Revelando Padrões na Avaliação de Carros: Uma Exploração de Mineração de Dados

Diogo Porta-Nova
Universidade de Évora
m53871@alunos.uevora.pt

1 Introdução

1.1 Revisão da Literatura

A avaliação de carros é uma área multifacetada que tem sido abordada de várias maneiras na literatura, envolvendo pesquisas com utilizadores, revisões de especialistas e até a aplicação de modelos de aprendizagem de máquina. Esta revisão terá como foco resumir a literatura anteriormente já existente, com destaque para as diferentes abordagens e com identificação de lacunas que a análise proposta pretende preencher.

- **Pesquisas com Utilizadores** : Estudos que recorrem a pesquisas com utilizadores oferecem uma perspectiva valiosa sobre as preferências e experiências dos consumidores em relação a carros. Estas pesquisas muitas vezes capturam aspetos subjectivos, como design, conforto e até preenchem requisitos de preferências individuais.

- **Revisões de Especialistas** : São frequentemente publicadas por revistas automóveis de renome, oferecendo avaliações críticas com base em testes práticos, nos quais se avaliam desempenho, segurança e até inovações tecnológicas.

- **Modelos de Aprendizagem de Máquina na Avaliação de Carros** : A aplicação destes modelos tem-se destacado como uma inovadora abordagem na avaliação de carros. O conjunto de dados em estudo proposto, Car Evaluation Database, que se propõe a treinar modelos que previssem a aceitabilidade de carros em que tivessem por base atributos como o preço, a manutenção, as características técnicas e até a segurança. Existem também estudos, como o de Zupan et al. (1997), no qual ficou demonstrada a eficácia do HINT (Hierarchy INduction Tool), usado na reconstrução de modelos hierárquicos para a avaliação de carros.

- **Descobertas e Metodologias Anteriores** : Estudos anteriores conseguiram revelar certos padrões interessantes na preferência do consumidor, com destaque para a importância de fatores como o preço, a segurança e o conforto. Abordagens de aprendizagem de máquina, em especial aquelas que incorporam estruturas com hierarquias, têm mostrado promessa no processo de modelagem de preferências de modo mais preciso e rigoroso. A análise de conjuntos de dados em específico, como é o caso do estudo proposto e mencionado do Car Evaluation Database, proporcionou insights valiosos sobre estruturas subjacentes dos modelos de precisão.

- **Estado Atual do Campo de Avaliação de Carros** : Este campo demonstra estar em constante evolução, com a incorporação de forma crescente de tecnologias avançadas, como veículos autónomos e sistemas de assitência ao condutor. Aparte disto, a integração de abordagens de aprendizagem de máquina têm proporcionado uma compreensão mais profunda das preferências dos consumidores. No entanto, ainda existem lacunas notórias, com especial foco na interpretabilidade dos modelos de aprendizagem de máquina e a consideração adequada de fatores éticos e de segurança.

- **Lacunas e Limitações na Pesquisa Existente** : Apesar dos avanços, a literatura existente atualmente sobre a avaliação de carros ainda enfrenta bastantes desafios. A interpretabilidade de modelos complexos continua a ser uma preocupação, especialmente quando procuramos entender de que forma as decisões são tomadas. Além disto, as pesquisas muitas vezes carecem de uma abordagem mais holística (abordagem na íntegra, como um todo) que integre tanto as preferências subjectivas dos consumidores quanto as métricas objectivas de desempenho do veículo. De realçar outro aspeto ético ainda subexplorado, como a segurança cibernética em veículos conectados.

Esta revisão procura consolidar o conhecimento existente sobre a avaliação de carros, destacando as diversas abordagens já empregues. Ao se identificarem lacunas e limitações, a análise proposta visa contribuir para o desenvolvimento futuro deste campo, incentivando uma abordagem integrada que tenha em consideração tanto a perspectiva do consumidor quanto os avanços tecnológicos emergentes.

2 Dados

2.1 Apresentação do Conjunto de Dados

- **Detalhes do Conjunto de Dados** : O conjunto de dados em estudo apresenta um total de 1728 instâncias, com 6 atributos (buying, maint, doors, persons, lug_boot e safety). A Car Evaluation Database é derivada de um modelo de decisão hierárquico originalmente desenvolvido para a demonstração de DEX, tendo como autores M. Bohanec e V.Rajkovic, que decidiram criar um sistema de Expertise para tomadas de decisões. Este modelo avalia carros de acordo com os atributos descritos anteriormente.

- **Relevância do Conjunto de Dados** : A relevância deste conjunto de dados reside na sua origem num modelo teórico com boa fundamentação e na sua estrutura, que preserva as relações entre os atributos. Esta estrutura única ofereca oportunidades valiosas para testar métodos de indução construtiva e descoberta de estruturas, especialmente se falarmos de contextos de avaliação de carros onde as relações de hierarquia são fundamentais. O conjunto de dados está assim disponível desde junho de 1997 e tem sido amplamente usado em estudos de aprendizagem de máquina, incluindo a avaliação do HINT (Hierarchy INduction Tool) e comparações com métodos como C4.5, como apresentado em conferências como o ICML-97 em Nashville, TN.

- **Base de Dados de Avalição de Carros** : Deriva de um modelo de decisão hierárquica desenvolvido originalmente para a avaliação de carros. Este modelo foi proposto por M.Bohanec e V.Rajkovic, apresentado na conferência "Knowledge acquisition and explanation for multi-attribute decision making"durante o 8º Workshop Internacional sobre Sistemas Especialistas e suas Aplicações, realizado em Avignon, França, em 1988. A estrutura original do modelo hierárquico é composta por conceitos inter-relacionados, iniciando com a avaliação geral de Carros (CAR) e descendendo para conceitos intermediários (PRICE, TECH, CONFORT) e atributos específicos (buying, maint, doors, persons, lug_boot, safety). Esta base de dados foi construída posteriormente a partir desse modelo, mantendo os atributos-chave enquanto removia a estrutura hierárquica original. Desta forma, o conjunto de dados relaciona diretamente o conceito de CAR aos 6 atributos de entrada: buying, maint, doors, persons, lug_boot, safety. A remoção da estrutura hierárquica oferece uma representação mais direta, tornando-a mais adequada para a aplicação de técnicas de mineração de dados.

- **Importância da Estrutura Hierárquica** : Fundamental para compreender nuances envolvidas nas decisões de avaliação de carros. A preservação desta estrutura hierárquica na origem do conjunto de dados reflete, e muito bem, a organização lógica das decisões de avaliação, fornecendo uma representação rica e detalhada das relações entre os diferentes atributos. Por exemplo, o conceito de PRICE encapsula aspectos relacionados ao custo, enquanto COMFORT engloba detalhes sobre o conforto do veículo. Esta estrutura permite uma análise mais granular e contextualizada, resultando numa representação mais fiel dos critérios considerados numa avaliação de carros. Com isto são apresentadas diversas vantagens tais como a Interdepência realista entre atributos ao manter as relações contextuais entre estes, a Interpretabilidade que fica facilitada ao se manterem as

relações entre os atributos e os Testes de Métodos Específicos que se tornam particularmente úteis uma vez que os dados têm uma organização subjacente bem definida.

- **Introdução e Descrição** : Conjunto de dados composto por 1728 instâncias que cobre de forma abrangente o espaço de atributos e garante assim uma representação diversificada de cenários de avaliação de carros. Os 6 atributos-chave são então os seguintes:

  1. **buying** : Referente ao preço de compra do carro, apresentando categorias como 'v-high', 'high', 'med' e 'low'.

Tabela 1: Contagens - Buying

	Count	**Percentage**
vhigh	432	25.0%
high	432	25.0%
med	432	25.0%
low	432	25.0%

  2. **maint** : Indica o custo de manutenção do carro, apresentando categorias como 'v-high', 'high', 'med' e 'low'.

Tabela 2: Contagens - Maint

	Count	**Percentage**
vhigh	432	25.0%
high	432	25.0%
med	432	25.0%
low	432	25.0%

  3. **doors** : Representa o número de portas do carro, apresentando categorias como '2', '3', '4' e '5-more'.

Tabela 3: Contagens - Doors

	Count	**Percentage**
2	432	25.0%
3	432	25.0%
4	432	25.0%
5more	432	25.0%

4. **persons** : Representa a capacidade de passageiros do carro, apresentando categorias como '2', '4' e 'more'.

Tabela 4: Contagens - Persons

	Count	Percentage
2	576	33.333%
4	576	33.333%
more	576	33.333%

5. **lug_boot** : Descreve o tamanho da bagageira do carro, apresentando categorias como 'small', 'med' e 'big'.

Tabela 5: Contagens - Lug_Boot

	Count	Percentage
small	576	33.333%
med	576	33.333%
big	576	33.333%

6. **safety** : Avalia a segurança estimada do carro, apresentando categorias como 'low', 'med' e 'high'.

Tabela 6: Contagens - Safety

	Count	Percentage
low	576	33.333%
med	576	33.333%
high	576	33.333%

Estes atributos abrangem aspectos cruciais na avaliação de carros, desde considerações financeiras, passando por características práticas como o número de portas e espaço da bagageira, até à segurança percecionada.

Com recurso a esta introdução e descrição, procuramos proporcionar uma visão inicial clara do conjunto de dados, por forma a preparar o terreno para análises mais profundadas e o entendimento do impacto potencial desses atributos na aceitabilidade geral de carros.

class : Define o grau de aceitabilidade do carro, apresentando estados como 'unacc', 'acc', 'good' e 'vgood'.

Tabela 7: Contagens - Class

	Count	**Percentage**
unacc	1210	70.023%
acc	384	22.222%
good	69	3.993%
vgood	65	3.762%

- **Relevância para o Problema Escolhido** : Intrinsecamente relevante para o problema abordado no estudo em causa, que se concentra na compreensão e predição da aceitabilidade de carros. Cada atributo do conjunto de dados, desde o preço de compra até à avaliação de segurança, está alinhado de forma direta com os aspetos-chave considerados pelos consumidores ao tomar decisões de compra de automóveis. Ao se analisar a estrutura hierárquica do modelo de decisão original, é percetível que os atributos presentes na base de dados estão entrelaçados de forma profunda com as diferentes camadas conceituais, proporcionando assim uma representação abrangente dos critérios considerados na avaliação de carros. A relação direta entre os atributos e as categorias do problema, como 'unacc'(inaceitável), 'acc'(aceitável), 'good'(bom) e 'v-good'(muito bom), consegue demonstrar a pertinência do conjunto de dados para a tarefa de avaliação de carros.

- **Tamanho, Estrutura e Peculiaridades** : Com um total de 1728 instâncias, esta base de dados é consideravelmente grande, proporcionando uma base estatisticamente significativa para análises detalhadas. A presença abrangente de instâncias no espaço de atributos garante portanto a representatividade dos cenários de avaliação de carros do mundo real. Este conjunto de dados apresenta uma estrutura composta por 6 atributos principais, nos quais cada um contribui de maneira única para a avaliação global. De realçar que a ausência de dados em faltas simplifica a análise e facilita a aplicação de técnicas de mineração de dados e de modelagem. Além disto, a presença de classes desbalanceadas, como está em evidência pela distribuição de classes, adiciona então uma camada adicional de complexidade, que torna o conjunto de dados adequado para abordagens de questões relacionadas com a desigualdade na aceitabilidade de carros. Estas peculiaridades, em conjunto com a estrutura hierárquica única, fazem desta base de dados de avaliação de carros uma escolha robusta e relevante para a exploração de diversos aspectos da avaliação de carros em estudo.

2.2 Referências Relacionadas

Durante a execução deste estudo, consultei e tive por base várias referências que achei relevantes e que contribuíram para uma compreensão mais aprofundada do conjunto de dados e das abordagens de modelagem usadas. Algumas das referências chave são:

- **Car Evaluation Database (UCI Machine Learning Repository)** : Este conjunto de dados serviu como principal base de estudo. Existem evidências de estudos anteriores que recorreram a este conjunto de dados e que exploraram a aplicação de algoritmos de aprendizagem de máquina com intuito de preverem a aceitabilidade de carros com base em diferentes características.

- **Machine Learning: A Probabilistic Perspective - Kevin P. Murphy** : Obra que fornece base teórica sólida para os algoritmos de aprendizagem de máquina, nos quais se incluem Decision

Trees, Random Forests e Neural Networks. A abordagem deste autor, com base em probabilidades, ajudou imenso na compreensão de métodos usados.

- **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow - Aurélien Géron** : Livro prático que ofereceu insights extremamente valiosos sobre a implementação de algoritmos em machine learning, com destaque para as melhores práticas, técnicas de pré-processamento e até avaliação de modelos. É presente neste livro uma seção sobre métodos de ensemble que foi de tal forma útil que me ajudou a compreender o Random Forest.

- **Artigos Científicos Relacionados** : Consultei também diversos artigos de origem científica nos quais foram aplicados algoritmos semelhantes em conjuntos de dados relacionados com a avaliação de carros. Estes estudos forneceram insights acerca de abordagens específicas, desafios decorrentes e até medidas de desempenho de relevância.

Ao comparar o meu trabalho com estas referências, consigo destacar a inclusão do algoritmo SVM (Support Vector Machine) como uma extensão significativa. A análise comparativa entre diferentes algoritmos, incluindo SVM, Decision Trees, Random Forest e Neural Networks, oferece uma visão mais abrangente das capacidades e limitações de cada um dos métodos. Esta abordagem diferenciada busca preencher lacunas que tenham sido identificadas em estudos passados e que permita contribuir para uma compreensão mais completa da aplicação de machine learning na avaliação de carros.

3 Algoritmos / Soluções Propostas

3.1 Definição da Abordagem

A minha principal abordagem será a classificação, tendo em conta que o problema em questão envolve a determinação da aceitabilidade de carros com base em diferentes características. A classificação é a escolha mais apropriada para a atribuição de uma classe específica a cada instância de carro, como ”aceitável”, ”inaceitável”, ”bom”ou até ”muito bom”. A justificação para esta escolha reside no facto de que este problema envolve a categorização de dados, algo que tem perfeito alinhamento com as diversas técnicas de classificação. No que toca a ferramentas de software, optei por recorrer ao uso de bibliotecas e frameworks populares em Python, tais como **scikit-learn, pandas, matplotlib, seaborn, entre outra**s. **Scikit-learn** oferece uma gama bastante ampla de algoritmos de machine learning e ferramentas de avaliação de desempenho, o **pandas** é uma poderosa biblioteca de manipulação e análise de dados que fornece estruturas de dados flexíveis como os DataFrames, o **matplotlib** é uma biblioteca de visualização de dados que oferece uma gama bastante ampla de funcionalidades que permitem a criação de gráficos 2D, gráficos de dispersão, histogramas, barras, linhas, entre outros sendo frequentemente usado para gerar gráficos estáticos em relatórios, apresentações e ambientes que peçam detalhadas personalizações e o **seaborn** que é aplicado sobre o **matplotlib** e permite ”embelezar”gráficos e tudo o que for criado sendo este mais recorrente o seu uso para visualizações estatísticas mais avançadas e esteticamente mais agradáveis.

3.2 Avaliação de Desempenho

Neste ponto, considerei o uso de uma matriz de confusão e índices de desempenho como a **precision, o recall, o F1-score e área sob a curva ROC (AUC-ROC)**. A matriz de confusão irá fornecer uma visão mais detalhada das previsões do modelo em relação às classes reais. Além disto, ao serem consideradas as características desbalanceadas do conjunto de dados, o recurso ao uso de uma matriz de custo pode

ser algo a ser explorado para a ponderação e descoberta de diferentes tipos de erros. Os objectivos específicos em termos de índices de desempenho irão sofrer ajustes em conformidade com as necessidades do problema em questão, como por exemplo, fornecendo mais ênfase à precisão caso seja crucial e crítico evitar os falsos positivos.

3.3 Seleção de Algoritmos

A seleção de algoritmos será diversificada para se poderem explorar diferentes abordagens. Pelo menos 4 tipos de algoritmos foram escolhidos:

- **Decision Trees** : O seu uso deriva da Interpretabilidade e capacidade de lidar com variáveis categóricas.
- **Random Forest** : O seu uso deriva de uma abordagem de ensemble que combina diversas Decision Trees com o intuito de melhorar o desempenho e reduzir o overfitting.
- **Neural Networks** : O seu uso deriva da complexidade que o problema abordado pode acarretar visto que uma rede pode capturar relações não lineares entre características.
- **SVM - Support Vector Machine** : O seu uso deriva de uma abordagem eficaz para problemas de classificação, especiamlmente se estivermos a tratar dados e classes não lineares.

3.4 Análise de Algoritmos

Cada algoritmo escolhido será analisado de forma cuidadosa:

Decision Trees :

- **Pontos Fortes** : Interpretabilidade e lida bem com dados categóricos.
- **Pontos Fracos** : Suscetibilidade a overfitting.

Random Forest :

- **Pontos Fortes** : Redução de overfitting e apresenta alta precisão.
- **Pontos Fracos** : Menor interpretabilidade em comparação com Decision Tree única.

Neural Networks :

- **Pontos Fortes** : Captura relações complexas e consegue lidar com grandes conjuntos de dados
- **Pontos Fracos** : Requer mais dados para treino e computacionalmente pode ser mais intensivo e pesado.

SVM - Support Vector Machine :

- **Pontos Fortes** : Eficácia em espaços de alta dimensão e apresenta versatilidade para diferentes funções de kernel.
- **Pontos Fracos** : Sensibilidade à escolha do kernel e requerimento de tuning de parâmetros.

Qualquer ausência de trabalhos anteriores que utilizassem a mesma abordagem será devidamente reconhecida, destacando sempre a originalidade do estudo. O uso de **SVM** vem adicionar uma dimensão adicional à análise, permitindo assim proceder com uma avaliação de como esse método se compara aos restantes.

4 Resultados

4.1 Experimentos

- **Estatísticas Descritivas** : Observei uma distribuição equilibrada nas categorias, indicando que o conjunto de dados é representativo em termos de diferentes características dos carros. A análise percentual por categoria fornece insights sobre a proporção de cada classe em relação ao total.

Tabela 8: Estatísticas da Base de Dados

	buying	**maint**	**doors**	**persons**	**lug_boot**	**safety**	**class**
count	1728	1728	1728	1728	1728	1728	1728
unique	4	4	4	3	3	3	4
top	vhigh	vhigh	2	2	small	low	unacc
freq	432	432	432	576	576	576	1210

- **Análise das Variáveis** : A análise das variáveis em relação à variável alvo (class) revela padrões importantes:

1. **buying x class** : A associação entre o preço de compra e a aceitabilidade do carro destaca que carros mais acessíveis têm maior probabilidade de serema aceites.

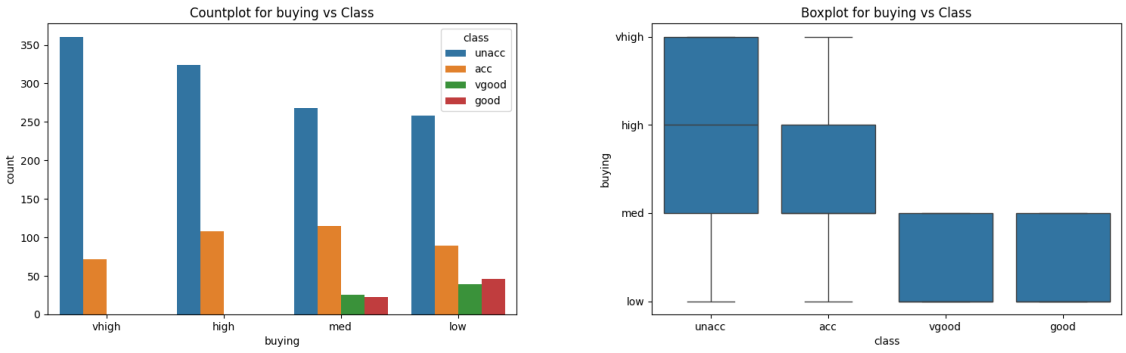

Figura 1: Buying X Class

2. **maint x class** : A relação entre os custos de manutenção e a aceitabilidade do carro mostra que custos mais baixos estão associados a carros aceites.

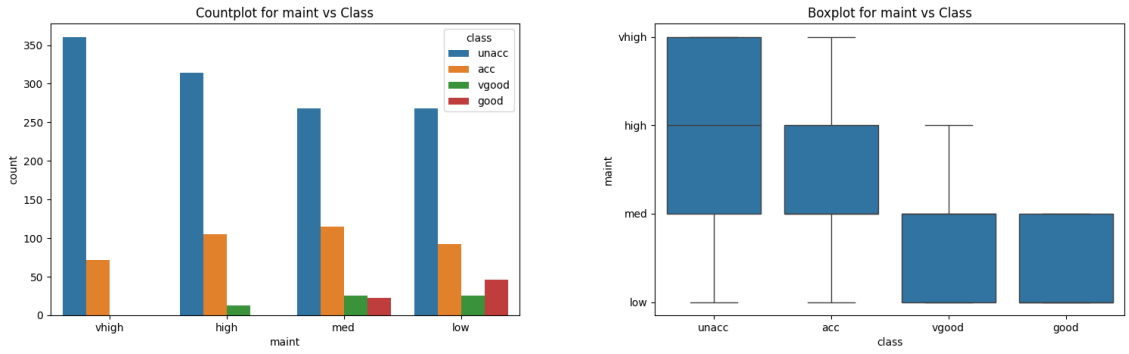

Figura 2: Maint X Class

3. **doors x class** : O número de portas influencia a aceitabilidade do carro, com 2 ou 4 portas a serem as mais aceitáveis.

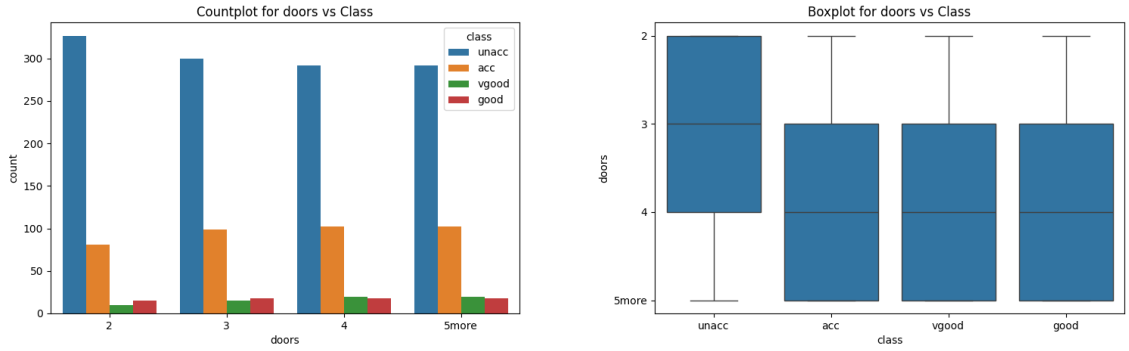

Figura 3: Doors X Class

4. **persons x class** : A capacidade de passageiros afeta a aceitabilidade, indicando que carros com capacidade para 2 ou 4 pessoas são mais aceites.

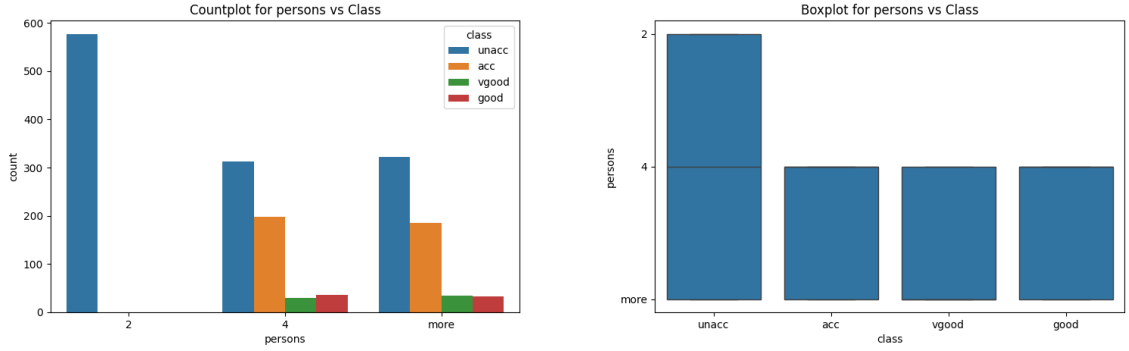

Figura 4: Persons X Class

5. **lug_boot x class** : O tamanho da bagageira impacta na aceitabilidade, sendo carros com bagageiras grandes mais aceites.

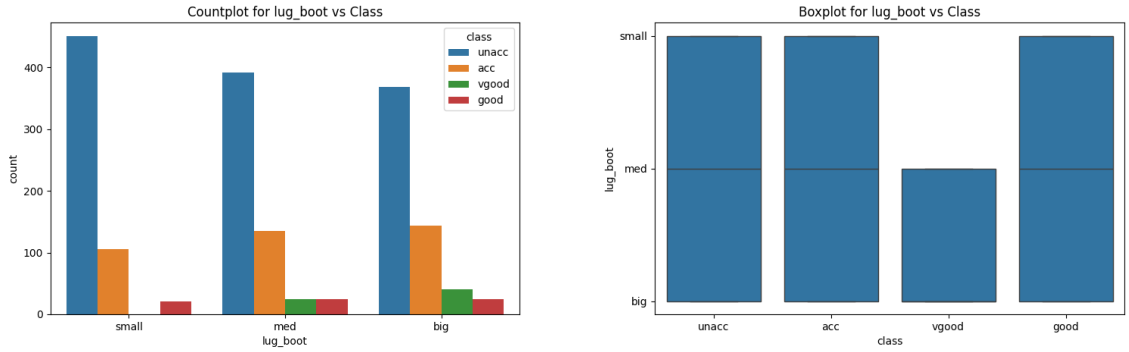

Figura 5: Lug_Boot vs Class

6. **safety x class** : A segurança é um factor crucial na aceitabilidade do carro, com carros mais seguros sendo mais aceites.

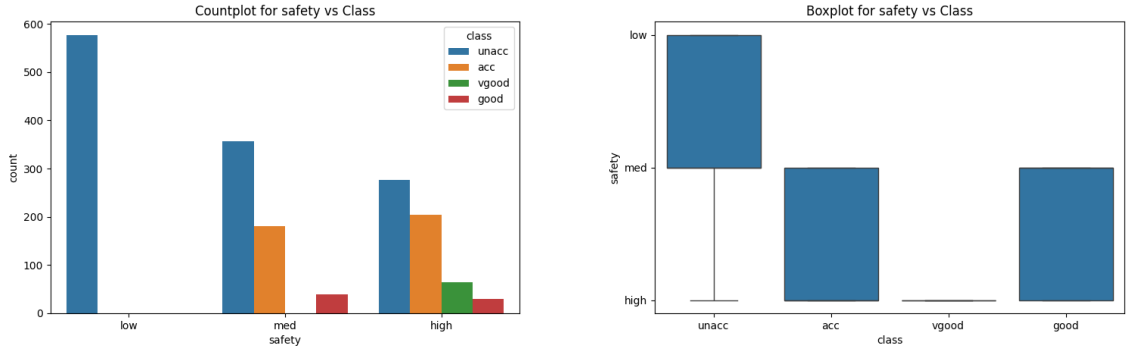

Figura 6: Safety vs Class

- **Análise das Cross-Tabulações** : A análise individual das cross-tabulações reforça as associações identificadas. Os testes qui-quadrado com valores de p próximos de zero reforça a ideia de que há uma dependência significativa entre quase todos atributos e a class atribuída a um carro. Todos os atributos, à excepção do **doors** (número de portas de um carro), apresentam dependência significativa com a **class** (estado de um carro). Quanto a dependência significativa entre os diversos atributos, é notória uma clara independência dos atributos entre si, visto que não há evidências de dependência significativa.

  Em resumo, a dependência significativa entre as variáveis indica que há uma relação estatisticamente significativa, enquanto a independência sugere falta de relação significativa. Essas conclusões são baseadas nos valores de qui-quadrado (Chi2) e p-values obtidos nos testes do qui-quadrado.

- **Seleção de Atributos e Pré-processamento**: No processo de preparação dos dados, optei por criar um conjunto de dados organizado, destacando atributos específicos relacionados com a avaliação de carros. A aplicação de técnicas de seleção de atributos, remoção de outliers e tratamento de

valores ausentes foram essenciais. A percentagem de valores únicos em cada coluna foi também calculada, fornecendo uma compreensão detalhada da distribuição dos dados.

- **Divisão do Conjunto de Dados** : O conjunto de dados foi dividido em conjuntos de treino e teste usando a função '**train-test-split**' do scikit-learn. A escolha de proporções visa garantir uma avaliação robusta do modelo em ambientes variados.

- **Medidas de Desempenho** : A avaliação do desempenho do modelo foi realizada utilizando quatro modelos diferentes: **Random Forest, SVM, Neural Networks (MLP) e Decision Trees**. A métrica de acurácia ('**accuracy**') foi empregue para avaliação. Visualizações como gráficos de contagem e boxplots foram geradas para melhor compreensão da distribuição dos dados e detecção de padrões. Além disso, métricas adicionais como **precision, recall, F1-score e AUC-ROC** foram aplicadas para fornecer uma compreensão mais abrangente do desempenho de cada modelo. Foram também analisadas matrizes de confusão para verificação do desempenho dos modelos de classificação apresentados acima.

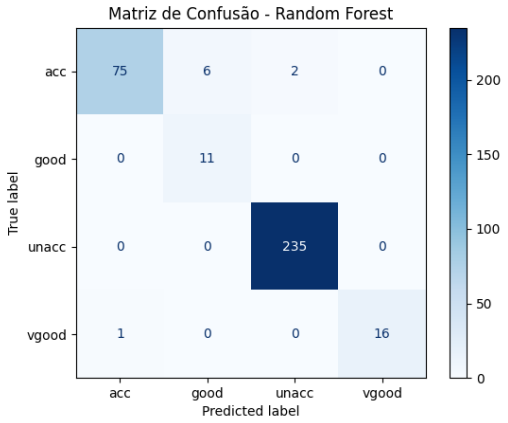

Figura 7: Confusion Matrix - Random Forest

	precision	recall	f1-score	support
acc	0.99	0.9	0.94	83
good	0.65	1.0	0.79	11
unacc	0.99	1.0	1.0	235
vgood	1.0	0.94	0.97	17
accuracy			0.97	346
macro avg	0.91	0.96	0.92	346
weighted avg	0.98	0.97	0.98	346

Figura 8: Report Classificação - Random Forest

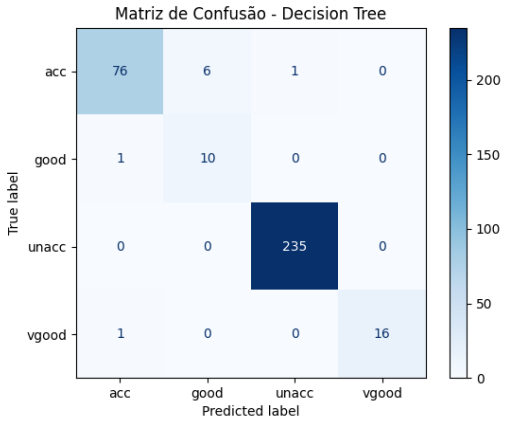

Figura 9: Confusion Matrix - Decision Tree

	precision	recall	f1-score	support
acc	0.97	0.92	0.94	83
good	0.62	0.91	0.74	11
unacc	1.0	1.0	1.0	235
vgood	1.0	0.94	0.97	17
accuracy			0.97	346
macro avg	0.9	0.94	0.91	346
weighted avg	0.98	0.97	0.98	346

Figura 10: Report Classificação - Decision Tree

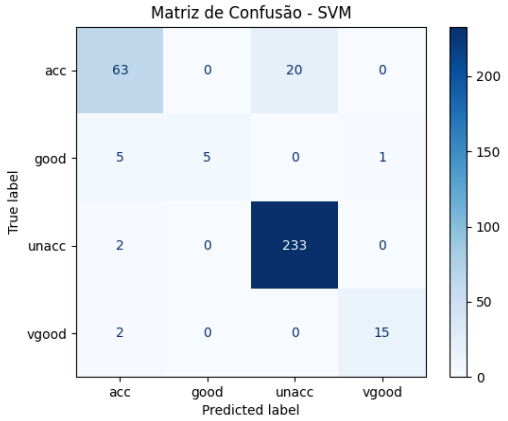

Figura 11: Confusion Matrix - SVM

	precision	recall	f1-score	support
acc	0.88	0.76	0.81	83
good	1.0	0.45	0.62	11
unacc	0.92	0.99	0.95	235
vgood	0.94	0.88	0.91	17
accuracy			0.91	346
macro avg	0.93	0.77	0.83	346
weighted avg	0.91	0.91	0.91	346

Figura 12: Report Classificação - SVM

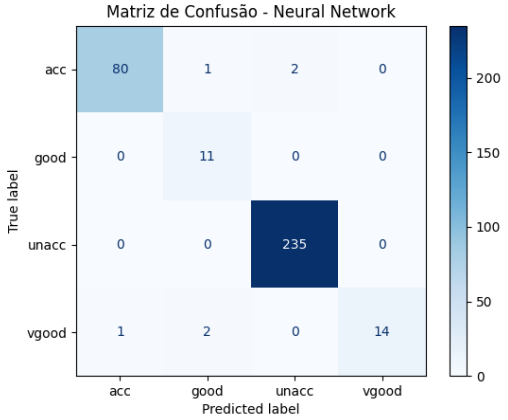

Figura 13: Confusion Matrix - Neural Networks

	precision	recall	f1-score	support
acc	0.99	0.96	0.98	83
good	0.79	1.0	0.88	11
unacc	0.99	1.0	1.0	235
vgood	1.0	0.82	0.9	17
accuracy			0.98	346
macro avg	0.94	0.95	0.94	346
weighted avg	0.98	0.98	0.98	346

Figura 14: Report Classificação - Neural Networks

- **Análise Geral** : Destaca-se a relevância de cada característica na previsão da aceitabilidade do carro. A distribuição equilibrada das classes e as associações identificadas indicam que todas as características analisadas podem ser preditoras significativas em modelos de aprendizagem de máquina. A análise de **recall** e **F1-score** revelaram alguns pontos de atenção em classes específicas. No **Random Forest** a classe 'unacc' (classe 3) apresenta um desempenho notável, com 235 verdadeiros positivos. No **SVM**, o modelo enfrenta desafios na classificação da classe 'good' (classe 2) com apenas 5 verdadeiros positivos e 5 falsos negativos. Na **Decision Tree**, similar à **Random Forest**, a classe 'unacc' (classe 3) tem um bom desempenho e a classe 'good' (classe 2) também apresenta um desempenho decente. Quanto às **Neural Networks**, é de denotar um desempenho notável em todas as classes, com destaque para a classe 'good' (classe 2) que tem 11 verdadeiros positivos e nenhum falso negativo, referindo também que a classe 'vgood' (classe 4) mostra 14 verdadeiros positivos, mas também 2 falsos positivos.

- **Conclusões** : A comparação entre os modelos indica que todos apresentam desempenho sólido, mas o modelo de **Neural Networks (MLP)** destaca-se com uma acurácia de 98.3%. No entanto, é crucial considerar outras medidas de desempenho, como **precision, recall, F1-score e AUC-ROC**,

para uma avaliação mais abrangente. A **AUC-ROC** de 0.999 destaca a capacidade do modelo MLP em distinguir diferentes classes.

Link para o Código no GitHub : `https://github.com/boynewdoor/Car-Evaluation`

4.2 Apresentação de Problemas

- Durante os experimentos, enfrentei alguns desafios significativos. Um dos principais foi a carga computacional e a velocidade dos algoritmos, especialmente ao lidar com conjuntos de dados de tamanho considerável. Busquei otimizar a eficiência do processo, considerando opções como a paralelização de tarefas e a utilização de recursos computacionais mais avançados, como GPUs, para acelerar o treinamento dos modelos.

- A análise exploratória dos dados revelou algumas características que podem impactar o desempenho do modelo. Identifiquei desafios na representação de classes desbalanceadas, especialmente na categoria "good", onde o número de amostras é menor. Estratégias adicionais, como técnicas de reamostragem ou o uso de métricas específicas para classes desbalanceadas, podem ser exploradas para lidar com essa questão.

- O ajuste fino dos hiperparâmetros também foi um ponto crítico. Embora tenha obtido resultados promissores, explorar técnicas mais avançadas de otimização de hiperparâmetros, como busca em grade ou otimização bayesiana, pode proporcionar melhorias adicionais no desempenho dos modelos.

- Adicionalmente, a interpretabilidade dos modelos, especialmente para redes neurais profundas, pode ser um desafio. Estratégias de interpretabilidade, como SHAP (SHapley Additive exPlanations) ou LIME (Local Interpretable Model-agnostic Explanations), poderiam ser aplicadas para melhor compreensão das decisões do modelo.

5 Conclusões e Trabalho Futuro

5.1 Conclusão

- **Principais Descobertas e Insights** :
  - Os experimentos com quatro modelos diferentes **(Random Forest, SVM, Neural Networks e Decision Trees)** forneceram resultados promissores na tarefa de prever a aceitabilidade de carros com base em várias características.
  - A **Random Forest** apresentou uma precisão notável de 97.4%, destacando sua eficácia na classificação.
  - **Neural Networks (MLP)** demonstraram uma performance excepcional, atingindo uma precisão de 98.3% e uma **AUC-ROC** de 99.9%, indicando uma capacidade excepcional de distinguir entre diferentes classes.
  - A **Decision Trees** também obteve resultados sólidos, com uma precisão de 97.4%.
  - O **SVM**, embora tenha uma precisão de 91.3%, mostrou ser uma opção viável, considerando o desempenho competitivo em um conjunto de dados complexo.
  - Todas as **matrizes de confusão** apresentadas mostram um bom desempenho na classe 'unacc' (classe 3) indicando uma boa capacidade dos modelos de identificar essa classe. O desempenho nas outras classes varia entre os modelos, destacando a importância de considerar métricas detalhadas como precisão, recall e f1-score para uma avaliação completa.

Tabela 9: Modelos - Resultados

Model	Precision Score	Percentage	AUC-ROC
Random Forest	0.974	97.4%	0.997
SVM	0.913	91.3%	0.948
Neural Networks	0.983	98.3%	0.999
Decision Tree	0.974	97.4%	

- **Implicações dos Resultados**:
  - A análise das variáveis destacou a importância de características como preço, custos de manutenção, número de portas, capacidade de passageiros, tamanho da bagageira e níveis de segurança na determinação da aceitabilidade de carros.
  - A distribuição equilibrada das classes e as associações identificadas sugerem que todas as características analisadas podem ser preditoras significativas em modelos de aprendizagem de máquina.
  - A interpretação dos resultados deve levar em consideração a complexidade do problema e as nuances associadas às diferentes classes de aceitabilidade.

5.2 Trabalhos Futuros

- **Tópicos Potenciais para Pesquisas Futuras**:
  - Exploração de técnicas avançadas de otimização de hiperparâmetros para cada modelo, buscando melhorias adicionais no desempenho.
  - Investigação de abordagens específicas para lidar com classes desbalanceadas, especialmente na categoria ”good”, para aprimorar a generalização do modelo.
  - Avaliação do impacto de diferentes estratégias de interpretabilidade, como SHAP ou LIME, para compreender melhor as decisões dos modelos, principalmente em redes neurais profundas.
  - Extensão do estudo para incluir conjuntos de dados adicionais ou considerar diferentes categorias de carros para uma análise mais abrangente.
  - Exploração de modelos de ensemble que combinam os pontos fortes de diferentes algoritmos para obter um desempenho ainda mais robusto.

- **Melhorias ou Extensões para o Trabalho Atual**:
  - Refinamento da seleção de atributos, considerando abordagens mais avançadas, como análise de importância de características.
  - Investigação de estratégias para lidar com possíveis outliers de maneira mais específica, buscando uma melhor compreensão de sua influência nos modelos.
  - Consideração de técnicas de validação cruzada mais avançadas, como validação cruzada estratificada, para garantir uma avaliação mais precisa do desempenho do modelo em diferentes conjuntos de dados.
  - Aplicação de técnicas de normalização ou padronização específicas para redes neurais, visando otimizar o treinamento desses modelos.

Resumo

Neste estudo de mineração de dados, explorei a Base de Dados de Avaliação de Carros, utilizando abordagens e técnicas estabelecidas para avaliar a aceitabilidade de carros. O artigo apresenta a relevância do conjunto de dados, referências importantes, a minha abordagem escolhida e os resultados experimentais, proporcionando insights sobre o desempenho de vários algoritmos.

Referências

[1] Michael Bohlke-Schneider Valentin Flunkert Jan Gasthaus Tim Januschowski Danielle C. Maddix Syama Sundar Rangapuram David Salinas J. Schulz Lorenzo Stella Ali Caner Türkmen Bernie Wang A. Alexandrov, Konstantinos Benidis. Gluonts: Probabilistic time series models in python. 2019.

[2] Punniya D. Ganesan A. Tzacheva, A. Bagavathi. Mr - random forest algorithm for distributed action rules discovery. *International Journal of Data Mining Knowledge Management Process*, 2016.

[3] M. Bohanec. Car evaluation database. 1997.

[4] V. Bhuvaneswari K. Arunprabha. Comparing k-value estimation for categorical and numeric data clustring. *International Journal of Computer Applications*, 2010.

[5] Toshihide Ibaraki Kazuya Haraguchi. Studies on classifiers based on iteratively composed features. 2007.

[6] Chinmay D. Pai Kedar Potdar, Taher S. Pardawala. A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications*, 2017.

[7] V. Rajkovič M. Bohanec. Knowledge acquisition and explanation for multi-attribute decision making. 1988.

[8] Yoshitaka Kameya Taisuke Sato, Keiichi Kubota. A logic-based approach to generatively defined discriminative modeling. 2014.

[9] Najmeh Forouzandehmehr V. Jalali, David B. Leake. Learning and applying case adaptation rules for classification: An ensemble approach. 2017.

[10] M. Seltzer Xiyang Hu, C. Rudin. Optimal sparse decision trees. 2019.