Politécnico
Castelo Branco
Escola Superior
de Tecnologia

# Análise Preditiva de Preços de Automóveis Usados
## Projeto 1

João Luís Lebre Ramos Antunes Mendes 20200705

Pedro Miguel da Fonseca e Silva 20200766

**Orientador**

Eduardo Sabina dos Santos Valente

Trabalho de Projeto apresentado à Escola Superior de Tecnologia de Castelo Branco do Instituto Politécnico de Castelo Branco para cumprimento dos requisitos necessários à obtenção do grau de Engenheiro em Engenharia Informática, realizada sob a orientação científica do Professor Doutor Eduardo Valente, do Instituto Politécnico de Castelo Branco.

**2025**

## Composição do júri

Presidente do júri

Doutora, Arminda Guerra Lopes

Vogais

Doutora, Arminda Guerra Lopes

Professor Coordenador Escola Superior de Tecnologia de Castelo Branco

Mestre, José Luís Silva Tavares da Cruz

Professor adjunto Escola Superior de Tecnologia de Castelo Branco

Doutor, Eduardo Sabina dos Santos Valente

Professor adjunto Escola Superior de Tecnologia de Castelo Branco

## Resumo

Este estudo investiga a análise preditiva de preços de automóveis usados, destacando a importância de prever o valor de mercado dos veículos com base em características como marca, quilometragem, ano de fabrico e tipo de combustível. A pesquisa explora diversas técnicas de machine learning para gerar previsões mais precisas, incluindo regressão linear, florestas aleatórias (*random forest*) e redes neuronais artificiais.

A metodologia adotada baseia-se na análise de *dataset*s do mercado automóvel do Reino Unido, examinando padrões de desvalorização e identificando os principais fatores que influenciam a variação dos preços ao longo do tempo. O estudo evidencia que a abordagem tradicional de regressão linear apresenta limitações na modelagem da depreciação dos veículos, sendo a curva em S um modelo mais adequado para representar esse comportamento de forma realista.

Os resultados obtidos indicam que modelos baseados em *random forest* oferecem maior precisão preditiva em comparação com métodos estatísticos convencionais. A análise gráfica, incluindo heatmaps e regressões, reforça a importância da quilometragem, marca e ano de fabrico como variáveis determinantes na precificação de veículos usados.

Por fim, o estudo propõe o desenvolvimento de uma aplicação funcional baseada nos modelos explorados, permitindo prever preços futuros de automóveis usados. Essa ferramenta visa auxiliar consumidores, concessionárias e seguradoras a tomarem decisões mais informadas no mercado de segunda mão, reduzindo incertezas e aumentando a eficiência das transações comerciais.



## Palavras-chave

Análise preditiva, Machine learning, Preços de automóveis usados, Regressão, Florestas aleatórias, Curva em S.

## Abstract

This study investigates the predictive analysis of used car prices, emphasizing the importance of estimating vehicle market value based on factors such as brand, mileage, year of manufacture, and fuel type. The research explores various machine learning techniques to generate more accurate predictions, including linear regression, random forests, and artificial neural networks.

The methodology is based on the analysis of datasets from the UK used car market, examining depreciation patterns and identifying the key factors influencing price variations over time. The study highlights that the traditional linear regression approach has limitations in modeling vehicle depreciation, while the S-curve proves to be a more suitable model for realistically representing this behavior.

The results indicate that random forest-based models provide higher predictive accuracy compared to conventional statistical methods. Graphical analysis, including heatmaps and regression models, reinforces the significance of mileage, brand, and year of manufacture as key variables in used vehicle pricing.

Finally, the study proposes the development of a functional application based on the explored models, enabling the prediction of future used car prices. This tool aims to assist consumers, dealerships, and insurance companies in making more informed decisions in the second-hand market, reducing uncertainties, and increasing transaction efficiency.

## Keywords

Predictive analysis, Machine learning, Used car prices, Regression, Random forests, S-curve.

# Índice geral

## Conteúdo

# Índice de figuras

# Lista de tabelas

# 1. Introdução

A análise preditiva de preços de automóveis usados tem-se tornado uma ferramenta essencial no setor automóvel, permitindo estimativas mais precisas do valor de mercado de veículos com base em diversas variáveis, como marca, quilometragem, ano de fabrico e tipo de combustível. A evolução das tecnologias de *machine learning* e análise de dados possibilita a identificação de padrões complexos que influenciam a desvalorização dos automóveis ao longo do tempo, proporcionando um suporte valioso para consumidores, concessionárias, seguradoras e plataformas de compra e venda de veículos.

Neste contexto, o presente estudo investiga a aplicação de técnicas avançadas de *machine learning* para prever preços de automóveis usados, comparando diferentes modelos preditivos, incluindo regressão linear, florestas aleatórias (*random forest*) e redes neuronais artificiais. A pesquisa é baseada em *dataset*s do mercado automóvel do Reino Unido e foca-se na identificação dos fatores que mais impactam a valorização ou desvalorização dos veículos.

Um dos destaques desta investigação é a exploração da curva em S como um modelo mais realista para descrever a depreciação dos automóveis ao longo do tempo, contrastando com as abordagens tradicionais baseadas em regressão linear [1]. Através da análise de dados históricos e do desenvolvimento de modelos preditivos, pretende-se compreender melhor a dinâmica do mercado e fornecer uma ferramenta robusta para estimar o preço futuro de veículos usados.

Além de apresentar os fundamentos teóricos e metodológicos da análise preditiva, este estudo propõe a criação de uma aplicação funcional que permitirá aos utilizadores obter previsões personalizadas de preços com base nos modelos explorados. Este recurso será de grande utilidade para otimizar transações comerciais, reduzir riscos e melhorar a tomada de decisões no mercado de segunda mão.

Por fim, serão discutidas as principais limitações dos métodos utilizados, como a qualidade dos dados e as variações do mercado, bem como as oportunidades para aprimorar a precisão das previsões, incluindo o uso de novas fontes de dados, integração com sensores IoT e técnicas de processamento de linguagem natural para análise de descrições de veículos [2]. O estudo contribui, assim, para o avanço das metodologias preditivas aplicadas ao setor automóvel, reforçando a importância da inovação tecnológica na avaliação e comercialização de automóveis usados.

## 2. Estado da Arte

A **depreciação** de um veículo é um fenômeno natural que ocorre ao longo do tempo, independentemente do uso. Nos primeiros anos, essa desvalorização é mais acentuada, podendo atingir entre 20% e 30% no primeiro ano, estabilizando nos anos seguintes. Marcas premium tendem a perder menos valor devido à perceção de qualidade e prestígio. Além disso, a quilometragem é um fator determinante na avaliação do desgaste de um carro [3].

Veículos com **quilometragem** elevada são vistos como menos confiáveis, enquanto quilometragens muito baixas podem levantar suspeitas de adulteração [4]. O **estado geral do veículo** também influencia diretamente o seu valor de revenda. Pintura danificada, interiores desgastados e danos estruturais podem desvalorizar significativamente um carro, enquanto modelos bem cuidados tendem a atrair compradores dispostos a pagar mais [5].

A **marca e o modelo** do veículo desempenham um papel importante na sua valorização ou desvalorização. Marcas conhecidas pela durabilidade e confiabilidade, como Toyota e Honda, geralmente mantêm melhor o valor ao longo dos anos, enquanto modelos de nicho ou edições limitadas podem ser mais procurados e menos afetados pela desvalorização. Além disso, o **histórico de manutenção** influencia diretamente o valor do carro. Um veículo com todas as revisões registadas em oficinas autorizadas transmite mais confiança ao comprador, enquanto a ausência dessa documentação pode gerar dúvidas sobre o seu estado mecânico [6].

**Alterações no mercado** também impactam diretamente a valorização dos veículos. O aumento da procura por determinado tipo de carro, como SUVs, pode elevar seu preço, enquanto a introdução de modelos tecnologicamente mais avançados pode depreciar rapidamente os anteriores [7].

O **tipo de combustível e a eficiência** energética são fatores cada vez mais relevantes, especialmente diante da flutuação dos preços dos combustíveis e dos incentivos para veículos elétricos e híbridos. Modelos movidos a gasolina ou diesel podem sofrer maior depreciação em mercados onde há forte incentivo à eletrificação [3].

**Fatores econômicos e políticos** também exercem influência no mercado de carros usados. Durante crises econômicas, muitos consumidores optam por veículos seminovos em vez de novos, aumentando a demanda e influenciando os preços. Da mesma forma, novas regulamentações ambientais ou aumentos de impostos podem reduzir a atratividade de determinados tipos de veículos, como os a diesel [7].

O nível de **equipamentos e a versão** do modelo são outros aspetos relevantes. Veículos com mais recursos tecnológicos, como sistemas de navegação, bancos de couro e assistências à condução, costumam manter melhor o valor, enquanto equipamentos obsoletos podem não agregar tanto à revenda [8].

A **localização geográfica** também afeta o valor de um veículo, pois certas regiões podem ter preferências específicas. Em áreas rurais, veículos utilitários e SUVs são mais valorizados, enquanto em centros urbanos, carros compactos são preferidos. O clima local também pode influenciar a conservação do veículo, como a corrosão em regiões litorâneas [7]. Além disso, as **tendências tecnológicas** desempenham um papel importante na valorização ou desvalorização de um carro. Modelos equipados com as tecnologias mais recentes, como conectividade avançada e sistemas de assistência à condução, tendem a manter melhor o seu valor de mercado [3]. Por outro lado, a **concorrência entre modelos** pode acelerar a depreciação, especialmente quando há saturação de um determinado segmento. Modelos populares e com bom custo-benefício podem pressionar a desvalorização de seus concorrentes diretos [8].

Todos esses fatores demonstram que a valorização ou desvalorização de um veículo depende de uma combinação de variáveis que vão desde características do próprio carro até fatores econômicos e tecnológicos do mercado.

## 2.1. Análise de Documentos

Para a condução deste estudo, adotou-se a metodologia PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*), amplamente utilizada em revisões sistemáticas e meta-análises. O processo envolveu a identificação, seleção e análise de artigos científicos relevantes publicados em bases de dados indexadas. Os critérios de inclusão consideraram estudos que abordam a previsão de preços de automóveis usados utilizando modelos estatísticos e técnicas de *machine learning*, enquanto estudos sem metodologias quantitativas ou sem aplicação prática foram excluídos. A extração de dados seguiu um protocolo padronizado, garantindo reprodutibilidade e confiabilidade dos resultados.

O documento "**An S-curve Model on the Maximum Predictive Pricing of Used Cars**" [9] tomou como estudo que a previsão de preços de bens e produtos, como carros usados, é uma área de grande relevância na economia, uma vez que influencia decisões de compra, venda e seguro. O estudo do comportamento dos preços ao longo do tempo é essencial, especialmente em um contexto no qual os carros novos se tornam mais caros, levando as famílias a optarem por alternativas mais acessíveis.

Historicamente, modelos de previsão têm sido amplamente utilizados para estimar esses preços, mas a identificação do modelo mais adequado ainda representa um desafio. Modelos lineares, em particular a regressão linear simples, são os métodos mais básicos e frequentemente aplicados em análises preditivas, estabelecendo uma relação linear entre uma variável independente e uma variável dependente.

Diversos estudos demonstraram que a regressão linear é eficaz para identificar relações positivas entre variáveis; entretanto, a sua aplicação é limitada pela suposição de linearidade, o que frequentemente resulta em erros residuais crescentes à medida que os dados se afastam da média central. Essa imprecisão torna os modelos lineares inadequados para situações reais em que os preços não se depreciam de forma constante ao longo do tempo.

Pesquisas, como a de Mamipour e Jezeie [10], evidenciaram que relações não-lineares entre variáveis econômicas e financeiras produzem resultados superiores, e Ferrari et al. [11] confirmaram que abordagens não-lineares apresentam melhor desempenho na análise de ciclos econômicos. Diante dessas limitações, modelos não-lineares oferecem uma flexibilidade maior para descrever comportamentos complexos, proporcionando maior precisão na previsão de preços e no tratamento dos erros de previsão. Entre esses, a curva em "S" se destaca por sua capacidade de capturar processos de crescimento e difusão, sendo amplamente utilizada para modelar fenômenos naturais e sociais.

A curva logística, uma forma específica da curva em "S", caracteriza-se por três fases principais: um crescimento inicial lento, seguido por uma fase de aceleração e, finalmente, uma estabilização. Essa abordagem tem sido aplicada em diversos contextos, como na análise de redes neurais artificiais, previsão de produção de gás natural [12] e gestão de projetos [13] embora sua aplicação na previsão de preços de bens, como carros usados, ainda seja pouco explorada.

A utilização da curva em "S" no contexto de preços de carros usados oferece uma alternativa prática e precisa aos modelos lineares, uma vez que a depreciação do valor de um carro antigo não ocorre de maneira linear, apresentando, ao contrário, uma estabilização que pode ser representada pelo comportamento assintótico da curva. Estudos que empregaram a função sigmoide em redes neurais demonstraram como a introdução da não-linearidade pode melhorar os modelos preditivos, e outras pesquisas utilizaram a curva em "S" para prever poupanças financeiras [14] e como ferramenta de monitorização e controle de projetos [13], reforçando seu potencial na modelagem de comportamentos complexos de mercado.

O estudo "**Used Cars Price Prediction using Machine Learning with Optimal Features**" comenta como a previsão de preços de carros usados é uma área crescente de pesquisa, impulsionada pela popularidade desse segmento no mercado automotivo e pela complexidade em determinar valores justos. O aumento do comércio eletrônico e de plataformas online intensificou a necessidade de sistemas preditivos que considerem atributos variados dos veículos, como tipo de combustível, quilometragem e condição geral. Nesse contexto, a aplicação de algoritmos de *machine learning* (ML) [15] tem mostrado avanços significativos, combinando técnicas estatísticas e computacionais para lidar com grandes volumes de dados e aprimorar a acurácia das previsões.

Os modelos de *machine learning* aplicados à previsão de preços podem ser divididos em abordagens supervisionadas e não supervisionadas. Entre as técnicas supervisionadas, destacam-se os modelos de regressão linear [15], *random forest* e redes neurais profundas. A regressão linear, amplamente utilizada como ponto de partida para prever preços de carros usados, apresentou limitações na captura de relações não lineares entre variáveis [1], o que motivou o desenvolvimento de variantes, como a regressão polinomial e modelos baseados em *regularization* (por

exemplo, Lasso e Ridge). Essas variantes contribuem para reduzir o sobre ajuste e melhorar a precisão preditiva.

O algoritmo *random forest* [16], por sua vez, é frequentemente adotado para a previsão de preços devido à sua capacidade de lidar com dados categóricos e alta dimensionalidade. Estudos indicam que esse modelo apresenta elevada acurácia, demonstrando resultados consistentes em diferentes conjuntos de dados. A combinação de *random forest* com técnicas de seleção de atributos, como o *Recursive Feature Elimination* (RFE) [16], tem-se mostrado uma abordagem robusta e eficiente para aprimorar a performance dos modelos preditivos.

Além disso, as redes neurais profundas, especialmente arquiteturas que combinam *long short-term memory* (LSTM) e redes convolucionais (CNN), têm sido exploradas para capturar padrões complexos e dependências temporais nos dados. Essa capacidade torna-as particularmente adequadas para cenários em que atributos como variações históricas de preços ou condições de mercado precisam ser considerados. Pesquisas demonstram que redes LSTM podem superar modelos mais tradicionais, especialmente na previsão dos intervalos mínimo e máximo de preços [17].

A avaliação de desempenho desses modelos é realizada por meio de métricas como o erro absoluto médio (MAE), o erro quadrático médio (MSE) e o coeficiente de determinação ($R^2$). Estudos recentes sugerem que o uso de ensemble methods, que combinam *random forest* e regressão, tende a produzir menores valores de erro absoluto, indicando maior precisão preditiva. Além disso, técnicas de validação cruzada e otimização de hiperparâmetros, como *grid search ou Bayesian optimization*, são fundamentais para assegurar que os modelos generalizem bem para dados fora da amostra [17].

Entretanto, a implementação desses modelos enfrenta desafios significativos. A qualidade dos dados é um dos principais obstáculos, pois a presença de atributos irrelevantes ou dados inconsistentes pode afetar negativamente o desempenho dos modelos. Problemas de multicolinearidade, decorrentes de fortes correlações entre variáveis independentes comuns em conjuntos de dados de carros usados, também exigem a realização de análises estatísticas, como o cálculo *do Variance Inflation Factor* (VIF). Adicionalmente, a exploração de dados não estruturados, como comentários de usuários e descrições das condições dos veículos, representa uma oportunidade para enriquecer as análises, embora exija avanços em técnicas de processamento de linguagem natural.

Por outro lado, as oportunidades para aprimorar a previsão de preços de carros usados incluem a integração com a Internet das Coisas (IoT), que pode fornecer dados em tempo real através de sensores veiculares, e o uso de plataformas baseadas em blockchain, que podem garantir a autenticidade e a transparência dos históricos de manutenção e uso dos veículos, aumentando a confiança nos modelos preditivos [2].

A previsão de preços de carros usados apresenta desafios significativos devido à diversidade de fatores que influenciam os valores, como ano do modelo,

quilometragem, tipo de combustível e condição geral. A crescente demanda por veículos de segunda mão e o aumento dos preços de combustíveis tornam esse tema ainda mais relevante. Este estudo analisa técnicas de aprendizado supervisionado, como regressão linear, regressão Lasso e *random forest*, além de métodos estatísticos, comparando seus resultados em termos de acurácia e eficiência.

No documento "**Used Car Price Prediction Using Machine Learning Techniques**" foram analisados diferentes modelos de machine learning para a previsão de preços de carros usados, com o objetivo de avaliar a eficácia e precisão de cada abordagem. Cada modelo apresenta características distintas, sendo alguns mais adequados para conjuntos de dados lineares, enquanto outros se destacam em cenários com múltiplas variáveis correlacionadas. [18] Dentre os modelos utilizados, a regressão linear é amplamente empregue para modelar relações entre a variável dependente (preço) e variáveis independentes (atributos do carro), mas sua simplicidade e a suposição de linearidade podem levar a resultados imprecisos em conjuntos de dados mais complexos. [19] Em contrapartida, a regressão Lasso ("Least Absolute Shrinkage and Selection Operator") introduz um fator de penalização que reduz a multicolinearidade e simplifica os modelos, sendo eficaz na seleção de variáveis relevantes ao eliminar atributos irrelevantes à medida que o parâmetro $\lambda$ aumenta. O algoritmo *random forest*, que combina múltiplas árvores de decisão, melhora a precisão das previsões e reduz o risco de *overfitting*, sendo especialmente eficaz para dados de alta dimensionalidade, embora apresente maior complexidade computacional quando comparado a modelos mais simples. Além destes, modelos estatísticos, como a regressão por métodos dos mínimos quadrados, também foram explorados para descrever matematicamente as relações entre as variáveis; contudo, apesar de oferecerem boas bases analíticas, geralmente não se adaptam tão bem quanto as técnicas de *machine learning.*

Uma análise por meio de *heat map* foi realizada para identificar correlações entre o preço e diversos atributos, como "car width", "wheel base" e "city mpg". Essa ferramenta facilitou a seleção de variáveis relevantes, evidenciando que o preço apresenta correlações positivas com dimensões do veículo e negativas com a eficiência de combustível. A metodologia adotada envolveu a coleta de dados a partir de fontes como o Kaggle, abrangendo mais de 20 mil amostras com atributos relevantes. Após a limpeza e transformação dos dados, foram realizadas etapas de análise exploratória dos atributos, construção e avaliação dos modelos de *machine learning* e, finalmente, a comparação do desempenho entre eles [19].

Em termos de resultados, a regressão linear revelou-se um modelo básico, com limitações em cenários complexos. Por outro lado, a regressão Lasso demonstrou melhor desempenho em situações com multicolinearidade, embora tenha apresentado uma capacidade preditiva inferior à do *random forest*. Este último destacou-se pela maior precisão, atingindo uma acurácia de 91,43%, comprovando-se a técnica mais eficaz no contexto do estudo [16].

## 2.4. Análise comparativa

A comparação entre os trabalhos analisados e o presente estudo evidencia semelhanças e distinções significativas no contexto da previsão de preços de automóveis usados, tanto no que se refere às abordagens metodológicas quanto aos resultados obtidos.

Uma das principais convergências entre os estudos reside na utilização de técnicas de *machine learning* para prever os preços de automóveis usados. Tanto o presente projeto quanto os trabalhos analisados exploram métodos como regressão linear, *random forest* e redes neurais profundas. No entanto, uma diferenciação fundamental do estudo em questão é a ênfase na curva em S como modelo alternativo à regressão linear, com o intuito de representar de forma mais realista a depreciação dos veículos ao longo do tempo. Essa abordagem também está presente no artigo "An S-curve Model on the Maximum Predictive Pricing of Used Cars" [9], que destaca a aplicabilidade da curva logística na modelagem do valor residual dos automóveis. Por outro lado, os artigos "Used Cars Price Prediction using Machine Learning with Optimal Features" [15] e "Used Car Price Prediction Using Machine Learning Techniques" [20] concentram-se principalmente na implementação e otimização de modelos supervisionados tradicionais, sem considerar explicitamente a aplicação da curva em S.

A metodologia empregada no presente estudo apresenta semelhanças com os trabalhos analisados no que diz respeito à seleção de atributos e análise exploratória de dados. Tal como no artigo "Used Cars Price Prediction using Machine Learning with Optimal Features" [15], são utilizados *heat maps* e métricas estatísticas para avaliar correlações entre variáveis como quilometragem, marca e ano de fabrico, permitindo a identificação dos fatores que mais influenciam a precificação. No entanto, o presente projeto distingue-se ao propor a implementação de uma aplicação funcional baseada nos modelos explorados, conferindo um caráter mais prático à pesquisa, enquanto os outros trabalhos analisados concentram-se essencialmente em avaliações experimentais e teóricas dos modelos preditivos.

*Tabela 1 - Análise comparativa*

| CRITÉRIO | FADZILAH SALIM | MUHAMMAD ASGHAR | SHYAMALI DAS |
|---|---|---|---|
| **FOCO PRINCIPAL** | Curva em "S" para previsão de preços com base na depreciação realista. | *Random Forest* como principal técnica, destacando sua acurácia. | Comparativo entre regressão linear, Lasso e *Random Forest*. |
| **FONTES DE DADOS** | Sites de venda de carros usados na Malásia. | Dados coletados do Kaggle e fontes abertas indianas. | Dados de fontes abertas com 20 mil amostras. |
| **MODELOS AVALIADOS** | Linear, Cúbica e Curva em "S". | *Random Forest*, Lasso, Linear Regression. | Linear Regression, Lasso, *Random Forest*. |
| **ANÁLISE DE CORRELAÇÃO** | Baseada em funções matemáticas. | Uso de Heat Map para identificar correlações-chave. | Heat Map destacando relações entre preço e atributos. |
| **ACURÁCIA PREDITIVA** | Melhor desempenho com a curva em "S" em relação aos modelos lineares. | *Random Forest* com 91,43% de acurácia. | *Random Forest* obteve melhor resultado entre as técnicas. |
| **ABORDAGEM DE SELEÇÃO DE ATRIBUTOS** | Base na experiência dos autores. | RFE (Recursive Feature Elimination) para selecionar atributos relevantes. | Lasso para redução de multicolinearidade. |

Uma diferença relevante diz respeito à origem dos dados analisados. Enquanto o presente estudo se baseia em *datasets* do mercado automóvel do Reino Unido, os artigos analisados utilizam dados provenientes de outros contextos, como Malásia e Índia. Essa distinção geográfica pode impactar a aplicabilidade e generalização dos modelos, uma vez que fatores como preferências de consumidores, políticas económicas e padrões de depreciação variam conforme o mercado. Assim, o presente

estudo contribui ao trazer uma perspetiva europeia, complementando as análises realizadas nos demais trabalhos.

No que se refere à avaliação dos modelos, os estudos analisados utilizam, na sua maioria, erro absoluto médio (MAE), erro quadrático médio (MSE) e coeficiente de determinação ($R^2$) como métricas de desempenho. Os estudos indicam que modelos baseados em *random forest* apresentam maior precisão preditiva, no entanto, a inclusão da curva em S como alternativa metodológica diferencia o presente estudo, uma vez que permite modelar melhor a dinâmica da depreciação, superando as limitações da regressão linear tradicional.

No que concerne aos desafios e oportunidades, o presente estudo e os trabalhos analisados identificam fatores semelhantes, tais como a qualidade dos dados, a presença de multicolinearidade e a necessidade de explorar fontes de dados não estruturadas. Entretanto, o presente projeto destaca-se ao propor a integração de tecnologias emergentes, como blockchain e *IoT*, para aprimorar a precisão e a confiabilidade das previsões de preços, abordagem não contemplada nos estudos analisados.

Em síntese, o presente estudo incorpora elementos metodológicos dos trabalhos analisados, mas distingue-se ao propor um modelo híbrido que combina técnicas avançadas de *machine learning* com uma abordagem baseada na curva em S. Além disso, a intenção de desenvolver uma aplicação prática para prever preços de automóveis usados adiciona um elemento de inovação e aplicabilidade comercial, tornando a pesquisa uma contribuição relevante para o avanço das técnicas preditivas no setor automóvel.

### 2.4.1. Principais Contribuições e Limitações

F. Salim et al. [9] apresentam uma contribuição importante ao introduzir um modelo baseado na curva em "S", que melhor representa a depreciação de veículos ao longo do tempo. Ele também fornece uma comparação entre modelos lineares e não-lineares, destacando a superioridade da abordagem não-linear. No entanto, a aplicabilidade desse modelo está restrita ao mercado automotivo da Malásia, e o estudo não explora suficientemente técnicas modernas de *machine learning*.

P. P. Shinde e D. S. Shah [21] concentram-se no uso do *random forest* como método principal, alcançando uma alta acurácia de 91,43%. Ele também utiliza ferramentas como *Heat Map* e RFE para identificar e selecionar os atributos mais relevantes, o que reforça a robustez de suas análises, algo que virá a ser utilizado neste projeto. Contudo, a dependência de dados específicos da índia limita a generalização de seus resultados, e o estudo carece de discussão sobre o impacto de modelos mais complexos, como redes neurais profundas.

Por sua vez, M. S. Das et al. [20] fornecem um comparativo abrangente entre modelos de regressão linear, Lasso e *random forest*, evidenciando a superioridade do *random forest* em relação aos demais métodos. Apesar disso, o estudo apresenta pouco

aprofundamento em técnicas não supervisionadas ou redes neurais profundas, e o foco limitado ao desempenho preditivo ignora aspetos como os custos computacionais.

# 3. Análise preditiva

## 3.1. Técnicas

Existem várias técnicas de análise preditiva que podem ser aplicadas para prever o preço de carros usados.

### 3.1.1. Regressão Linear

A regressão linear é o algoritmo supervisionado de aprendizado de máquina mais usado, que funciona em trem para prever uma saída bem estabelecida que depende dos dados de entrada. Esses algoritmos geralmente treinam o conjunto e resultam na saída. A Análise de Regressão trata-se de uma metodologia de modelagem preditiva que tem como objetivo investigar a relação entre diversos dados de entrada. Para problema de regressão simples (um único x e um único y), o modelo de formato segue como Y = B0 + B1* X [20].

### 3.1.2. Árvores de Decisão e Florestas Aleatórias

A importância das árvores de decisão na área de *classifiers* decorre, essencialmente, de duas características:

- Árvores de decisão são excelentes preditores, como será mostrado, adiante, através de um exemplo.
- Permitem não só uma visualização gráfica, mas também geram regras de classificação do tipo IF condição THEN resultado.

Para ilustrar essas características, será utilizado aquele que talvez seja o conjunto de dados mais famoso da Estatística: os dados de íris de Fisher. Esse conjunto de dados consiste em 150 observações de três tipos de íris: setosa, versicolor e virginica. Na base têm-se 50 observações de cada tipo.

Cada observação consiste não só na classificação nos três tipos, mas também em quatro medidas:

- Comprimento da pétala.
- Largura da pétala.
- Comprimento da sétala.
- Largura da sétala.

É sabido que as duas variáveis que melhor se discriminam entre os tipos de íris são comprimento e largura da pétala.

Assim, a Figura 1 apresenta o diagrama de dispersão dos dados, indicando os tipos de íris.

Nota-se que as setosas formam um grupo claramente distinto à esquerda e abaixo na Figura 1. Já as versicolors e as virginicas apresentam uma "zona cinzenta" entre elas, o que dificulta o trabalho do classificador **[22]**.

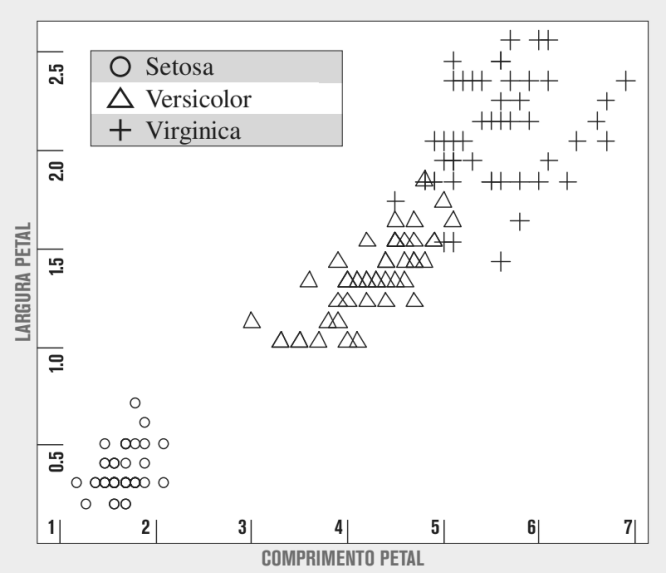


*Figura 1 Dados de Íris*

## 3.2 Aplicações

A análise preditiva de preços de carros usados é amplamente utilizada por consumidores, empresas e instituições financeiras para tomada de decisões estratégicas. Com base em dados históricos, algoritmos de aprendizado de máquina e fatores de mercado, essas previsões ajudam a definir preços mais justos e competitivos.

### 3.2.1. Plataformas de compra e venda de veículos

Sites especializados na compra e venda de automóveis, como Standvirtual, AutoScout24 e mobile.de, utilizam algoritmos de análise preditiva para calcular preços de veículos com base em fatores como ano de fabricação, quilometragem, estado de conservação, localização e tendências de mercado. Essas plataformas oferecem recomendações automáticas para compradores e vendedores, ajudando-os a definir um preço competitivo que atraia interessados sem subestimar o valor do automóvel. Além disso, algumas dessas plataformas utilizam inteligência artificial para detetar anúncios com preços fora do padrão do mercado, alertando os usuários sobre possíveis fraudes ou negócios desvantajosos.

### 3.2.2. Seguradoras e financeiras

Empresas de seguros e instituições financeiras dependem da análise preditiva para avaliar o valor futuro dos veículos, o que impacta diretamente o cálculo de apólices de seguro e os termos de financiamento. No setor de seguros, prever a desvalorização de um veículo ajuda a ajustar o valor das indenizações em caso de sinistro. Já no setor financeiro, bancos e empresas de leasing utilizam essas análises para definir taxas de juros e condições de crédito para a compra de veículos usados, minimizando riscos de perdas.

### 3.2.3. Concessionárias e empresas de leasing

Concessionárias e empresas de leasing utilizam a análise preditiva para determinar os valores de recompra e revenda dos veículos. Essas previsões são essenciais para

avaliar quando é mais vantajoso vender um veículo usado, evitando perdas devido à desvalorização acelerada. Empresas de leasing, por exemplo, precisam prever o valor dos automóveis no final do contrato de locação para definir as melhores condições de recompra ou revenda. Além disso, redes de concessionárias analisam a oferta e a demanda de determinados modelos para planejar estoques de veículos usados de maneira estratégica.

## 3.3. Desafios e limitações

Embora a análise preditiva de preços de carros usados ofereça uma estimativa valiosa, alguns desafios e limitações devem ser considerados.

### 3.3.1. Qualidade dos Dados

A precisão da análise preditiva depende diretamente da qualidade dos dados coletados. Erros nos registos, falta de informações sobre a manutenção do veículo ou inconsistências nos anúncios podem levar a previsões imprecisas. Além disso, algumas bases de dados podem conter informações desatualizadas, não refletindo corretamente a realidade do mercado. Para minimizar esse problema, é fundamental utilizar fontes confiáveis, atualizar periodicamente os bancos de dados e empregar técnicas de limpeza e validação de dados.

### 3.3.2. Mudanças de Mercado

Fatores externos, como mudanças econômicas (inflação, taxas de juros) ou crises (como a pandemia de COVID-19, que impactou a produção de carros novos), afetam a demanda e, consequentemente, os valores dos automóveis.

### 3.3.3. Diferenças Regionais

Os preços dos veículos variam significativamente de uma região para outra, e capturar essa variação pode ser um grande desafio. Alguns exemplos de fatores regionais que influenciam os preços incluem a demanda local que em grandes centros urbanos, carros compactos e elétricos tendem a ser mais valorizados devido à mobilidade e incentivos fiscais. Em regiões rurais, SUVs e picapes têm maior demanda.

Às condições climáticas visto que áreas costeiras podem acelerar a corrosão dos veículos devido à exposição ao sal, enquanto regiões com invernos rigorosos podem afetar componentes mecânicos, desvalorizando os carros.

A infraestrutura rodoviária vista que os locais com estradas precárias podem exigir veículos mais robustos, elevando a procura por modelos com suspensão reforçada e tração 4x4.

Alguns países ou estados possuem taxações diferenciadas para veículos conforme o tipo de combustível ou emissão de poluentes, o que pode impactar a valorização ou desvalorização de determinados modelos.

Para lidar com essas diferenças, modelos preditivos precisam considerar variáveis regionais e integrar bases de dados locais para garantir previsões mais precisas.

# 4. Análise de *Datasets*

O *dataset* intitulado "100,000 UK Used Car Dataset" é uma fonte rica de informações sobre o mercado de automóveis usados no Reino Unido. Ele contém variáveis cruciais como marca, modelo, ano de fabrico, preço, transmissão, quilometragem, tipo de combustível, taxa anual, consumo médio e tamanho do motor. Inicialmente, os dados estavam organizados em ficheiros separados por marca (e.g., Audi, BMW, Ford, entre outros). Para consolidar esta informação, foi utilizado um código em Python que unificou os ficheiros num único *dataset*, adicionando uma nova coluna para identificar a marca de cada veículo. Este processo garantiu uma estrutura mais robusta para a análise e permitiu explorar os dados de forma abrangente.

Após a unificação, o *dataset* passou por uma análise exploratória, onde foram identificados padrões e anomalias nos dados. Verificou-se a qualidade dos dados, tratando valores ausentes, *outliers* e inconsistências. Foram exploradas as distribuições dos principais atributos, como preço, quilometragem, consumo médio e ano de fabrico, com análises segmentadas por marca e modelo. Além disso, foram gerados mapas de correlação (*heatmaps*) para identificar relações importantes entre variáveis, como o impacto da quilometragem no preço ou a influência do tipo de combustível no valor de mercado.

A análise preliminar revelou que marcas premium, como BMW e Mercedes-Benz, apresentam menor desvalorização ao longo do tempo, enquanto marcas generalistas sofrem uma maior redução de valor. Adicionalmente, veículos a diesel mostraram-se mais valorizados em regiões rurais devido à sua eficiência em longas distâncias, enquanto automóveis elétricos e híbridos estão em crescente valorização em áreas urbanas, impulsionados pelas tendências ecológicas e incentivos governamentais. A quilometragem demonstrou uma forte correlação negativa com o preço, reforçando que veículos com maior desgaste são desvalorizados no mercado de segunda mão.

Além dessa análise exploratória dos dados, foi realizada uma análise gráfica detalhada para visualizar os padrões identificados no *dataset*. Foram gerados *heatmaps* para destacar a correlação entre variáveis numéricas, gráficos de dispersão para compreender a relação entre quilometragem e preço, e gráficos de regressão para modelar tendências lineares e não lineares. A análise do modelo S-Curve revelou padrões de estabilização nos preços de veículos mais antigos, capturando de forma mais realista a desvalorização ao longo do tempo do que os modelos de regressão linear simples.

Através da análise gráfica, foram também identificadas diferenças significativas entre marcas e categorias de veículos, bem como a influência de fatores como o tamanho do motor e o tipo de combustível. Essas informações foram fundamentais para a implementação dos modelos de *machine learning,* permitindo um ajuste mais preciso das previsões de preços. Modelos como *Random Forest* foram utilizados para destacar as variáveis mais relevantes, enquanto a *técnica Recursive Feature Eliminatio*n (RFE) ajudou a otimizar a seleção de atributos essenciais para as previsões.

A combinação da análise estatística dos dados com técnicas gráficas e de *machine learning* permitiu obter uma visão abrangente do mercado de automóveis usados, fornecendo insights valiosos para a construção de um modelo preditivo robusto e funcional.

## 4.1. Atributos e estatística

O conjunto de dados analisado contém informações detalhadas sobre veículos usados, incluindo atributos essenciais para a avaliação do seu valor de mercado. Entre os dados registados, encontram-se o modelo do veículo, o ano de fabrico e o preço em libras esterlinas (£). Além disso, o *dataset* inclui características mecânicas, como o tipo de transmissão, distinguindo entre caixas de velocidades manuais e automáticas, bem como a quilometragem percorrida (em milhas) e o tipo de combustível utilizado (gasolina, gasóleo, entre outros).

Adicionalmente, são apresentados dados sobre o imposto rodoviário (£), um fator relevante no custo de posse do veículo, a eficiência de combustível, medida em milhas por galão (mpg), e a cilindrada do motor (em litros). O *dataset* também regista a marca do automóvel, permitindo uma segmentação por fabricantes.

Por fim, foi identificada uma coluna designada *tax*(£), que aparenta ser uma duplicação da coluna do imposto rodoviário e poderá necessitar de revisão ou eliminação no processo de tratamento dos dados.

**Ano de Fabrico (*year*):**

Média: 2017.09 - A maioria dos carros são modelos recentes.

Moda: 2019 - O ano mais comum nos dados.

Mínimo: 1970 | Máximo: 2020 - Existem alguns carros muito antigos na base.

**Preço (*price*) (em libras esterlinas - £):**

Média: £16,805 - O preço médio dos carros usados na amostra.

Moda: £9,995 - O preço mais frequente.

Mínimo: £450 | Máximo: £159,999 - Inclui desde carros muito baratos até veículos de luxo.

**Quilometragem (*mileage*) (em milhas):**

Média: 23,058 mi - Quilometragem média dos veículos listados.

Moda: 10 mi - Pode indicar registros incorretos ou veículos recém-saídos de fábrica.

Mínimo: 1 mi | Máximo: 323,000 mi - Variação extrema de quilometragem.

**Imposto Rodoviário (*tax*) (em libras esterlinas - £):**

Média: £120 - Imposto médio anual pago pelos veículos.

Moda: £145 - O valor mais frequente.

Mínimo: £0 | Máximo: £580 - Alguns carros não pagam imposto (provavelmente elétricos).

**Eficiência de Combustível (*mpg*) (milhas por galão):**

Média: 55.16 mpg - Consumo médio relativamente eficiente.

Moda: 60.1 mpg - O valor mais comum.

Mínimo: 0.3 mpg | Máximo: 470.8 mpg - Pode haver erros de entrada de dados.

**Tamanho do Motor (*engineSize*) (em litros):**

Média: 1.66 L - O tamanho médio dos motores.

Moda: 2.0 L - O tamanho mais comum.

Mínimo: 0.0 L | Máximo: 6.6 L - Há alguns registos incorretos com tamanho zero.

### 4.1.1. Interpretação geral

A análise dos dados revela que a maioria dos veículos incluídos no conjunto de dados foram fabricados entre 2016 e 2019, o que indica que o foco do *dataset* está em automóveis relativamente recentes. Observa-se uma grande variação de preços, abrangendo desde veículos mais acessíveis até modelos de luxo.

No que respeita à quilometragem, existem registos extremamente baixos, como 10 milhas, que podem corresponder a veículos praticamente novos ou a erros na introdução dos dados. A eficiência de combustível apresenta valores geralmente razoáveis, embora alguns registos pareçam inconsistentes ou irreais, exigindo uma análise mais aprofundada.

Relativamente ao tamanho do motor, os veículos com motores de 2.0L são os mais comuns, mas há uma amplitude significativa, com alguns casos extremos que chegam aos 6.6L, sugerindo a presença de modelos de alto desempenho ou veículos de nicho no conjunto de dados.

# 5. Análise gráfica

A análise gráfica do *dataset* é uma etapa essencial para compreender os padrões e relações entre as variáveis. Abaixo encontram-se as análises dos principais gráficos gerados a partir do *dataset* consolidado.

## 5.1. Heatmap

O heatmap de correlação foi utilizado para identificar relações entre variáveis numéricas, como o ano de fabrico, preço, quilometragem, consumo médio, taxa anual e tamanho do motor. Este gráfico, gerado com o código apresentado, permite visualizar de forma intuitiva a força e a direção das correlações. Valores próximos de 1 ou -1 indicam correlações fortes, enquanto valores próximos de 0 sugerem ausência de relação significativa.

A análise da Figura 2 – Heatmap revelou que o preço dos veículos apresenta uma correlação positiva moderada com o ano de fabrico, indicando que carros mais recentes tendem a ser mais valorizados. Por outro lado, uma correlação negativa significativa foi observada entre o preço e a quilometragem, confirmando que veículos com maior desgaste têm menor valor de mercado. A análise do *heatmap* ajuda a selecionar as variáveis mais relevantes para os modelos preditivos.

Comparando com o *heatmap* presente no documento “Used Car Price Prediction Using Machine Learning Techniques” [15] conseguimos perceber que de acordo com o heatmap, a variável alvo (preço) está negativamente correlacionada com “mpg em cidade”, “mpg em autoestrada” e positivamente correlacionada com “distância entre eixos” e “largura do carro”, “comprimento do carro”, etc [20].

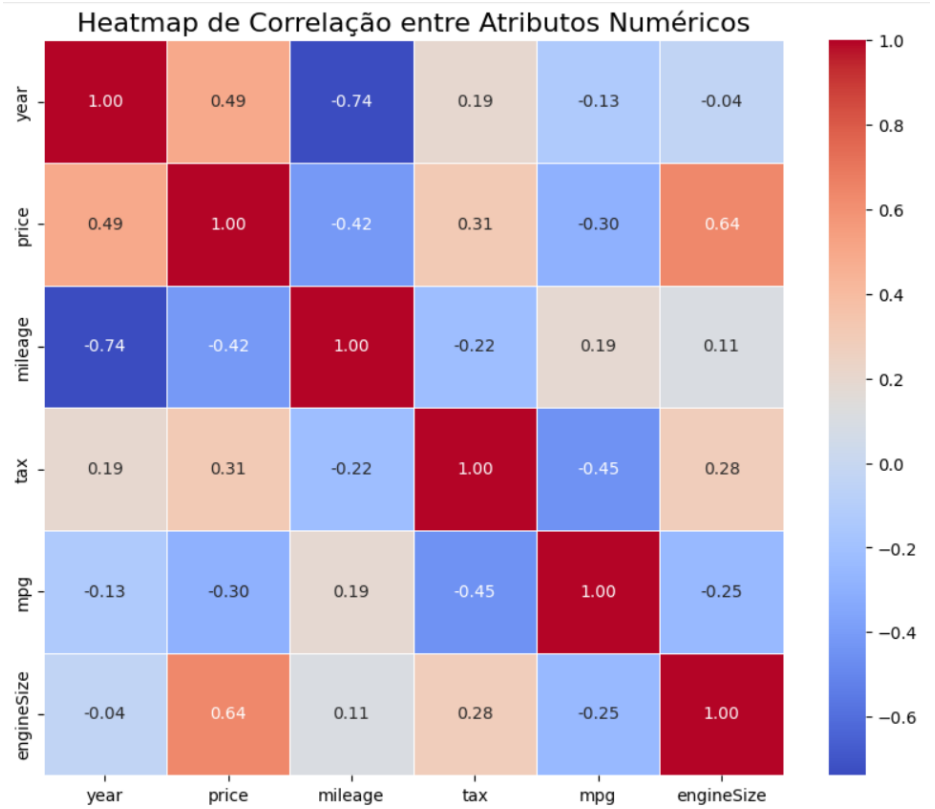


*Figura 2 – Heatmap*

A análise do *heatmap* é fundamental no presente estudo, permitindo identificar as variáveis mais relevantes na previsão do preço dos automóveis usados. Conforme evidenciado na investigação, este método possibilitou a seleção de fatores determinantes, como a quilometragem, o ano de fabrico e o tamanho do motor, contribuindo para a melhoria da precisão dos modelos preditivos aplicados.

Adicionalmente, a comparação com o estudo de M. S. Das et al. permitiu validar que a variável alvo (preço) apresenta correlações significativas com atributos como a eficiência de combustível e as dimensões do veículo. Esta análise reforça a pertinência da utilização do *heatmap* no presente trabalho, garantindo que os modelos de *machine learning* se baseiem em características com maior impacto na definição do valor dos automóveis.

## 5.2. Regressão Linear

O gráfico de regressão linear para a marca **Mercedes-Benz** utiliza o ano de fabrico como variável preditora e o preço como variável dependente. A linha de regressão ajustada demonstra a relação direta entre essas variáveis, evidenciando que veículos mais recentes da marca tendem a ter preços mais elevados.

A equação da regressão linear gerada pelo modelo quantifica essa relação, permitindo prever o preço aproximado de um veículo com base no seu ano. Apesar de a regressão linear ser útil para identificar tendências gerais, é importante reconhecer que este modelo assume uma relação linear perfeita, o que pode não captar todas as nuances do mercado automóvel.

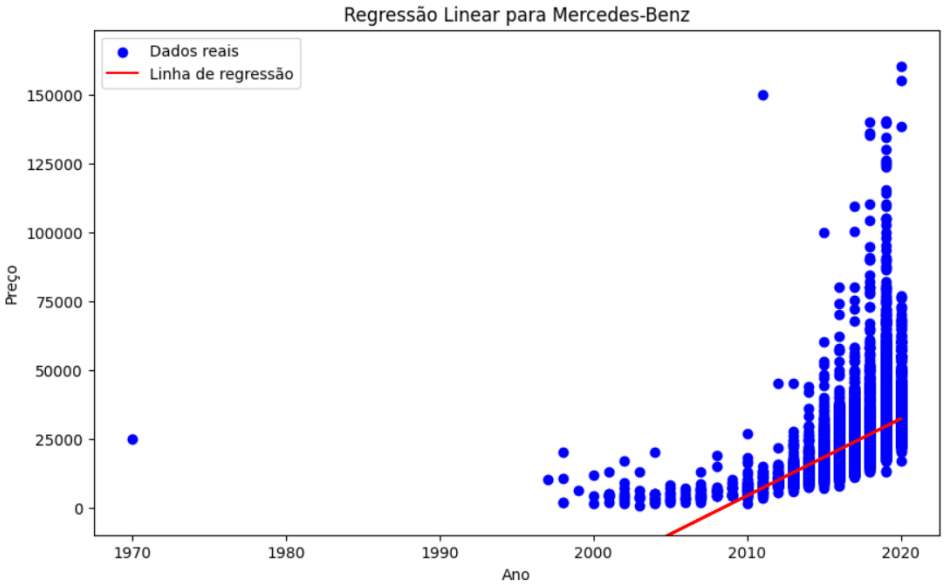


*Figura 3 - Regressão Linear*

## 5.3. S-curve

A análise do gráfico S-Curve para a marca **Mercedes-Benz** permite modelar a evolução dos preços de forma mais realista, capturando comportamentos não-lineares. A curva sigmoidal ajustada demonstra três fases distintas: uma fase inicial de crescimento lento nos preços, seguida por uma aceleração à medida que os veículos atingem anos recentes, e finalmente uma estabilização dos preços para modelos mais novos.

Este comportamento reflete a realidade do mercado, onde carros muito antigos têm preços baixos, os modelos mais recentes têm maior valorização, e os mais novos apresentam uma estabilização nos valores devido à falta de desvalorização significativa. Este modelo é particularmente útil para prever os limites superiores e inferiores do preço, ajustando-se melhor à realidade do que a regressão linear.

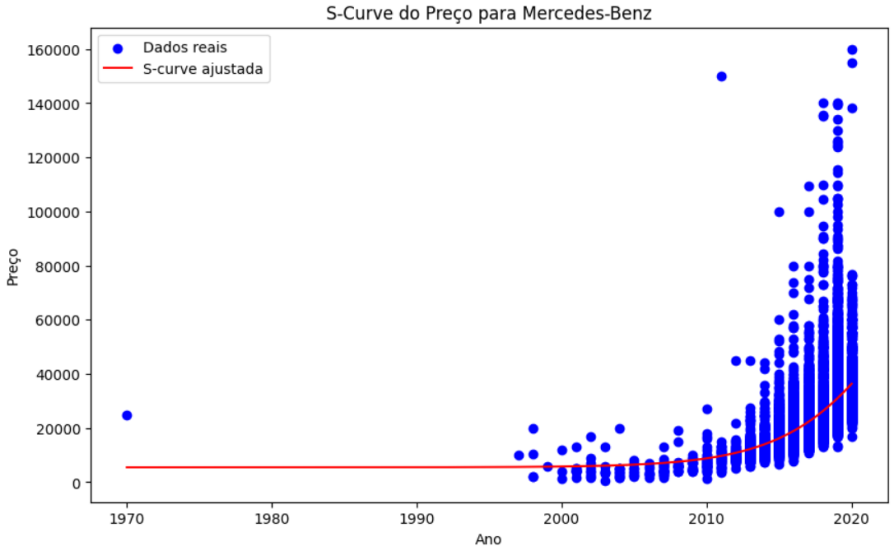


*Figura 4 - S-Curve*

## 5.4. Relações Visuais: Marca, Ano, Quilometragem, Tipo de Combustível e Tamanho do Motor

Foram criados cinco gráficos adicionais para analisar as relações entre variáveis-chave e o preço dos automóveis.

### 5.4.1. Marca vs. Preço

A análise revelou que marcas premium, como Mercedes-Benz e BMW, possuem preços médios significativamente mais altos em comparação com marcas generalistas, como Ford ou Skoda. Estas diferenças refletem o impacto da perceção de qualidade, confiabilidade e prestígio associados a cada marca.

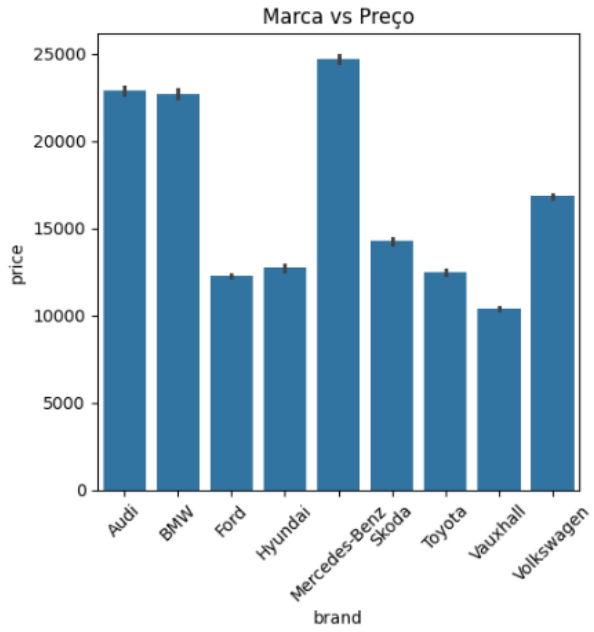


*Figura 5 - Gráfico de barras Marca vs Preço*

### 5.4.2. Ano vs. Preço

Existe uma relação clara entre o ano de fabrico e o preço. Carros mais recentes apresentam preços mais altos devido à menor desvalorização e ao facto de incorporarem tecnologias mais modernas. No entanto, anos muito antigos mostram uma estabilização dos preços, representando o valor residual.

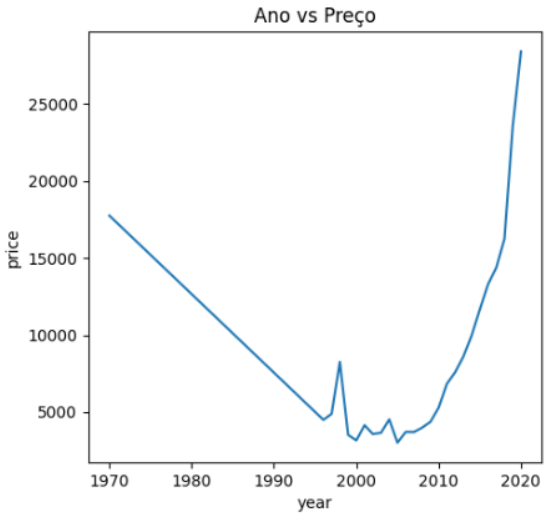


*Figura 6 - Gráfico de linha Ano vs Preço*

### 5.4.3. Quilometragem vs. Preço

Este gráfico confirma a correlação negativa entre quilometragem e preço. Carros com maior quilometragem são desvalorizados, pois representam maior desgaste e maior probabilidade de avarias. A dispersão dos dados também sugere que outros fatores, como marca e estado geral, influenciam a desvalorização.

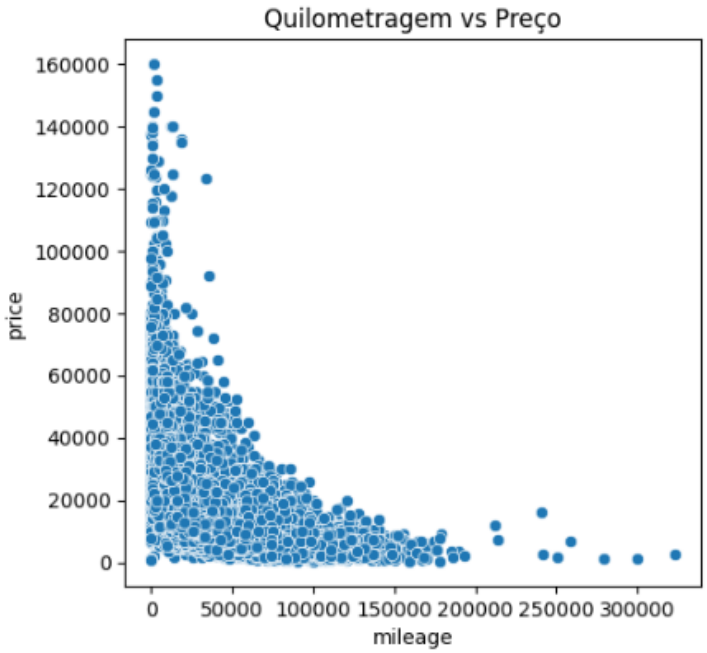


*Figura 7 - Gráfico de Dispersão Quilometragem vs Preço*

### 5.4.4. Tipo de Combustível vs. Preço

Veículos a diesel mostraram maior valorização em comparação com veículos a gasolina, refletindo a eficiência para longas distâncias. Por outro lado, veículos elétricos e híbridos estão em crescente valorização, especialmente em regiões urbanas, devido à procura por opções mais ecológicas.

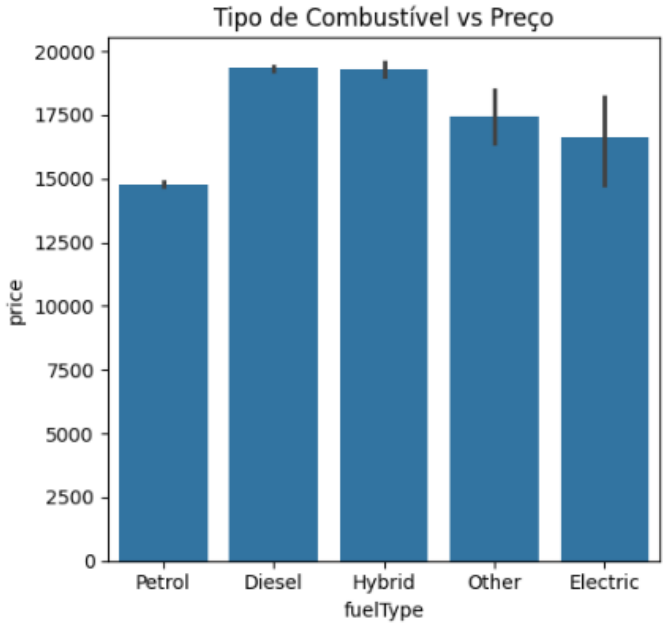


*Figura 8 - Gráfico de barras Tipo de Combustivel vs Preço*

### 5.4.5. Tamanho do Motor vs. Preço

Existe uma relação positiva entre o tamanho do motor e o preço, especialmente para carros de luxo. No entanto, motores muito grandes podem sofrer desvalorização em mercados onde a eficiência energética é uma preocupação crescente.

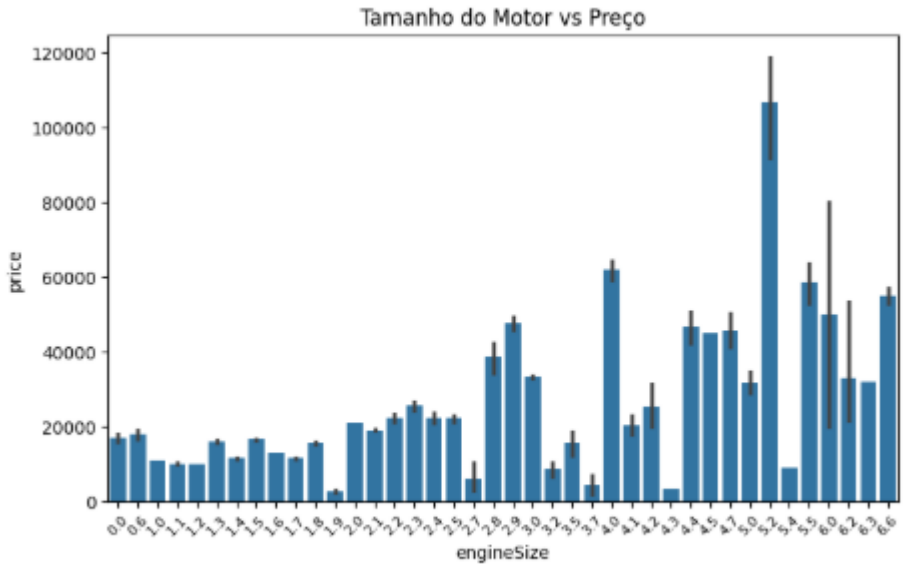


*Figura 9 – Gráfico de barras Tamanho do Motor vs Preço*

## 5.5. Implementação de algoritmos

O gráfico gerado apresenta a relação entre os preços reais e os preços previstos pelo modelo de regressão linear, permitindo avaliar a precisão das previsões realizadas. No gráfico, os pontos azuis representam os valores previstos em comparação com os valores reais, enquanto a linha vermelha tracejada representa a referência ideal, onde

as previsões seriam idênticas aos valores observados. A proximidade dos pontos em relação a essa linha indica a qualidade do modelo preditivo, sendo que uma dispersão elevada sugere imprecisão nas estimativas.

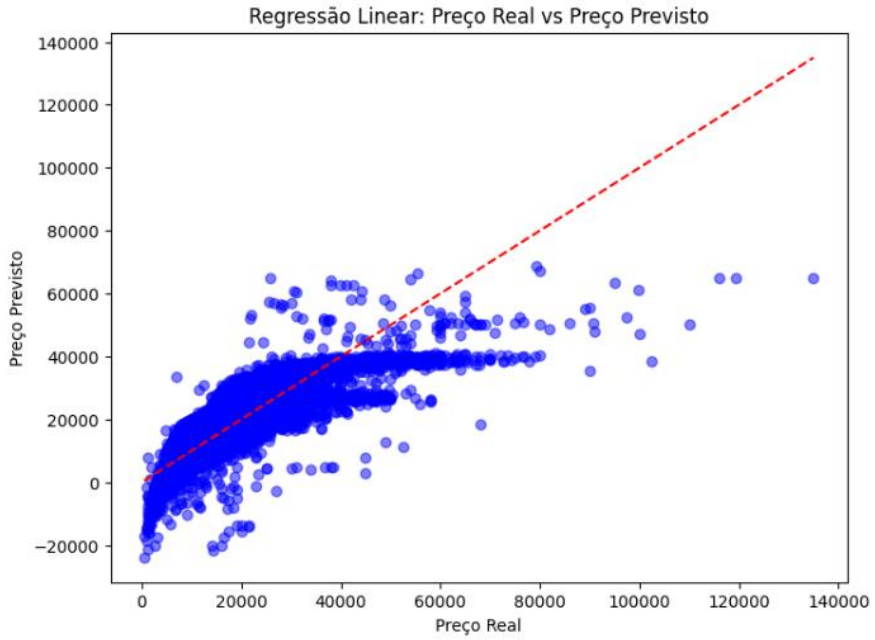


*Figura 10 - Regressão Linear: Preço Real vs Preço Previsto*

Caso os pontos estejam predominantemente acima ou abaixo da linha de referência, pode-se inferir a existência de um viés sistemático no modelo, resultando na subestimação ou superestimação dos preços. Além disso, a presença de padrões não lineares pode indicar que um modelo linear pode não ser a abordagem mais adequada, sendo recomendável a utilização de modelos não lineares, como *Random Forest* ou Redes Neurais, para melhorar a capacidade preditiva.

```
Características selecionadas: Index(['model', 'year', 'mileage', 'mpg', 'engineSize'], dtype='object'
Erro médio absoluto (MAE) com Regressão Linear: 3649.6942479697996
      Feature   Coefficient
4  engineSize  11825.514983
1        year   1583.181868
0       model      4.231789
2     mileage     -0.107557
3         mpg    -27.196275
```

*Figura 11 - Coeficiente da Regressão Linear*

A análise também inclui a métrica de erro médio absoluto (MAE), que quantifica a discrepância média entre os valores previstos e os valores reais. Um MAE elevado sugere que o modelo apresenta limitações na sua capacidade de previsão, enquanto um MAE reduzido indica um desempenho mais satisfatório. Com base nesses resultados, recomenda-se a análise da seleção de variáveis, a possível introdução de modelos mais complexos e a avaliação de eventuais valores atípicos que possam estar a influenciar negativamente o desempenho do modelo.

## 5.6. *Random Forest*

O modelo *Random Forest* apresentado permite analisar a influência das variáveis independentes na previsão do preço dos automóveis. Diferentemente da regressão

linear, onde cada variável tem um coeficiente que indica sua influência direta na variável dependente, o *Random Forest* avalia a importância relativa de cada variável no modelo preditivo.

```
Erro médio absoluto (MAE): 1150.9937293402922
        Feature  Importance
8    engineSize    0.461330
4       mileage    0.213116
2          year    0.143254
1         model    0.074371
7           mpg    0.051475
0         brand    0.027906
6           tax    0.012279
3  transmission    0.009589
5      fuelType    0.006680
```

*Figura 12 - Random Forest*

A tabela gerada pelo código exibe a importância de cada variável na previsão do preço, indicando quais características exercem maior impacto nas estimativas. Valores mais elevados de importância sugerem que a variável tem um peso significativo na tomada de decisão do modelo, enquanto valores mais baixos indicam uma contribuição menor.

Ao analisar essa tabela, é possível identificar os principais fatores que influenciam o preço dos automóveis. Por exemplo, espera-se que variáveis como "*year*" e "*mileage*" tenham alta relevância, uma vez que carros mais novos tendem a ser mais caros e quilometragens mais altas normalmente reduzem o valor do veículo. Outras variáveis, como "*engineSize*" e "*fuelType*", também podem ter influência considerável, dependendo das características do mercado analisado.

## 5.7. Wrapper (*Random Forest* e RFE)

O modelo *Random Forest* apresentado incorpora a técnica de *Recursive Feature Elimination* (RFE) para selecionar as variáveis mais relevantes na previsão do preço dos automóveis. O RFE é um método que elimina gradualmente as características menos importantes, resultando na seleção das variáveis que mais contribuem para a precisão do modelo.

```
Características selecionadas: Index(['model', 'year', 'mileage', 'mpg', 'engineSize'], dtype='object')
Erro médio absoluto (MAE) com as características selecionadas: 1222.5732809053754
      Feature  Importance
4  engineSize    0.469994
2     mileage    0.219816
1        year    0.146967
0       model    0.102889
3         mpg    0.060334
```

*Figura 13 - Wrapper (*Random Forest *e RFE)*

A tabela gerada exibe as cinco variáveis mais significativas identificadas pelo RFE. Essas variáveis foram escolhidas por apresentarem maior impacto na previsão do preço, eliminando aquelas que contribuem menos para a performance do modelo. Esse processo melhora a interpretabilidade do modelo e pode reduzir o risco de *overfitting*, tornando a previsão mais robusta.

Após a seleção das variáveis, o modelo *Random Forest* é treinado novamente, agora considerando apenas as características escolhidas. O erro médio absoluto (MAE) calculado reflete o desempenho do modelo após essa otimização. Caso o MAE seja reduzido em comparação com um modelo que utiliza todas as variáveis, isso indica que a eliminação de características irrelevantes ajudou a melhorar a precisão do modelo.

A tabela de importâncias das variáveis mostra a contribuição relativa de cada uma das características selecionadas. Quanto maior o valor da importância, maior o impacto dessa variável na determinação do preço dos automóveis. Variáveis como "*year*", "*mileage*" e "*engineSize*" tendem a ser fatores dominantes, pois influenciam diretamente o valor de mercado dos veículos.

# 6. Conclusão

O presente projeto explorou a análise preditiva de preços de automóveis usados, abrangendo desde os fundamentos teóricos e metodológicos até à identificação das variáveis mais influentes e dos desafios inerentes ao setor. Através de uma análise detalhada dos *datasets* disponíveis e da revisão do estado da arte, foram investigadas as principais abordagens utilizadas no mercado, incluindo técnicas estatísticas tradicionais e métodos avançados de aprendizagem automática.

Na fase inicial do estudo, procedeu-se a um levantamento criterioso dos fatores determinantes na precificação dos veículos, tais como a depreciação natural, a quilometragem, o estado geral do automóvel e outros elementos de impacto. Foram igualmente analisadas diversas técnicas preditivas, incluindo regressão linear, árvores de decisão, florestas aleatórias e redes neuronais, comparando o seu desempenho e aplicabilidade em cenários reais.

Com base nestas análises, delineou-se um plano para a segunda fase do projeto, que não se limitará ao desenvolvimento de uma aplicação funcional para previsão de preços, mas procurará, também, melhorar os resultados obtidos. Para tal, será fundamental a exploração de novos algoritmos, a experimentação com diferentes *datasets* e a fusão de bases de dados, bem como a adoção de metodologias complementares que possam otimizar a precisão dos modelos preditivos.

Deste modo, o presente projeto não se restringe à criação de uma ferramenta simples de previsão de preços, mas visa aprofundar a investigação na área da análise preditiva, contribuindo para o desenvolvimento de soluções mais robustas e eficazes. A continuidade deste trabalho será essencial para refinar os modelos propostos e garantir a escalabilidade e aplicabilidade da solução no contexto do mercado automóvel.

# Referências

[1] S. Sinha, R. Azim e S. Das, "Linear Regression on Car Price Prediction," 2020.

[2] Khan, M. Nasir e F. Naseer, "IoT based university garbage monitoring system for healthy environment for students," 2020.

[3] "Mobycar," 2021. [Online]. Available: https://www.mobycar.pt/quanto-um-carro-desvaloriza-por-ano-entenda-a-depreciacao/. [Acedido em 12 2024].

[4] T. consultas, "Trackcar," [Online]. Available: https://trakcar.com.br/desvalorizacao-de-um-veiculo/. [Acedido em 12 2024].

[5] "SOS Baterias," [Online]. Available: https://sosbateria.com.br/depreciacao-veiculo-principais-fatores/. [Acedido em 01 2025].

[6] G. Sheidt, "bsoft blog," 6 11 2024. [Online]. Available: https://blog.bsoft.com.br/calculo-de-depreciacao-de-veiculo. [Acedido em 1 2025].

[7] Jaque, "valornoticias," 17 05 2023. [Online]. Available: https://valornoticias.com/depreciacao-de-veiculos/. [Acedido em 12 2024].

[8] E. OLX, "OLX," 18 05 2018. [Online]. Available: https://dicas.olx.com.br/autos/para-voce/depreciacao-de-carros. [Acedido em 12 2024].

[9] F. Salim e N. A. Abu, "An S-curve Model on the Maximum Predictive Pricing of Used Cars," 2020.

[10] S. Mamipour e F. V. Jezeie, "Non-linear relationships among oil price , gold price and stock market returns in Iran : A multivariate regime-switching approach," 2015.

[11] C. Ferrari, M. Marchese e A. Tei, "Shipbuilding and economic cycles: a non-linear econometric approach," 2018.

[12] G. Aydin, "Forecasting natural gas production using various regression models," 2015.

[13] J. R. S. Cristóbal, "The S-curve envelope as a tool for monitoring and control of projects," 2017.

[14] P. R. Mahalingam e S. Vivek, “Predicting financial savings decisions using sigmoid function and information gain ratio,” 2016.

[15] E. Gegic, B. Isakovic, D. Keco, Z. Masetic e J. Kevric, “Car Price Prediction using Machine Learning Techniques,” 2021.

[16] A. Pandey, V. Rastogi e S. Singh, “Car’s Selling Price Prediction using Random Forest Machine Learning Algorithm,” 2020.

[17] D. Štifanić, J. Musulin, A. Miočević, S. B. Šegota, R. Šubić e Z. Car, “Impact of covid-19 on forecasting stock prices: an integration of stationary wavelet transform and bidirectional long short-term memory,” 2020.

[18] E. Gegic, B. Isakovic, D. Keco, Z. Masetic e J. Kevric, “Car Price Prediction using Machine Learning Techniques,” 2019.

[19] K. Noor e S. Jan, “Vehicle Price Prediction System using Machine Learning Techniques,” 2017.

[20] M. S. Das, M. A. Laha, M. A. Jena e M. P. Samal, “Used Car Price Prediction Using Machine Learning Techniques,” 2021.

[21] P. P. Shinde e D. S. Shah, “A Review of Machine Learning and Deep,” 2018.

[22] L. C. D. S. Lucas, “ÁRVORES,FLORESTAS E SUA FUNÇÃO COMO PREDITORES: UMA APLICAÇÃO NA AVALIAÇÃO DO GRAU DE MATURIDADE DE EMPRESAS,” 2011.