAE: 0.18674 MSE: 0.07511 RMSE: 0.27406 R2: 0.92535	MAE: 0.19355 MSE: 0.38898 RMSE: 0.62368 R2: 0.61395
Conj. Dados B (HP Modific.)	MAE: 0.30771 MSE: 0.17620 RMSE: 0.41977 R2: 0.82084	MAE: 0.14095 MSE: 0.05062 RMSE: 0.22498 R2: 0.94854	MAE: 0.12901 MSE: 0.03917 RMSE: 0.19792 R2: 0.96017	MAE: 0.13514 MSE: 0.04100 RMSE: 0.20249 R2: 0.95831	MAE: 0.17752 MSE: 0.06980 RMSE: 0.26420 R2: 0.92903	MAE: 0.16983 MSE: 0.23850 RMSE: 0.48837 R2: 0.75650
Conj. Dados C (HP Modific.)	MAE: 0.31915 MSE: 0.21889 RMSE: 0.46786	MAE: 0.12778 MSE: 0.04336 RMSE: 0.20824	MAE: 0.12585 MSE: 0.03947 RMSE: 0.19868	MAE: 0.13117 MSE: 0.04296 RMSE: 0.20726	MAE: 0.14475 MSE: 0.05214 RMSE: 0.22834 R2: 0.94766	MAE: 0.14459 MSE: 0.05177 RMSE: 0.22753 R2: 0.94798

	R2: 0.78024	R2: 0.95646	R2: 0.96036	R2: 0.95687		
Conj. Dados D (HP Modific.)	MAE: 0.30001 MSE: 0.17530 RMSE: 0.41869 R2: 0.82378	MAE: 0.12619 MSE: 0.03944 RMSE: 0.19860 R2: 0.96035	**MAE: 0.12389 MSE: 0.03577 RMSE: 0.18913** R2: 0.96404	MAE: 0.13039 MSE: 0.03922 RMSE: 0.19804 R2: 0.96057	MAE: 0.14779 MSE: 0.05050 RMSE: 0.22473 R2: 0.94923	MAE: 0.14747 MSE: 0.04777 RMSE: 0.21856 R2: 0.95192
(Narayana et al., 2021)	MAE: 0.37998 MSE: 0.36805 RMSE: 0.60667	MAE: 0.19780 MSE: 0.10122 RMSE: 0.31816				
(C. Chen et al., 2017) Model 2	NMSE: 0.26000	NMSE: 0.05200				
(Longani et al., 2021)		MAE: 1.13000 MSE: 11.89000 RMSE: 3.44000	MAE: 0.17000 MSE: 0.28000 RMSE: 0.53000			
(Gupta et al., 2022)	MAE: 1100.7714 RMSE: 1442.1927 R2: 0.9652	MAE: 644.8759 RMSE: 1131.8468 R2: 0.9785				
(Jin, 2021)	R2: 0.72354	R2: 0.90416				
(Hankar et al., 2022)	RMSE: 63933.52 R2: 0.57000	RMSE: 44939.79 R2: 0.74000	RMSE: 44516.20 R2: 0.80			
(H. Zhang, 2022)	MAE: 2.5552 MAPE: 0.1652 RMSE: 1.9126	MAE: 1.5769 MAPE: 0.1669 RMSE: 1.8982	MAE: 1.5247 MAPE: 0.1602 RMSE: 1.8103	MAE: 1.4903 MAPE: 0.1591 RMSE: 1.7739		
(Narayana et al., 2022)	MAE: 9.5351 MSE: 3.1199 RMSE: 6.4762	MAE: 4.1918 MSE: 0.4069 RMSE: 2.4082				

Através da análise da tabela acima é possível verificar-se que:

- Relativamente à métrica MAE:

- Apenas 5 (Gupta et al., 2022; Longani et al., 2021; Narayana et al., 2021, 2022; H. Zhang, 2022) dos 9 artigos estudados apresentam esta métrica;
- O melhor resultado desta métrica é alcançado pelo algoritmo XGBoost no conjunto de dados D;

- Relativamente à métrica MSE e RMSE:
  - 6 (Gupta et al., 2022; Hankar et al., 2022; Longani et al., 2021; Narayana et al., 2021, 2022; H. Zhang, 2022) dos 9 artigos estudados apresentam esta métrica;
  - Dos 6 artigos mencionados, apenas 4 (Longani et al., 2021; Narayana et al., 2021, 2022; H. Zhang, 2022) normalizaram os seus dados;
  - Dentro dos 4 artigos cujos dados se encontram normalizados, o melhor resultado destas métricas é alcançado pelo algoritmo XGBoost no conjunto de dados D;

- Relativamente à métrica R2:
  - Apenas 3 (Gupta et al., 2022; Hankar et al., 2022; Jin, 2021) dos 9 artigos estudados apresentam esta métrica;
  - O melhor resultado desta métrica é alcançado pelo algoritmo RF do artigo elaborado por (Gupta et al., 2022). No entanto, uma vez que este artigo apenas usou 205 carros, considerou-se o algoritmo XGBoost do conjunto de dados A como o que foi melhor sucedido;

- Relativamente à métrica NMSE:
  - Apenas 1 (C. Chen et al., 2017; Gupta et al., 2022) dos 9 artigos estudados apresenta esta métrica;
  - Apenas se extraiu esta métrica para os algoritmos XGBoost do conjunto A (0.03568) e D (0.03596) e LightGBM do conjunto A (0.03652) e D (0.03943);
  - Comparando os resultados obtidos, o melhor resultado desta métrica é alcançado pelo algoritmo LightGBM no conjunto de dados A;

- O artigo Arora et al., 2022, ao apresentar apenas os resultados obtidos para a métrica *accuracy*, não foi comparado com nenhum dos restantes artigos nem com os modelos desenvolvidos;

- O algoritmo RL obteve a sua melhor performance quando testado com o conjunto de dados A (sublinhado a azul), com hiperparâmetros padrão;

- Os algoritmos RF, MLP e CNN obtiveram a sua melhor performance quando testado com o conjunto de dados D (sublinhado a roxo, cor-de-laranja e cor-de-rosa, respetivamente), com hiperparâmetros modificados;

- Os algoritmos XGboost e LightGBM obtiveram a sua melhor performance quando testado com os conjuntos de dados A e D (sublinhado a verde a amarelo, repetivamente), com hiperparâmetros modificados.

Nenhum dos artigos estudados obteve tão bons resultados quanto os desenvolvidos no âmbito da presente dissertação.

Os resultados obtidos podem justificar-se, entre outros motivos, pelo número de características e o número de carros que compõe o conjunto de dados usados pelos artigos estudados e pelos utilizados no desenvolvimento dos modelos criados. O número de características e o número de carros do conjunto de dados usados em cada um dos artigos analisados, bem como dos usados no contexto da presente dissertação encontram-se detalhados na Tabela 34.

Tabela 34 – Número de características e carros usados nos modelos analisados e desenvolvidos

	N.º Características	**N.º Carros**
(Narayana et al., 2021)	8	>4000
(C. Chen et al., 2017)	19	102959
(Longani et al., 2021)	11	2425
(Gupta et al., 2022)	26	205
(Jin, 2021)	6	100000
(Hankar et al., 2022)	6	8000
(H. Zhang, 2022)	16	-
(Arora et al., 2022)	8	∼300
(Narayana et al., 2022)	10	>4250
Conjunto A	50	57038
Conjunto B	30	70253
Conjunto C	10	192799
Conjunto D	13	144702

Como se pode analisar pela Tabela 34, os conjuntos de dados A e B são os que apresentam um maior número de características, 50 e 30, respetivamente. Por outro lado, os conjuntos de dados C e D, com 10 e 13 características respetivamente, são os que possuem mais carros, 192799 e 144702 carros, respetivamente. Excluindo os conjuntos de dados A e B, o único artigo que apresenta um maior número de características é o Gupta et al., 2022. Contudo, este apenas contém um conjunto de dados de 205 carros, enquanto que os conjuntos de dados A e B contém 57038 e 70253 carros, respetivamente. Por outro lado, os artigos que possuem, com exceção dos conjuntos de dados C e D, um maior conjunto de dados são C. Chen et al., 2017 e Jin, 2021 , com 102959 e 100000 carros, respetivamente. Contudo, estes artigos apenas contêm 19 e 6 características, respetivamente.

Outros fatores que condicionam a performance dos algoritmos é o tratamento de dados ao qual os mesmos foram sujeitos, bem como os hiperparâmetros usados aquando do teste dos algoritmos.

Uma vez que num contexto real é mais provável que um vendedor/comprador possua mais facilmente as características do conjunto D do que do conjunto A, optou-se por testar em contexto real o modelo criado com o algortimo XGBoost, introduzindo as características do conjunto de dados D. Desta forma, a Tabela 35 apresenta o preço obtido para 15 carros constantes do conjunto de testes, aquando da introdução das características dos mesmos no modelo desenvolvido.

A análise da referida tabela permite verificar que, de facto, o modelo desenvolvido cumpre o propósito para o qual foi feito, dada a proximidade dos valores faculdados e obitdos aquando do teste do algortimo em causa.

Assim, pode afirmar-se que o objetivo inicial foi cumprido, podendo o mesmo contribui para a negociação de carros usados a preços mais justos e razoáveis, reduzindo os riscos associados às transações neste setor.

Tabela 35 – Comparação do preço original e do preço obtido aquando do teste do algoritmo XGBoost com dados treino do conjunto de dados D

Tipo De Carro	Cilindrada	Potência	Ano	Combustível	Tipo De Caixa	Marca	Quilometragem	Modelo	Detalhe Da Versão	Versão	Capacidade	Preço Original	Preço
carrinha	1560.0	100	2018	gasóleo	manual	Peugeot	57000	308_sw	1.6 BlueHDi Active	ver_1_6_bluehdi_active	5.0	17950	17906
carrinha	1461.0	90	2018	gasóleo	manual	Renault	111956	clio_sport_tourer	1.5 dCi GT Line	ver_1_5_dci_gt_line	5.0	15900	15602
sedan	2143.0	150	2017	gasóleo	automática	alfa_romeo	66000	giulia	2.2 D Super AT8	ver_2_2_d_super_at8	5.0	26900	27002
sedan	2995.0	218	1994	gasolina	manual	BMW	183000	730	i V8	ver_i_v8	5.0	9999	7713
carrinha	1968.0	190	2016	gasóleo	automática	Audi	155000	a6_avant	2.0 TDi Business Line S_line S tronic	ver_2_0_tdi_business_line_s__line_s_tronic	5.0	26500	28285
compacto	1422.0	90	2017	gasóleo	manual	VW	112280	polo	1.4 TDi Connect	ver_1_4_tdi_connect	5.0	13140	14359
monovolume	1461.0	110	2014	gasóleo	manual	Renault	212000	grand_scenic	1.5 dCi Bose Edition SS	ver_1_5_dci_bose_edition_ss	5.0	10499	10997
carrinha	1598.0	136	2015	gasóleo	manual	Mercedes Benz	140000	c_200	BlueTEC Avantgarde	ver_bluetec_avantgarde	5.0	20000	22051
sedan	1968.0	150	2017	gasóleo	automática	Audi	143000	a5_sportback	2.0 TDi S_line S tronic	ver_2_0_tdi_s__line_s_tronic	4.0	30900	33194
coupé	1995.0	204	2008	gasóleo	manual	BMW	190000	123	d	ver_d	4.0	15950	15733
citadino	998.0	95	2020	gasolina	manual	Ford	1568	Fiesta	1.0 EcoBoost ST_Line	ver_1_0_ecoboost_st__line	5.0	21990	21698
citadino	1499.0	102	2019	gasóleo	manual	Peugeot	22000	208	1.5 BlueHDi Signature	ver_1_5_bluehdi_signature	5.0	17950	17673
carrinha	1598.0	120	2017	gasóleo	automática	VW	110667	passat_variant	1.6 TDI Confortline DSG	ver_1_6_tdi_confortline_dsg	5.0	21980	22131
monovolume	1461.0	110	2017	gasóleo	manual	Renault	97000	scenic	1.5 dCi GT Line	ver_1_5_dci_gt_line	5.0	18500	19045
compacto	1560.0	100	2018	gasóleo	manual	Peugeot	57000	308	1.6 BlueHDi Active	ver_1_6_bluehdi_active	5.0	17950	17895

4.6 Conclusão

Após a análise e tratamento de dados foram testados os algoritmos RF, XGBoost, LightGBM, RL, MLP e CNN nos quatro conjuntos de dados A, B, C e D. O conjunto de dados A possui 50 características e 52995 carros, o conjunto de dados B possui 30 características e 65050 automóveis, o conjunto de dados C possui 10 características e 182119 veículos. O conjunto de dados D é constituído por 13 características, as quais, para além do preço, foram obtidas através do método *Shapely Additive exPlanations* (SHAP), aplicado ao algoritmo XGBoost do conjunto de dados A possui 135701 carros.

Os algoritmos aplicados aos conjuntos de dados A, B e C foram testados duas vezes, com hiperparâmetros padrão e hiperparâmetros modificados. Todos os algoritmos dos quatro conjuntos de dados foram sujeitos a uma metodologia de 80% treino de e 20% testes e avaliados, maioritariamente, através das métricas R2, MSE, RMSE e MAE.

Os algoritmos testados com os conjuntos de dados A, B e C obtiveram melhores resultados aquando da alteração de hiperparâmetros padrão, com a exceção do algoritmo MLP no conjunto de dados A e o algoritmo RL nos conjuntos A, B, C e D.

Dentro dos algoritmos testados, os algoritmos XGboost e LightGBM foram os que apresentaram melhores resultados, tendo os mesmos sido muito idênticos entre si nos 4 conjuntos de dados. Entre os dois algoritmos, o XGBoost foi o que apresentou melhores resultados.

Por fim, o algoritmo XGBoost do conjunto de dados A e D foi o que apresentou melhores resultados entre os algoritmos testados, bem como quando comparados com os algoritmos estudados aquando da revisão do estado da arte, tendo apresentado um valor de R2 de 0.96432 e 0.96404, respetivamente.

O número de características, o tamanho do conjunto de dados usado, o tratamento do mesmo e os hiperparâmetros utilizados são alguns dos principais fatores que influenciam a performance dos algoritmos estudados.

5 Conclusões Finais

Recorde-se as principais motivações, finalidades e etapas do estudo realizado.

O avanço da Inteligência Artificial tem fomentado o lançamento de automóveis com especificações cada vez mais inovadoras e, consequentemente, a preços mais elevados.

Tal aumento de preços conduz a uma maior procura na compra/venda de carros usados. Esta procura leva, muitas vezes, à atribuição de preços irrealistas aos mesmos, aumentando o número de fraudes neste setor, e a uma elevada discrepância nos preços praticados.

De facto, um dos maiores problemas associados a este mercado prende-se com a dificuldade em perceber qual o preço justo a pagar/vender por um determinado veículo, devido à disparidade de preços praticados dentro da mesma gama ou até dentro do mesmo modelo de carro.

Neste âmbito, a área de *Machine Learning* pode ter um papel preponderante, nomeadamente na elaboração de modelos de previsão de preços de carros usados. Assim, o objetivo do presente trabalho prendeu-se com a análise dos modelos já desenvolvidos neste contexto, do grau de precisão dos mesmos e com a criação de um modelo que colmatasse as falhas nos já existentes, de forma a se aumentar o referido grau de precisão.

Desta forma, a concretização do objetivo mencionado pretende, assim, contribuir para a negociação de carros usados a preços mais justos e razoáveis, reduzindo os riscos associados às transações neste setor.

De modo a orientar a investigação realizada, foram elaboradas quatro questões, as quais, consequentemente, conduziram à criação de cinco objetivos principais:

- **Q1** – Como se encontra atualmente o estado de arte no que respeita a sistemas de previsão, particularmente no âmbito da previsão de preços para carros usados?

- **Q2** – Poder-se-á considerar um sistema de previsão de preço de carros usados preciso, utilizando apenas um certo número de características e um conjunto de dados relativos a transações anteriores?

- **Q3** – Será o resultado de um modelo de previsão diferente consoante algoritmos utilizados?

- **Q4** – É um algoritmo mais preciso se fizer uso da combinação de várias técnicas?

Assim, os objetivos identificados com base nas questões acima formuladas são os seguintes:

- **O1** – Investigar o atual estado de arte no âmbito de modelos de previsão e o seu enquadramento no setor automóvel;

- **O2** – Detetar as características que impactam a criação de um modelo de previsão de preço de carros usados;

- **O3** – Criar um *dataset* com a informação necessária para aplicação num modelo de previsão de preço de carros usados;

- **O4** – Comparar e avaliar a performance dos diferentes modelos e técnicas criadas;

- **O5** – Desenvolver um modelo para previsão de preços de carros usados, com base nos resultados obtidos das comparações efetuadas.

Tabela 36 – Questões de investigação e objetivos correspondentes

Questões	**Objetivos**
Q1	O1
Q2	O2; O3; O4; O5
Q3	O4; O5
Q4	O4; O5

De seguida, apresenta-se um resumo do trabalho realizado que visou cumprir com os objetivos supracitados, bem como as suas conclusões e as respostas às perguntas que os geraram.

Relativamente à **Q1**, é possível encontrar-se vários artigos na literatura acerca de sistemas de previsão de preços. Tais sistemas de previsão têm por base um conjunto de dados, com diversas características, os quais dependem da área de estudo em questão. Com estes dados e características é possível, através de algoritmos de ML, tentar determinar-se o preço mais justo e real para um determinado produto, bem ou objeto.

Alguns dos contextos para os quais se têm desenvolvido sistemas de previsão de preços prendem-se, por exemplo, com a compra e venda de ouro (Farahani & Mehralian, 2013), moedas digitais (McNally et al., 2018), ações (K. V. Kumar & Anitha, 2022) e de casas (Banerjee & Dutta, 2018). Alguns dos algoritmos utilizados no desenvolvimento destes sistemas são o RF, RNA, SVM, KNN, NB, LSTM.

É também possível encontrar-se vários artigos na literatura acerca de sistemas de previsão de preços de automóveis usados, nomeadamente Arora et al., 2022; C. Chen et al., 2017; Gupta et al., 2022; Hankar et al., 2022; Jin, 2021; Longani et al., 2021; Narayana et al., 2021, 2022; H. Zhang, 2022, tendo os mesmos sido os artigos mais estudados no contexto deste estudo. Considerando todos os artigos estudados, os principais algoritmos estudados foram RL, AD, RF, XGBoost, LightGBM, SVM e KNN.

Através da análise dos mesmos, pôde concluir-se que:

- O algoritmo RL foi, na grande maior parte dos algoritmos estudados, o que apresenta piores resultados;

- O algoritmo RF foi considerado como o melhor, sempre que os algoritmos XGBoost e LightGBM não foram testados, com exceção do artigo Gupta et al., 2022 onde o algoritmo AD foi o melhor;

- No caso em que foram testados, simultaneamente, os algoritmos RF e XGBoost, mas em que não foi testado o algoritmo LightGBM, o algoritmo que apresentou melhor performance foi o XGBoost;

- No caso em que foram testados, simultaneamente, os algoritmos XGBoost e LightGBM, o algoritmo que apresentou melhor performance foi o LightGBM;

- A maior parte do artigos foram avaliados através de métricas como R2, MSE, RMSE ou MAE.

Note-se que todos os artigos estudados utilizaram conjuntos de dados pertencentes ao mercado estrangeiro. De facto, não foram encontrados quaisquer artigos neste âmbito com conjuntos de dados nacionais.

Dado o exposto, pode concluir-se que o **O1** foi alcançado com sucesso.

O **O2** foi, nesta fase, parcialmente concretizado aquando da análise da sugestão das principais melhorias de investigação futura, as quais incluíram o teste de algoritmos com um maior dataset maior ou com um maior número de características ou o estudo de um maior número de algoritmos.

Dado o exposto, e de forma a concretizar o **O3** e a colmatar tal o supracitado, procedeu-se à aproximação a várias empresas do ramo automóvel que pudessem fornecer dados para o desenvolvimento do modelo pretendido. Neste sentido, a empresa Standvirtual prontificou-se a fornecer os dados necessários ao desenvolvimento de um modelo de previsão de preço de carros usados aplicados ao contexto do mercado nacional, tendo facultado dados de 198107 carros com 230 características, os quais foram organizados numa tabela.

Os dados facultados foram divididos em três conjuntos A, B e C, os quais continham 50, 30 e 10 características, respetivamente. Posteriormente, procedeu-se à eliminação de carros cujas

descrições das características estavam ausentes ou corrompidas em cada um dos conjuntos de dados A, B e C, os quais ficaram reduzidos a 57038, 70253 e 192799 carros, respetivamente.

De seguida, procedeu-se à conversão em linguagem numérica da descrição das características previamente apresentadas em linguagem alfabética, à remoção de *outliers e* à normalização dos dados de determinadas características. Findo o processo de tratamento de dados, o conjunto de dados A passou a ter um total de 52995 carros, conjunto de dados B 65050 carros e o conjunto de dados C 182119 carros.

Para cada um dos conjuntos de dados A, B e C procedeu-se ao teste dos algoritmos RF, XGBoost, LightGBM e RL, uma vez que, dos artigos estudados, os primeiros três foram os que obtiveram melhores resultados e o RL o mais comumente usado. Para além destes, procedeu-se ainda aos testes das redes neuronais MLP e CNN, de forma a se incluir técnicas de *deep learning*.

De forma a dar resposta ao **O4**, todos os algoritmos foram testados através de técnicas de treino, validação e teste, com hiperparâmetros padrão/comumente usados e introduzidos manualmente antes da fase de treino, tendo sido atribuídos 80% do conjunto inicial de dados à fase de treino e 20% à fase de teste.

Os algoritmos testados obtiveram melhores resultados aquando da alteração de hiperparâmetros padrão, com a exceção do algoritmo MLP no conjunto de dados A e o algoritmo RL nos conjuntos A, B, C e D. Assim, o valor atribuído aos hiperparâmetros é também uma das características que impacta o sucesso de um modelo, complementado o **O2**.

Dentro dos algoritmos testados, os algoritmos XGBoost e LightGBM com hiperparâmetros modificados foram os que apresentaram melhores resultados, tendo os mesmos sido muito idênticos entre si nos 4 conjuntos de dados. Ainda assim, e entre os dois algoritmos, o XGBoost do conjunto A foi o que apresentou melhores resultados, com o valor de R2 de 0.96432. Por outro lado, o algoritmo RL foi o que apresentou piores resultados no três conjuntos de dados. Os resultado obtidos permitem responder à **Q3**, afirmando que os resultados obtidos diferem, efetivamente, consoante os algoritmos testados.

Uma vez que o algoritmo XGBoost aplicado ao conjunto de dados A foi que apresentou melhores resultados, decidiu-se elaborar um quarto conjunto de dados, D, constituído por 13 características, as quais, para além do preço, foram obtidas através do método *Shapely Additive exPlanations* (SHAP), aplicado ao algoritmo XGBoost do conjunto de dados A. Também no conjunto de dados D se aplicou o tratamento de dados previamente usado nos conjuntos de dados A, B e C. Após o mesmo, o conjunto de dados D ficou composto por 135701 carros.

Posteriormente, testaram-se novamente os mesmo algoritmos para o novo conjunto de dados D, com os mesmos hiperparâmetros (modificados) usados aquando do teste nos conjuntos de dados A, B e C. À semelhança do que ocorreu para os conjuntos de dados A, B e C, também no conjunto de dados D o algortimo que obteve melhores resultados foi o XGBoost, enquanto que o pior foi o RL.

Os hiperparâmetros que conduziram à melhor performance do algortimo XGBoost nos quatro conjuntos de dados A, B, C e D foram *nr_estimator* de 1000, *max_depth* de 7, *learning_rate* de 0.1, *reg_alpha* de 0.0 e *reg_lambda* de 1.0. No caso dos conjuntos de dados A e B, o valor de *subsample* que conduziu às melhor performance foi de 1.0, enquanto que no conjunto de dados C e D foi de 0.8.

O algoritmo XGBoost dos conjuntos de dados A e D com os hiperparâmetros acima descritos foi o que apresentou melhores resultados entre os algoritmos testados, bem como quando comparados com os algoritmos estudados aquando da revisão do estado da arte, tendo apresentado um valor de R2 de 0.96432 e 0.96404, respetivamente. De facto, nenhum dos algoritmos estudados apresentou resultados tão promissores quanto os alcançados através do algoritmo XGBoost aplicado aos conjuntos de dados A e D. Assim, o tratamento de dados aplicado, bem como os hiperparâmetros utilizados conduziram a melhores resultados, comprovando que a combinação de diversas técnicas conduz a uma maior precisão dos modelos desenvolvidos, respondendo positivamente à **Q4**.

Os resultados obtidos podem justificar-se, entre outros motivos, pelo número de características e o número de carros que compõe o conjunto de dados usados pelos artigos estudados e pelos utilizados no desenvolvimento dos modelos criados. Como tal, e respondendo à **Q2**, quanto maior o número de dados e de características analisadas, melhor será o desempenho de um modelo, sendo que um baixo número destes fatores podem invalidar os resultados obtidos.

Por último, pode afirmar-se que o modelo desenvolvido cumpre o propósito para o qual foi feito e que o objetivo da presente dissertação foi concluído com sucesso, cumprindo-se, assim, o último obetivo, **O5.**

Assim, pode afirmar-se que o objetivo global do presente trabalho foi cumprido, podendo o mesmo contribui para a negociação de carros usados a preços mais justos e razoáveis, reduzindo os riscos associados às transações neste setor.

6 Referências

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, *22*(5), 717–727. https://doi.org/10.1016/S0731-7085(99)00272-1

Anusuya, M. A., & Katti, S. K. (2009). Speech Recognition by Machine: A Review. *IJCSIS) International Journal of Computer Science and Information Security*, *6*(3). http://sites.google.com/site/ijcsis/

Arora, P., Gupta, H., & Singh, A. (2022). Forecasting resale value of the car: Evaluating the proficiency under the impact of machine learning model. *Materials Today: Proceedings*, *69*, 441–445. https://doi.org/10.1016/J.MATPR.2022.09.074

Asada, M. (2003). Robotics. *Encyclopedia of Information Systems*, 707–722. https://doi.org/10.1016/B0-12-227240-4/00150-7

Banerjee, D., & Dutta, S. (2018). Predicting the housing price direction using machine learning techniques. *IEEE International Conference on Power, Control, Signals and Instrumentation Engineering, ICPCSI 2017*, 2998–3000. https://doi.org/10.1109/ICPCSI.2017.8392275

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197–227. https://doi.org/10.1007/S11749-016-0481-7/FIGURES/4

Caruana, R. (2006). *An Empirical Comparison of Supervised Learning Algorithms*. www.cs.cornell.edu

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/GMD-7-1247-2014

Chaudhary, N. L., & Lee, W. J. (2016). Detecting and Removing Outliers in Production Data to Enhance Production Forecasting. *SPE Hydrocarbon Economics and Evaluation Symposium*, *2016-January*. https://doi.org/10.2118/179958-MS

Chen, C., Hao, L., & Xu, C. (2017). Comparative analysis of used car price evaluation models. *AIP Conference Proceedings*, *1839*. https://doi.org/10.1063/1.4982530

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, 1–24. https://doi.org/10.7717/PEERJ-CS.623/

DiPietro, R., & Hager, G. D. (2019). Deep learning: RNNs and LSTM. *Handbook of Medical Image Computing and Computer Assisted Intervention*, 503–519. https://doi.org/10.1016/B978-0-12-816176-0.00026-0

Dorri, A., Kanhere, S. S., & Jurdak, R. (2018). Multi-Agent Systems: A Survey. *IEEE Access*, *6*, 28573–28593. https://doi.org/10.1109/ACCESS.2018.2831228

el Naqa, I., & Murphy, M. J. (2015). What Is Machine Learning? *Machine Learning in Radiation Oncology*, 3–11. https://doi.org/10.1007/978-3-319-18305-3_1

Eye, A. von., & Clogg, C. C. (1996). *Categorical variables in developmental research : methods of analysis*. Academic Press. http://www.sciencedirect.com:5070/book/9780127249650/categorical-variables-in-developmental-research

Farahani, M. K., & Mehralian, S. (2013). Comparison between Artificial Neural Network and neuro-fuzzy for gold price prediction. *13th Iranian Conference on Fuzzy Systems, IFSC 2013*. https://doi.org/10.1109/IFSC.2013.6675635

Gandhi, M. K., Chaudhari, C., & Ghosh, K. (2022). To study the challenges faced in application of artificial intelligence in automobile industry. *AIP Conference Proceedings*, *2519*. https://doi.org/10.1063/5.0111115

Ganesh, M. (2019). Used Cars Price Prediction using Supervised Learning Techniques Article in International Journal of Engineering and Advanced Technology · December. *International Journal of Engineering and Advanced Technology (IJEAT)*, *9*, 2249–8958. https://doi.org/10.35940/ijeat.A1042.1291S319

Gegic, E., Isakovic, B., Keco, D., Masetic, Z., & Kevric, J. (2019). Car Price Prediction using Machine Learning Techniques. *TEM Journal*, *8*(1), 113–118. https://doi.org/10.18421/TEM81-16

Goldberg, X. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *6*, 1–116. https://doi.org/10.2200/S00196ED1V01Y200906AIM006

Google. (2022). *What is Clustering? | Machine Learning | Google Developers*. What Is Clustering? https://developers.google.com/machine-learning/clustering/overview

Gupta, R., Sharma, A., Anand, V., & Gupta, S. (2022). Automobile Price Prediction using Regression Models. *5th International Conference on Inventive Computation Technologies, ICICT 2022 - Proceedings*, 410–416. https://doi.org/10.1109/ICICT54344.2022.9850657

Hankar, M., Birjali, M., & Beni-Hssane, A. (2022). Used Car Price Prediction using Machine Learning: A Case Study. *11th International Symposium on Signal, Image, Video and Communications, ISIVC 2022 - Conference Proceedings*. https://doi.org/10.1109/ISIVC54825.2022.9800719

Hopke, P. K. (2003). The evolution of chemometrics. *Analytica Chimica Acta*, *500*(1–2), 365–377. https://doi.org/10.1016/S0003-2670(03)00944-9

Huang, X., Wu, L., & Ye, Y. (2019). A Review on Dimensionality Reduction Techniques. *Https://Doi.Org/10.1142/S0218001419500174*, *33*(10). https://doi.org/10.1142/S0218001419500174

Ibañez, A. (2019, May 29). *Semi-Supervised Learning… the great unknown - Think Big*. Telefónica. https://business.blogthinkbig.com/semi-supervised-learning-the-great-unknown/

Idris, N. O., Achban, A., Utiarahman, S. A., Karim, J., & Pontoiyo, F. (2020). Predicting the selling price of cars using business intelligence with the feed-forward backpropagation algorithms. *2020 5th International Conference on Informatics and Computing, ICIC 2020*. https://doi.org/10.1109/ICIC50835.2020.9288594

Jin, C. (2021). Price Prediction of Used Cars Using Machine Learning. *Proceedings of 2021 IEEE International Conference on Emergency Science and Information Technology, ICESIT 2021*, 223–230. https://doi.org/10.1109/ICESIT53460.2021.9696839

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, *4*, 237–285. https://doi.org/10.1613/JAIR.301

Károly, A. I., Fullér, R., & Galambos, P. (2018). Unsupervised Clustering for Deep Learning: A tutorial survey. *Acta Polytechnica Hungarica*, *15*(8).

Kasturi, S. N. (2019). *XGBOOST vs LightGBM: Which algorithm wins the race !!! | by Sai Nikhilesh Kasturi | Towards Data Science*. Towards Data Science. https://towardsdatascience.com/lightgbm-vs-xgboost-which-algorithm-win-the-race-1ff7dd4917d

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, *30*. https://github.com/Microsoft/LightGBM.

Kim, C. (2022). Deep Q-Learning Network with Bayesian-Based Supervised Expert Learning. *Symmetry*, *14*(10). https://doi.org/10.3390/SYM14102134

Kumar, I., Dogra, K., Utreja, C., & Yadav, P. (2018). A Comparative Study of Supervised Machine Learning Algorithms for Stock Market Trend Prediction. *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018*, 1003–1007. https://doi.org/10.1109/ICICCT.2018.8473214

Kumar, K. V., & Anitha, R. (2022). A Detailed Survey to Forecast the Stock Prices by Applying Machine Learning Predictive Models and Artificial Intelligence Techniques. *Proceedings of International Conference on Computing, Communication, Security and Intelligent Systems, IC3SIS 2022*. https://doi.org/10.1109/IC3SIS54991.2022.9885309

Lent, B., Swami, A., & Widom, J. (1997). Clustering association rules. *Proceedings - International Conference on Data Engineering*, 220–231. https://doi.org/10.1109/ICDE.1997.581756

Liao, S. H. (2005). Expert system methodologies and applications-a decade review from 1995 to 2004. *Expert Systems with Applications*, *28*(1), 93–103. https://doi.org/10.1016/J.ESWA.2004.08.003

Longani, C., Potharaju, S. P., & Deore, S. (2021). Price prediction for pre-owned cars using ensemble machine learning techniques. *Advances in Parallel Computing*, *39*, 178–187. https://doi.org/10.3233/APC210194

Mahesh, B. (2018). Machine Learning Algorithms-A Review Self Flowing Generator View project Machine Learning Algorithms-A Review View project Batta Mahesh Independent Researcher Machine Learning Algorithms-A Review. *International Journal of Science and Research*. https://doi.org/10.21275/ART20203995

Mahesh, B. (2019, January). *Machine Learning Algorithms - A Review | Enhanced Reader*. Independent Researcher.

Maulud, D., & Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, *1*(4), 140–147. https://doi.org/10.38094/JASTT1457

Mccarthy, J. (2007). *WHAT IS ARTIFICIAL INTELLIGENCE?* http://www-formal.stanford.edu/jmc/

McNally, S., Roche, J., & Caton, S. (2018). Predicting the Price of Bitcoin Using Machine Learning. *Proceedings - 26th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, PDP 2018*, 339–343. https://doi.org/10.1109/PDP2018.2018.00060

Metz, L., Brain, G., Poole, B., Pfau, D., Deepmind, G., & Sohl-Dickstein, J. (2016). Unrolled Generative Adversarial Networks. *ICLR Conference Papers*. https://arxiv.org/abs/1611.02163v4

Narayana, C. V., Likhitha, C. L., Bademiya, S., & Kusumanjali, K. (2021). Machine Learning Techniques to Predict the Price of Used Cars: Predictive Analytics in Retail Business. *Proceedings of the 2nd International Conference on Electronics and Sustainable Communication Systems, ICESC 2021*, 1680–1687. https://doi.org/10.1109/ICESC51422.2021.9532845

Narayana, C. V., Madhuri, N. O. G., Nagasindhu, A., Aksha, M., & Naveen, C. (2022). Second Sale Car Price Prediction using Machine Learning Algorithm. *7th International Conference on Communication and Electronics Systems, ICCES 2022 - Proceedings*, 1171–1177. https://doi.org/10.1109/ICCES54183.2022.9835872

Navada, A., Ansari, A. N., Patil, S., & Sonkamble, B. A. (2011). Overview of use of decision tree algorithms in machine learning. *Proceedings - 2011 IEEE Control and System Graduate Research Colloquium, ICSGRC 2011*, 37–42. https://doi.org/10.1109/ICSGRC.2011.5991826

Nguyen, G., and, M. W.-I. Worksh. on A.-V. C., & 2005, undefined. (2005). Similarity based visualization of image collections. *Citeseer*. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=badd6b8ccac299b5517d31faea9d3192d75a833d

Nkikabahizi, C., Cheruiyot, W., & Kibe, A. (2022). Chaining Zscore and feature scaling methods to improve neural networks for classification. *Applied Soft Computing*, *123*, 108908. https://doi.org/10.1016/J.ASOC.2022.108908

Nunes, D. F. (2021, February 14). *Carros usados ganham peso em ano de pandemia*. Diário de Notícias. https://www.dn.pt/dinheiro/carros-usados-ganham-peso-em-ano-de-pandemia-13350628.html

O'Shea, K., & Nash, R. (2015). An Introduction to Convolutional Neural Networks. *International Journal for Research in Applied Science and Engineering Technology*, *10*(12), 943–947. https://doi.org/10.22214/ijraset.2022.47789

Ostertagová, E. (2012). Modelling using Polynomial Regression. *Procedia Engineering*, *48*, 500–506. https://doi.org/10.1016/J.PROENG.2012.09.545

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *The BMJ*, *372*. https://doi.org/10.1136/BMJ.N71

Pal, N., Arora, P., Kohli, P., Sundararaman, D., & Palakurthy, S. S. (2019). How Much is my car worth? A methodology for predicting used cars' prices using random forest. *Advances in Intelligent Systems and Computing*, *886*, 413–422. https://doi.org/10.1007/978-3-030-03402-3_28/COVER

Poola, I. (2017). How Artificial Intelligence in Impacting Real life Everyday. *International Journal for Advance Research and Development*, *2*(10), 96–100. https://doi.org/xx.xxx/ijariit-v2i10-1170

Potdar, K., S., T., & D., C. (2017). A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. *International Journal of Computer Applications*, *175*(4), 7–9. https://doi.org/10.5120/IJCA2017915495

Prakash, V. J., & Nithya, L. M. (2014). A Survey On Semi-Supervised Learning Techniques. *International Journal of Computer Trends and Technology*, *8*(1). www.internationaljournalssrg.org

Proteção de Dados Pessoais | UCP. (2016). https://www.ucp.pt/pt-pt/catolicainstitucional/protecao-de-dados-pessoais

Pudaruth, S. (2014). Predicting the Price of Used Cars using Machine Learning Techniques. *International Journal of Information & Computation Technology*, *4*(7), 753–764. http://www.irphouse.com

Ranjan, N., Mundada, K., Phaltane, K., & Ahmad, S. (2016). A Survey on Techniques in NLP. *International Journal of Computer Applications*, *134*(8), 975–8887.

Ray, S. (2019). A Quick Review of Machine Learning Algorithms. *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019*, 35–39. https://doi.org/10.1109/COMITCON.2019.8862451

Rish, I. (2005). *An empirical study of the naive Bayes classifier*.

Samruddhi, K., & Ashok Kumar, R. (2020). Used Car Price Prediction using K-Nearest Neighbor Based Model. *International Journal of Innovative Research in Applied Sciences and Engineering*, *4*(3), 686–689. https://doi.org/10.29027/IJIRASE.V4.I3.2020.686-689

Saravanan, R., & Sujatha, P. (2019). A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification. *Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems, ICICCS 2018*, 945–949. https://doi.org/10.1109/ICCONS.2018.8663155

Shubhendu, S., & Vijay, J. (2013). *Applicability of Artificial Intelligence in Different Fields of Life*. *1*(1), 2347–3878. www.ijser.in

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Sorzano, C. O. S., Vargas, J., & Pascual-Montano, A. (2014). *A survey of dimensionality reduction techniques*.

Su, M., Liang, B., Ma, S., Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview You may also like Classification of Rock Mineral in Field X based on Spectral Data (SWIR & TIR) using Supervised Machine Learning Methods S A Pane and F M H Sihombing-Automatic Machine Learning Method for Hyper-parameter Search Machine Learning from Theory to Algorithms: An Overview. *J. Phys*, 12012. https://doi.org/10.1088/1742-6596/1142/1/012012

Suhaib Kamran, S., Haleem, A., Bahl, S., Javaid, M., Prakash, C., & Budhhi, D. (2022). Artificial intelligence and advanced materials in automotive industry: Potential applications and perspectives. *Materials Today: Proceedings*, *62*, 4207–4214. https://doi.org/10.1016/J.MATPR.2022.04.727

Sutton, R. S. (1992). Introduction: The Challenge of Reinforcement Learning. *Reinforcement Learning*, 1–3. https://doi.org/10.1007/978-1-4615-3618-5_1

Sutton, R. S., & Barto, A. G. (1999). *Book Reviews Reinforcement Learning*.

Tian, H., Wang, T., Liu, Y., Qiao, X., & Li, Y. (2020). Computer vision technology in agricultural automation —A review. *Information Processing in Agriculture*, *7*(1), 1–19. https://doi.org/10.1016/J.INPA.2019.09.006

University, L. N.-S. of Computing. S., & 2005, undefined. (2005). Multilayer perceptron tutorial. *Citeseer*. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=4c8339b893423f1e14e34cc1543faee4e5ee4244

Van den Broeck, G., Lykov, A., Schleich, M., & Suciu, D. (2022). On the Tractability of SHAP Explanations. *Journal of Artificial Intelligence Research*, *74*, 851–886. https://doi.org/10.1613/JAIR.1.13283

van Otterlo, M., & Wiering, M. (2012). Reinforcement learning and markov decision processes. *Adaptation, Learning, and Optimization*, *12*, 3–42. https://doi.org/10.1007/978-3-642-27645-3_1/COVER

Yin, W., Kann, K., Yu, M., SchützeSch, H., & Munich, L. (2017). *Comparative Study of CNN and RNN for Natural Language Processing*.

Zhang, H. (2022). Prediction of Used Car Price Based on LightGBM. *2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*, 327–332. https://doi.org/10.1109/AEMCSE55572.2022.00073

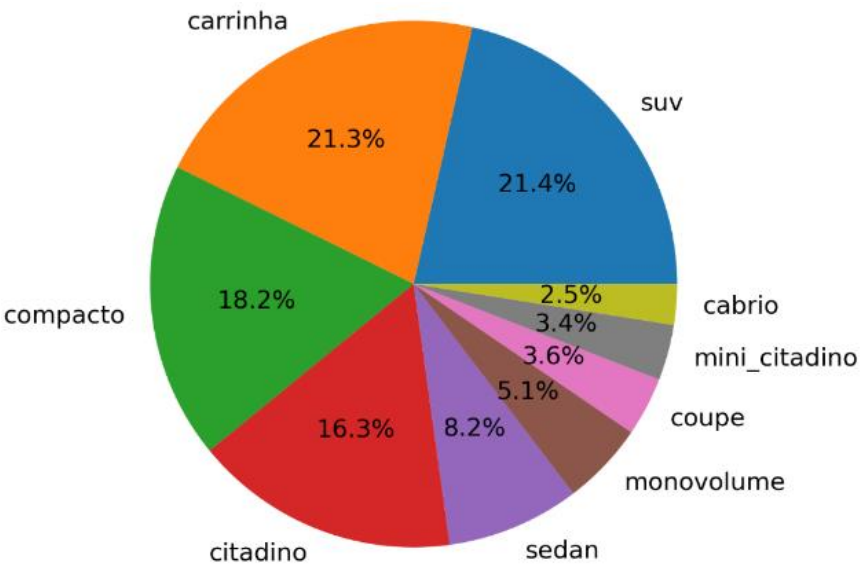
Zhang, M. L., & Zhou, Z. H. (2007). ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, *40*(7), 2038–2048. https://doi.org/10.1016/J.PATCOG.2006.12.019

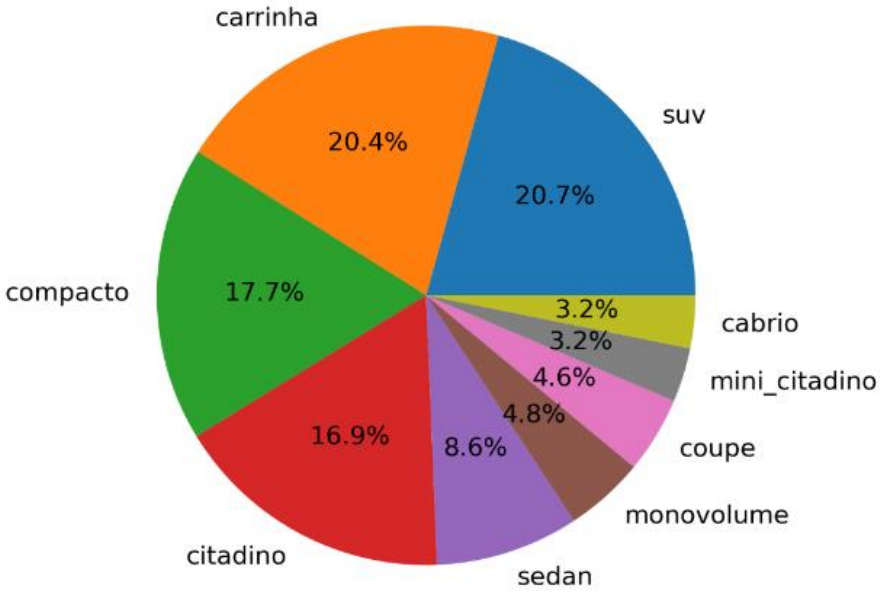
Zhang, Z., & Jung, C. (2019). *GBDT-MO: Gradient Boosted Decision Trees for Multiple Outputs*.

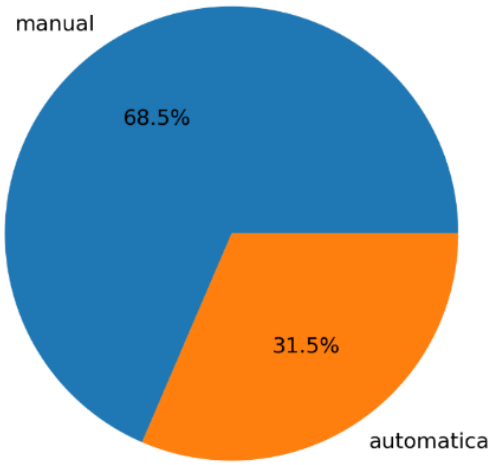
Anexo A

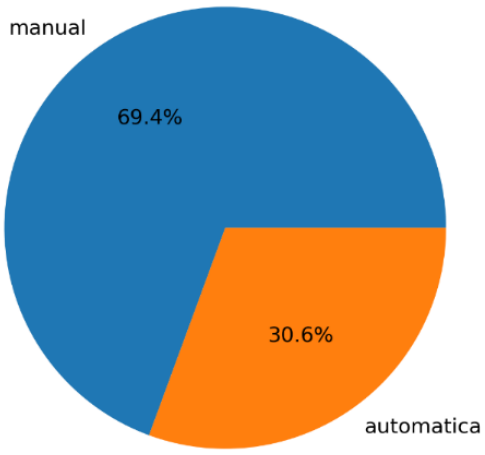
Os gráficos neste anexo mostram a percentagem de carros consoante o tipo de carro e o tipo de caixa e mostram também o número de carros consoantes o tipo de combustível e a marca do carro:

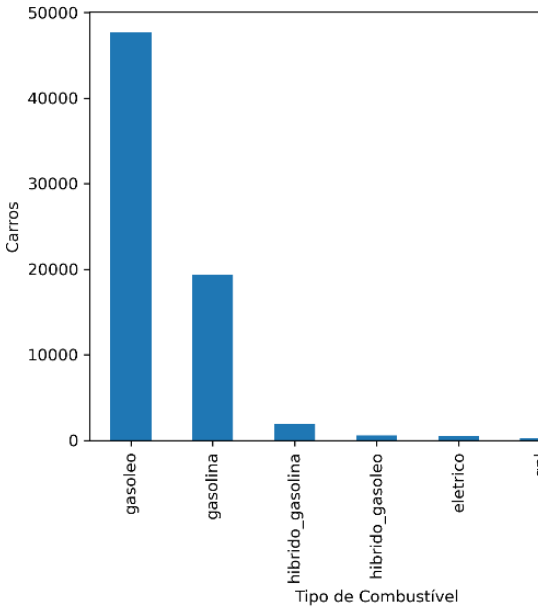
- A primeira fila apresenta a percentagem de cada tipo de carro existente no conjunto de dados B;
- A segunda fila apresenta a percentagem de cada tipo de carro existente no conjunto de dados C;
- A terceira fila apresenta a percentagem do número de carros com caixa manual e automática existente no conjunto de dados B (esquerda) e no conjunto de dados C (direita);
- A quarta fila apresenta o número de carros de cada tipo de combustível existente no conjunto de dados B (esquerda) e no conjunto de dados C (direita);
- A quinta fila apresenta o número de carros de cada marca existente no conjunto de dados B;
- A sexta fila apresenta o número de carros de cada marca existente no conjunto de dados C.

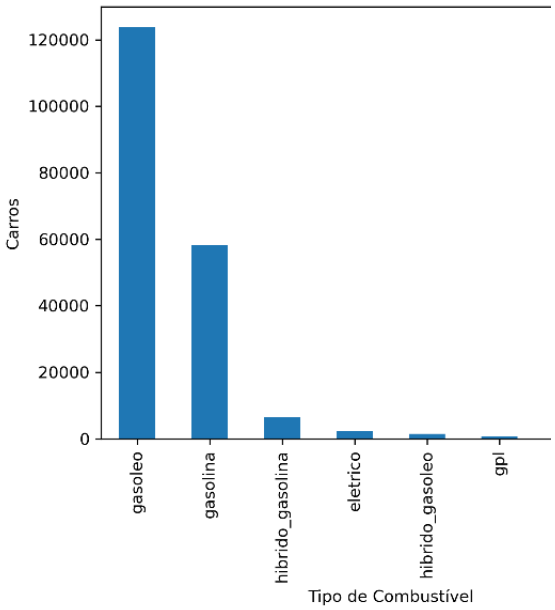


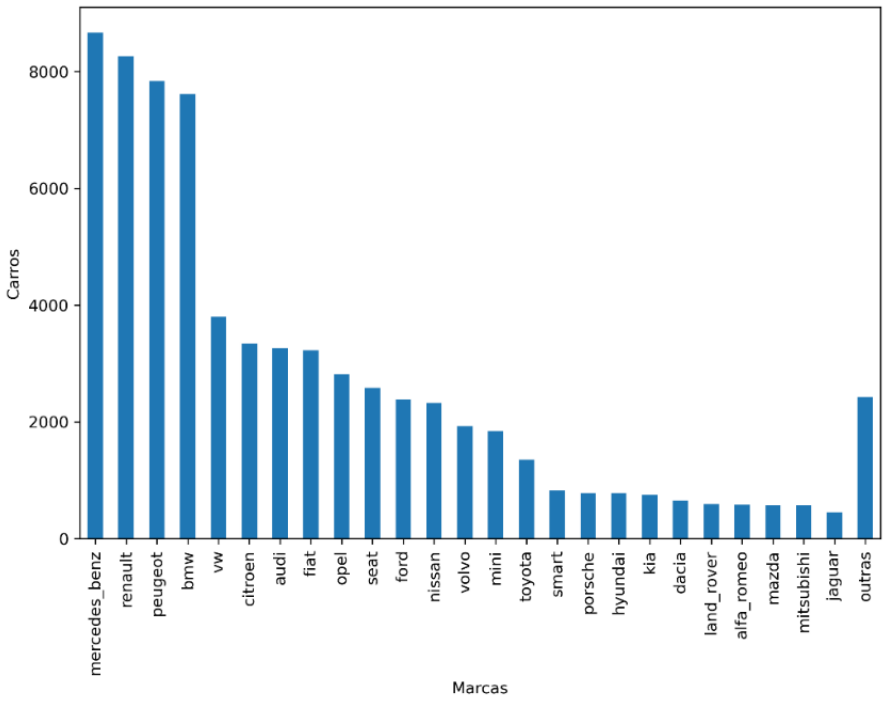

carrinha
20.4%
suv
20.7%
compacto
17.7%
3.2%
cabrio
3.2%
mini_citadino
4.6%
coupe
4.8%
monovolume
16.9%
8.6%
citadino
sedan

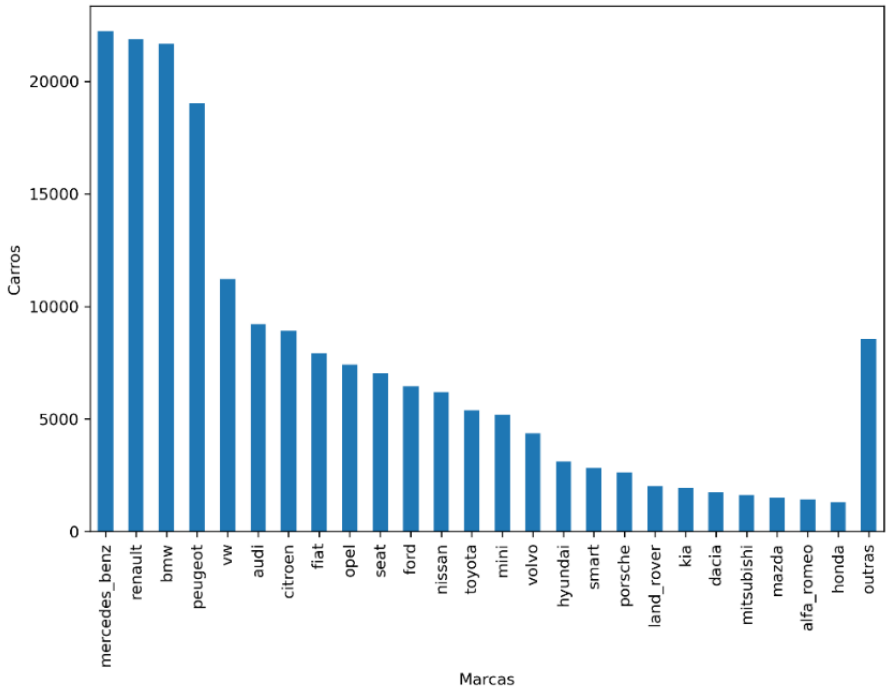

manual
68.5%
31.5%
automatica


manual
69.4%
30.6%
automatica


50000
40000
30000
20000
10000
0
Carros
gasoleo
gasolina
hibrido_gasolina
hibrido_gasoleo
eletrico
Tipo de Combustível


120000
100000
80000
60000
40000
20000
0
Carros
gasoleo
gasolina
hibrido_gasolina
eletrico
hibrido_gasoleo
gpl
Tipo de Combustível

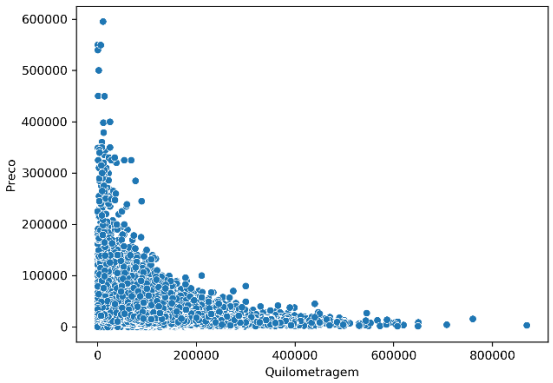

8000
6000
4000
2000
0
Carros
mercedes_benz
renault
peugeot
bmw
vw
citroen
audi
fiat
opel
seat
ford
nissan
volvo
mini
toyota
smart
porsche
hyundai
kia
dacia
land_rover
alfa_romeo
mazda
mitsubishi
jaguar
outras
Marcas

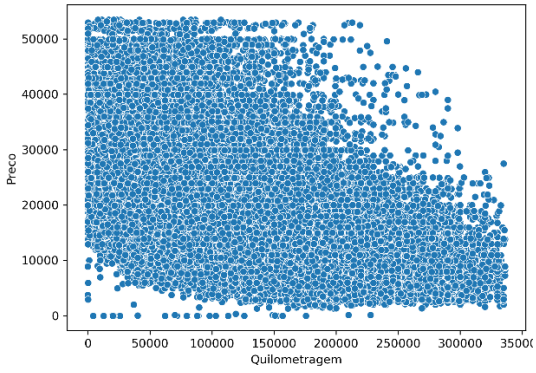

Carros
20000
15000
10000
5000
0
mercedes_benz
renault
bmw
peugeot
vw
audi
citroen
fiat
opel
seat
ford
nissan
toyota
mini
volvo
hyundai
smart
porsche
land_rover
kia
dacia
mitsubishi
mazda
alfa_romeo
honda
outras
Marcas

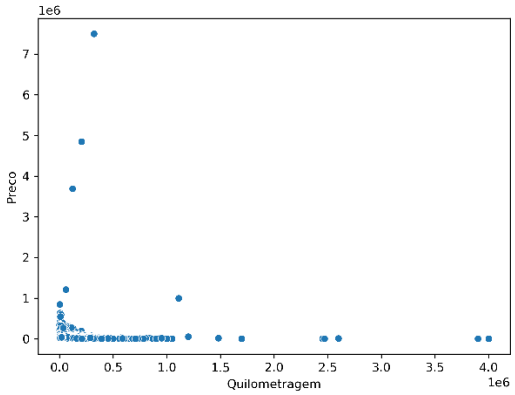
Anexo B

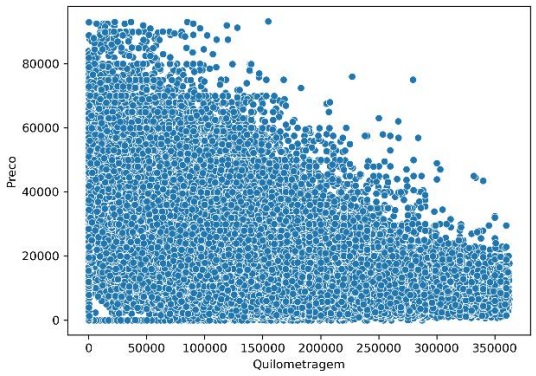
Os gráficos do seguinte anexo apresentam a relação entre o preço e quilometragem, ano, potência e cilindrada, antes (esquerda) e depois (direita) de serem submetidos à remoção de outliers:

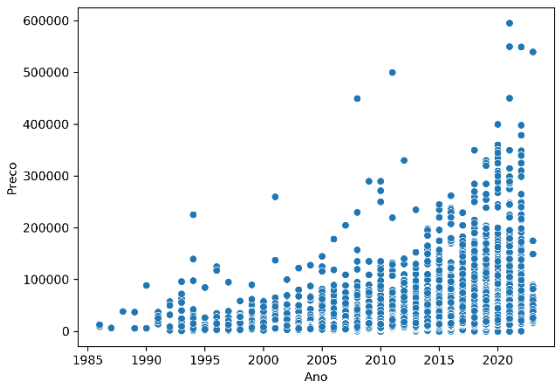
- a primeira fila apresenta a relação entre preço e quilometragem do conjunto de dados B;
- a segunda fila apresenta a relação entre preço e quilometragem do conjunto de dados C;
- a terceira fila apresenta a relação entre preço e ano do conjunto de dados B;
- a quarta fila apresenta a relação entre preço e ano do conjunto de dados C;
- a quinta fila apresenta a relação entre preço e potência do conjunto de dados B;
- a sexta fila apresenta a relação entre preço e potência do conjunto de dados C;
- a sétima fila apresenta a relação entre preço e cilindrada do conjunto de dados B;
- a oitava fila apresenta a relação entre preço e cilindrada do conjunto de dados C;

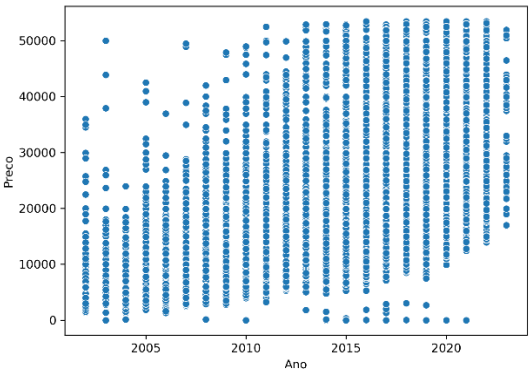


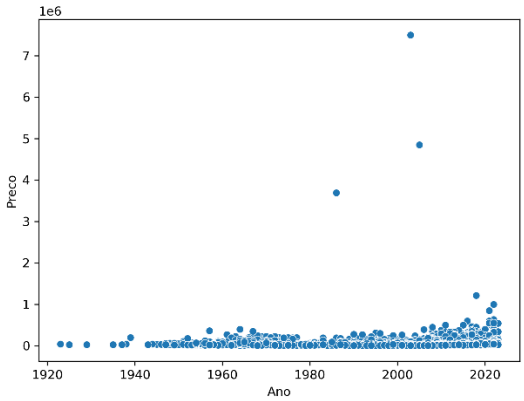


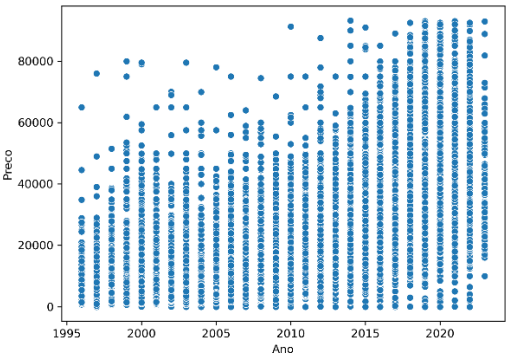


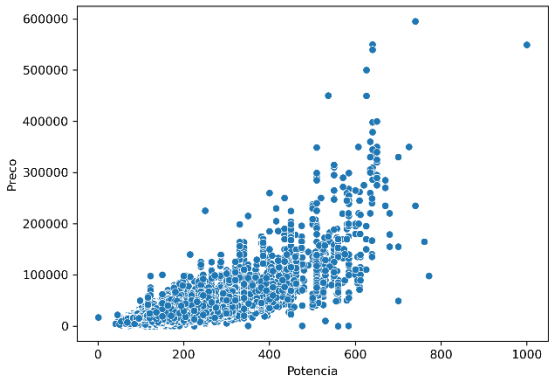


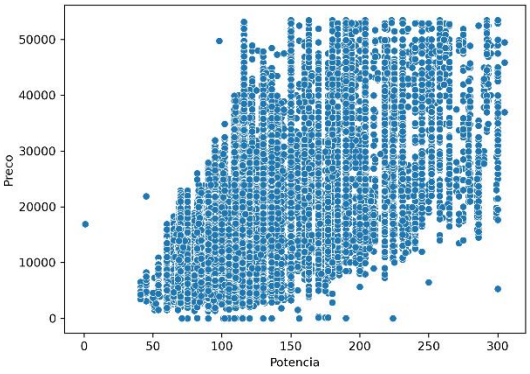

Preco
Ano

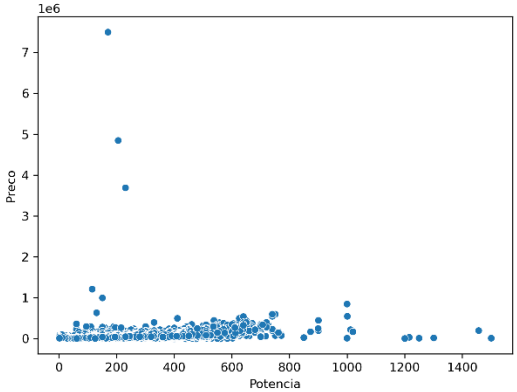

Preco
Ano

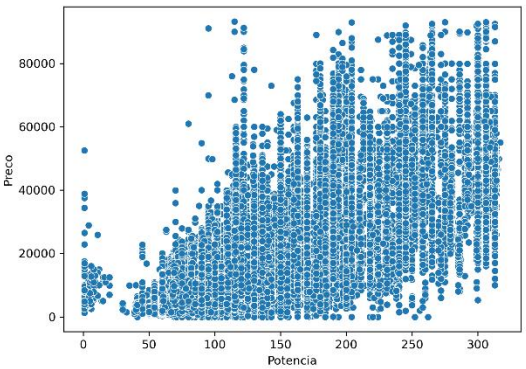

1e6
Preco
Ano

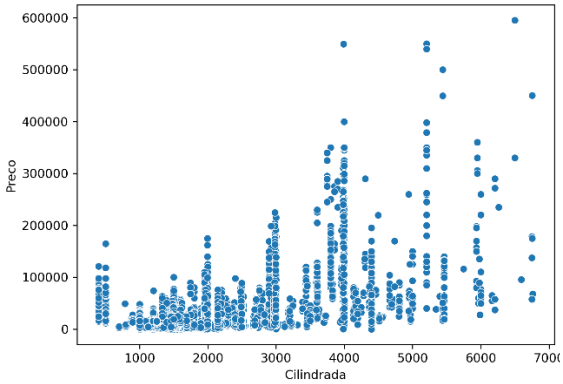

Preco
Ano

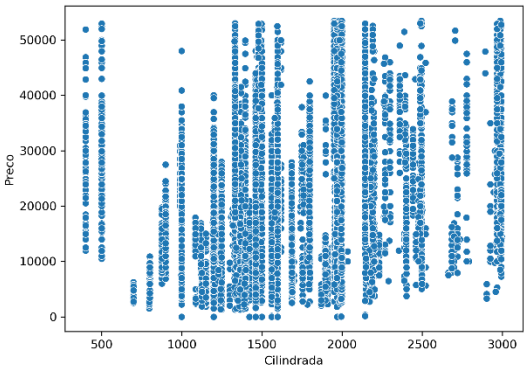

Preco
Potencia

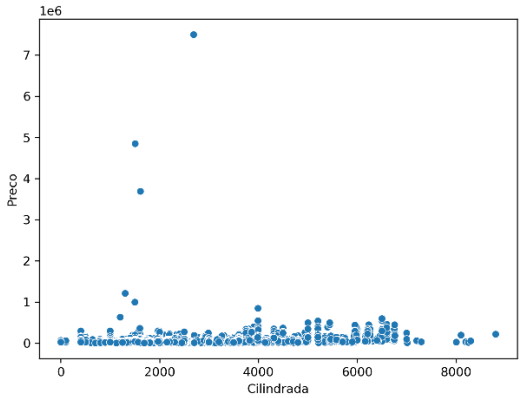

Preco
Potencia

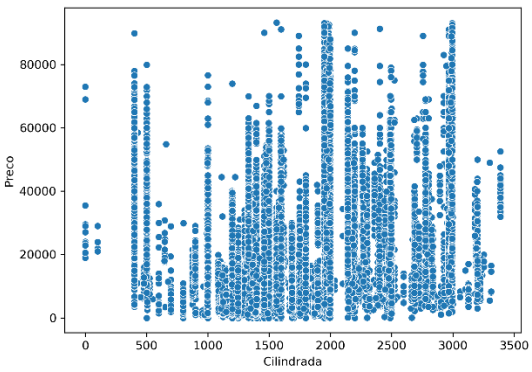

1e6
Preco
Potencia


Preco
Potencia


Preco
Cilindrada
600000
500000
400000
300000
200000
100000
0
1000
2000
3000
4000
5000
6000

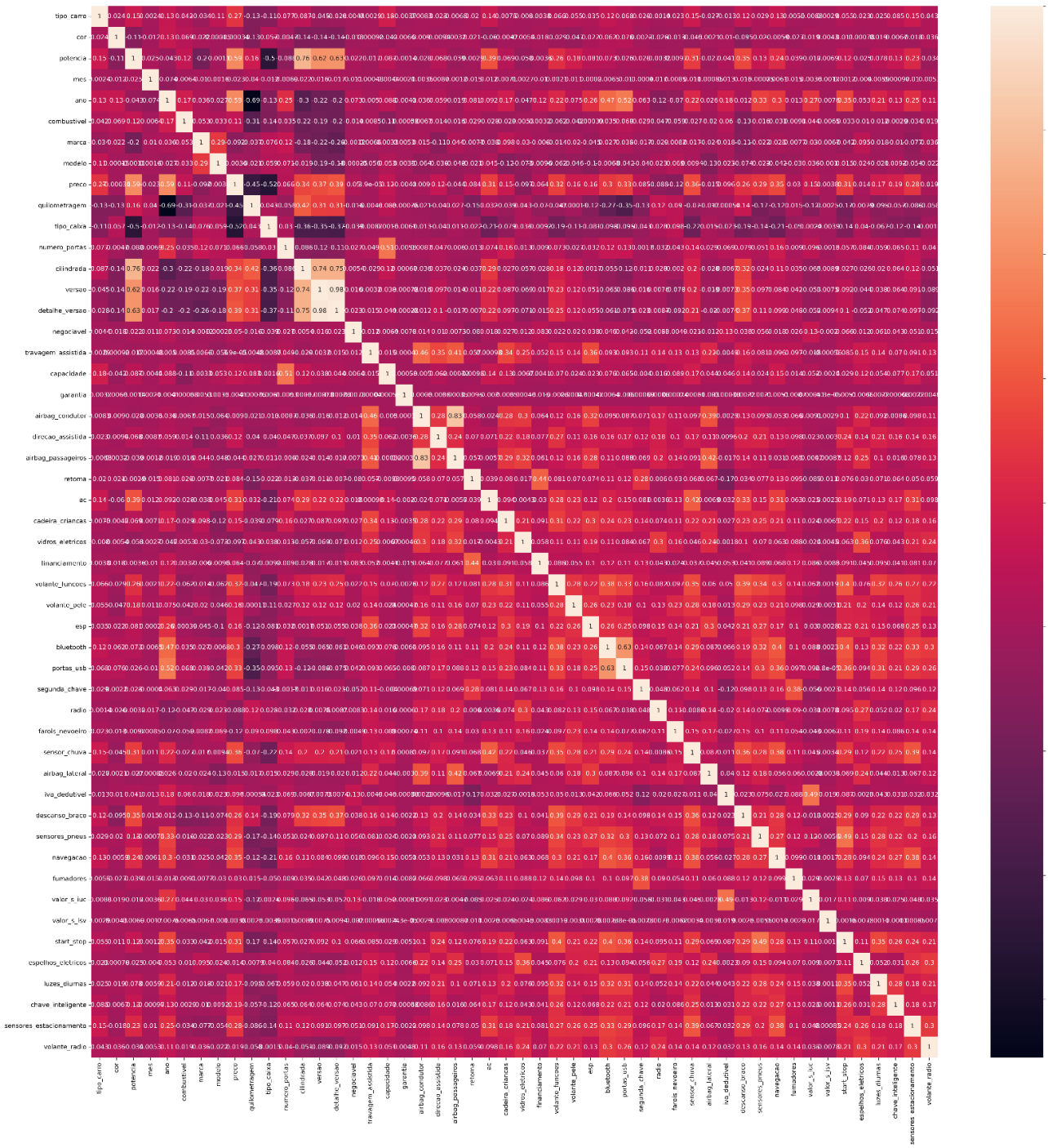

Preco
Cilindrada
50000
40000
30000
20000
10000
0
500
1000
1500
2000
2500
3000

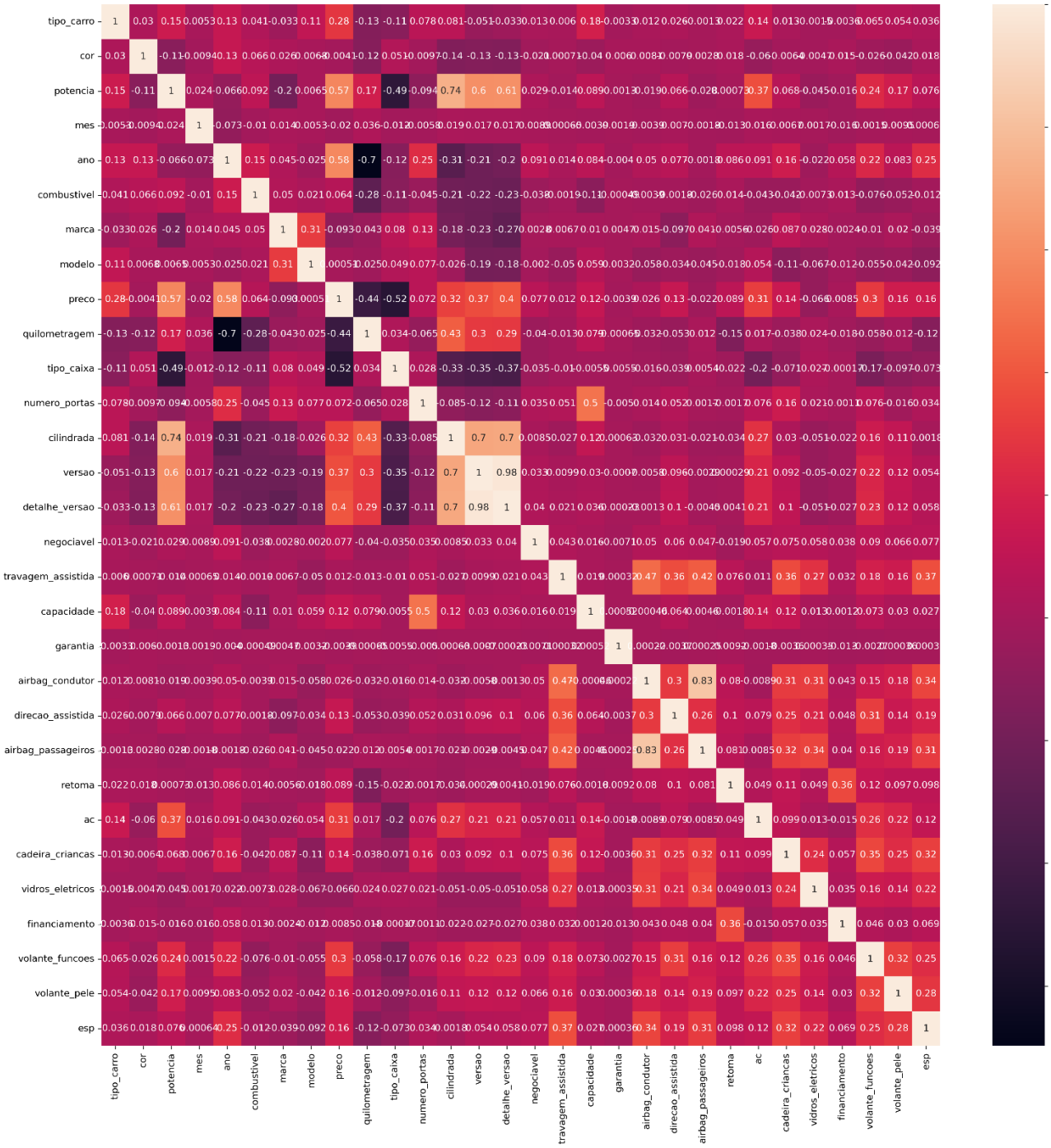

1e6
Preco
Cilindrada
7
6
5
4
3
2
1
0
0
2000
4000
6000
8000


Preco
Cilindrada
80000
60000
40000
20000
0
0
500
1000
1500
2000
2500
3000
3500

Anexo C

Os gráficos do seguinte anexo apresentam a matriz de correlação do conjunto de dados A com as 50 características e a matriz de correlação do conjunto de dados B com 30 características, na ordem nos quais os mesmos aparecem abaixo.




tipo_carro
cor
potencia
mes
ano
combustivel
marca
modelo
preco
quilometragem
tipo_caixa
numero_portas
cilindrada
versao
detalhe_versao
negociavel
travagem_assistida
capacidade
garantia
airbag_condutor
direcao_assistida
airbag_passageiros
retoma
ac
cadeira_criancas
vidros_eletricos
financiamento
volante_funcoes
volante_pele
esp
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6