PRACTICAL TEST - DATA SCIENTIST

PROBLEM 1

A) Present you exploratory data analysis.

Realizando a análise exploratória dos dados apresentada, identificamos insights relevantes a partir do conjunto de dados examinados.

Inicialmente, ao observar outliers e níveis de concentração, descobriu-se que determinados sensores, como PT08.S2, PT08.S3, e PT08.S1, não possuem limites de concentração definidos. Contudo, após a exclusão de dados ausentes, a média da concentração de monóxido de carbono (CO) ajustou-se para 2.182 mg/m³, valor dentro do aceitável e alinhado às diretrizes da Organização Mundial da Saúde, que estipula um limite máximo de 4 mg/m³ para os níveis de CO.

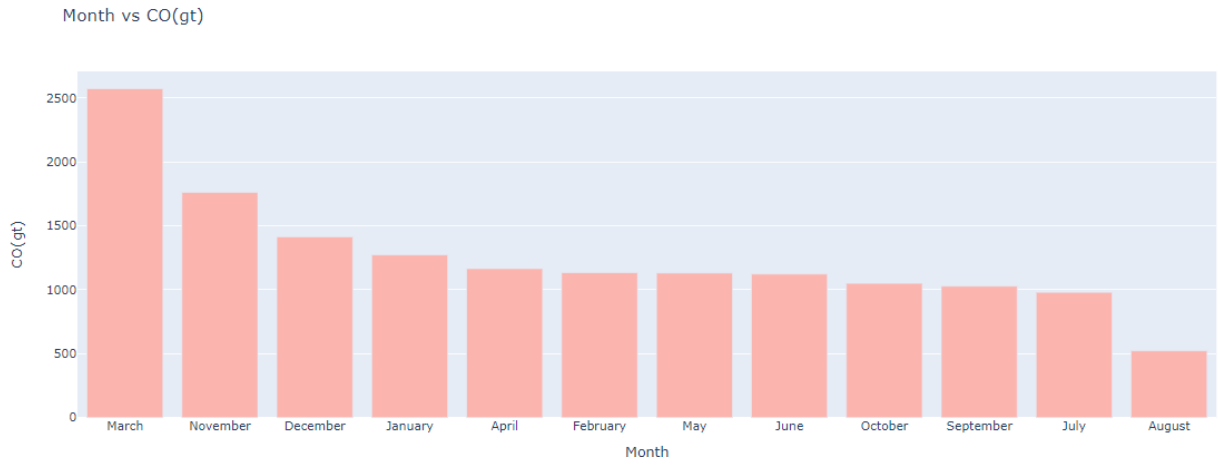


Figura 1 - Analise Temporal

Durante a análise temporal (Figura *1*), notou-se uma variação significativa nos níveis de CO ao longo do ano, com março apresentando os maiores índices e agosto, os menores, refletindo uma diferença marcante que pode ser atribuída às mudanças sazonais. A análise por dias da semana revelou que os níveis de CO tendem a ser mais elevados durante os dias úteis, com uma redução notável durante os finais de semana.

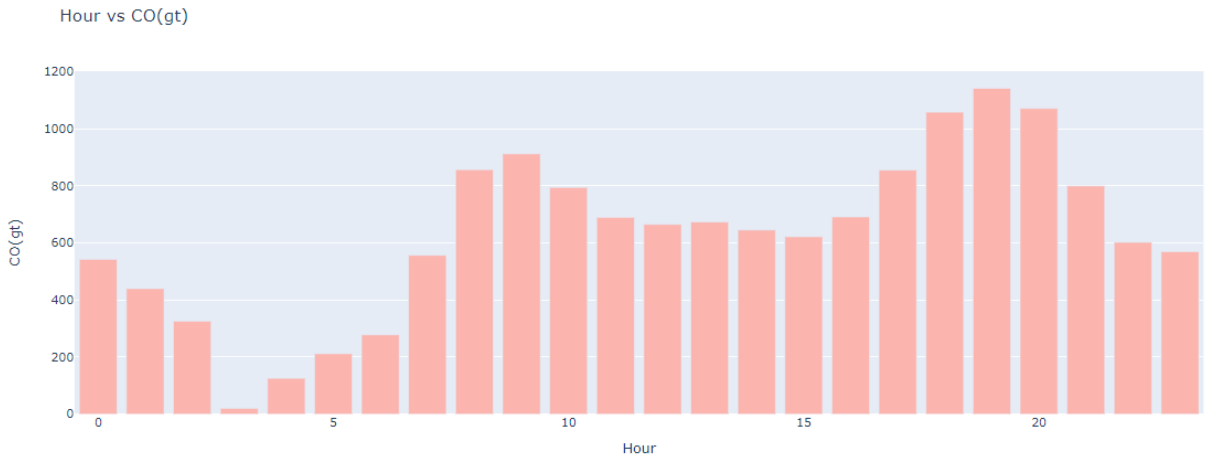


Figura 2 - Hour vs CO

Quanto aos horários de pico de concentração de CO (Figura 2), foram identificados períodos específicos no início da manhã e no final da tarde, possivelmente correlacionados aos horários de maior atividade humana, como o deslocamento para escolas e locais de trabalho. Os períodos de menor concentração de CO foram registrados nas primeiras horas da manhã, sugerindo uma diminuição das atividades.

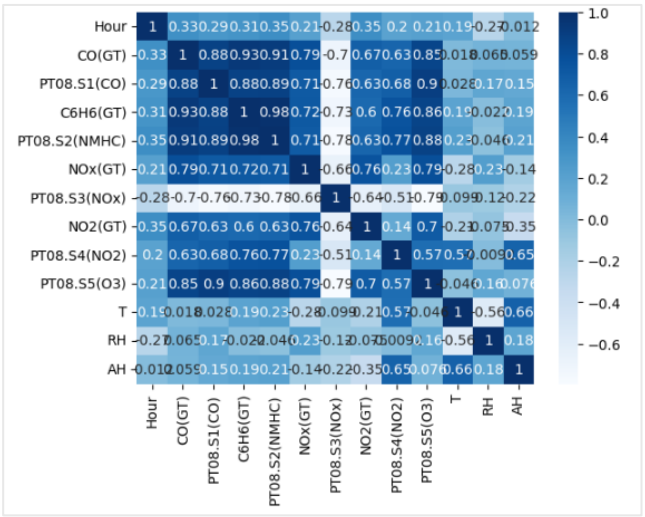


Figura 3 - Corr

A análise também revelou uma correlação significativa entre a umidade relativa e fatores ambientais, com uma relação notável entre a temperatura e a umidade relativa. Essa descoberta indica que períodos mais quentes podem apresentar menor umidade relativa, o que tem implicações importantes para a gestão ambiental, especialmente em áreas urbanas afetadas pelo efeito de ilha de calor. Além disso, a análise sugeriu uma

interação entre a umidade relativa e os níveis de óxidos de nitrogênio, indicando possíveis interações complexas entre a poluição do ar e as condições meteorológicas.

B) Estimate Relative Humidity behavior based on its answer to other parameters.

Para estimar o comportamento da Umidade Relativa (RH) em relação a outros parâmetros principais observa-se alguns dados importantes:

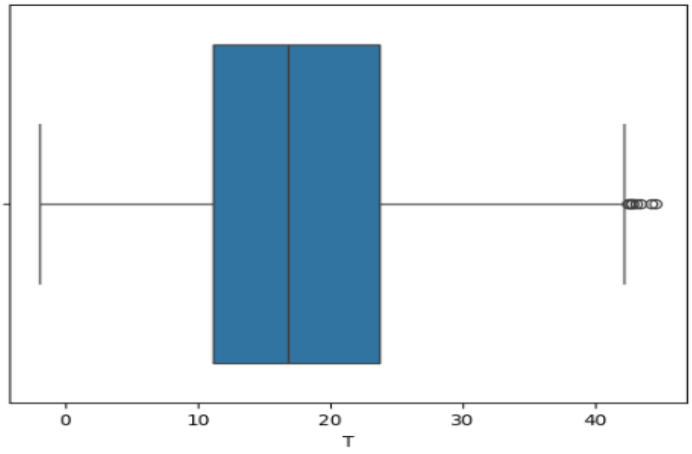


Figura 4 - BoxPlot Temperature

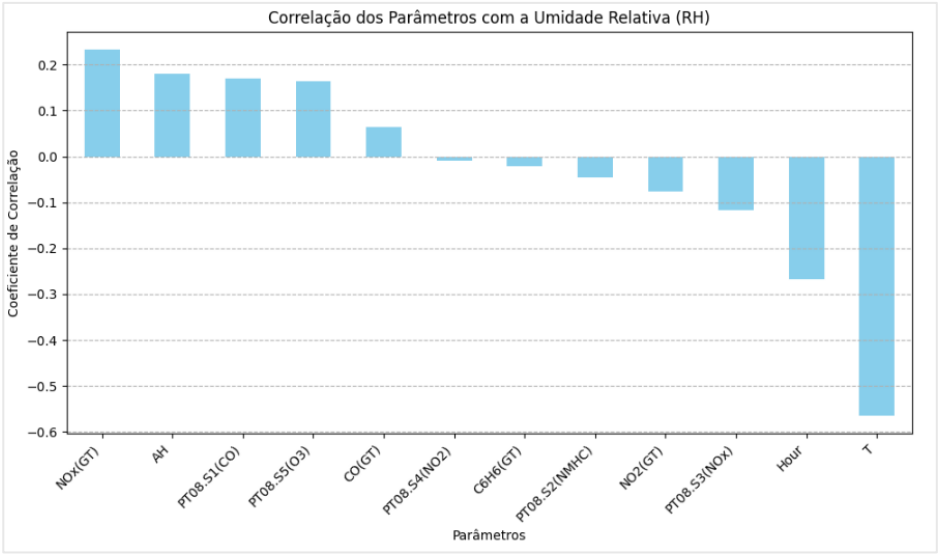


Figura 5 - Pearson Corr

Primeiro, a Temperatura (T) (Figura 4) mostrou uma grande variação, indo de frio a quente (-1.9°C a 44.6°C), com uma média de 17.8°C. Descobrimos que a temperatura tem uma forte relação inversa com a RH. Isso significa que, quando a temperatura

aumenta, geralmente a umidade relativa diminui. Portanto, em dias mais quentes, podemos esperar que o ar seja mais seco.

A Umidade Relativa (RH), por sua vez, varia bastante (9.2% a 88.7%), com uma média de 48.9%. Essa variação indica que a umidade pode mudar muito dependendo das condições climáticas e da hora do dia.

Com relação aos poluentes, a concentração de Monóxido de Carbono (CO) varia de 0.1 a 11.9 mg/m³, com uma média próxima de 2.18 mg/m³. Com essa exploração nota-se a relação da RH com os óxidos de nitrogênio (NOx), onde encontramos uma correlação positiva moderada (0.23). Isso sugere que em ambientes com mais umidade, a concentração de NOx tende a ser maior. Também observamos correlações positivas leves entre a RH e os sensores de CO e ozônio (O3), indicando que a umidade pode afetar ligeiramente a presença desses poluentes no ar.

Além disso, a hora do dia influencia a RH, com uma tendência de menor umidade relativa durante as horas mais quentes, o que reforça a relação entre temperatura e umidade.

Quanto aos modelos analíticos, o Gradiente Boosting se destacou com ótima performance (98), mostrando que é uma ferramenta confiável para prever a RH a partir desses parâmetros. A remoção dos outliers não mudou significativamente os resultados, o que indica que nossas observações são estáveis e confiáveis.

Conclui-se que a temperatura é um fator crucial que afeta inversamente a RH. A presença de certos poluentes como NOx também aumenta com a umidade. Essas descobertas nos ajudam a entender melhor como a umidade relativa se comporta e é influenciada por diferentes condições ambientais e poluentes.

Sem a remoção de outliers:

```
results_sorted = results.sort_values(by='Test R2', ascending=False)

# Exibir os melhores valores ordenados
print(results_sorted)
```

	Method	Training MSE	Training R2	Test MSE	Test R2
3	Gradient Boosting	2.951906	0.990239	3.458273	0.988831
0	Linear Regression	36.374550	0.879725	34.080248	0.889936
1	Huber Regression	38.935391	0.871258	35.434060	0.885564
4	Gaussian Process	37.591669	0.875701	35.520956	0.885283
6	Ada Boost	36.187602	0.880344	39.166687	0.873509
9	MLP	68.065265	0.774938	67.539955	0.781877
5	K-Neighbors	64.184545	0.787770	96.915838	0.687006
2	Random Forest	97.705572	0.676931	102.727439	0.668237
8	Decision Tree	110.316721	0.635231	116.418195	0.624022
7	SVR	193.345707	0.360691	195.634082	0.368190

Com remoção de outliers:

```
results_sorted = results.sort_values(by='Test R2', ascending=False)

# Exibir os melhores valores ordenados
print(results_sorted)
```

	Method	Training MSE	Training R2	Test MSE	Test R2
3	Gradient Boosting	3.025345	0.990143	3.869986	0.986630
9	MLP	21.689596	0.929331	22.210800	0.923266
0	Linear Regression	33.839091	0.889745	31.831988	0.890027
4	Gaussian Process	35.152506	0.885466	32.920851	0.886265
1	Huber Regression	39.870559	0.870094	37.697269	0.869763
6	Ada Boost	35.994745	0.882722	37.975108	0.868803
5	K-Neighbors	62.771240	0.795478	96.954307	0.665042
2	Random Forest	99.762547	0.674953	101.698491	0.648651
8	Decision Tree	112.437955	0.633654	117.110192	0.595407
7	SVR	197.237813	0.357359	184.398454	0.362939

PROBLEM 2

A) Provide some insights on the data such as shape, distribution and cross-category comparisons (data exploration)

Ao avaliar o conjunto de dados compostos por 1728 linhas e 7 colunas, obteve-se uma visão detalhada sobre as características determinantes para a classificação de automóveis. O conjunto de dados está completo, sem valores nulos ou duplicatas.

Dentre os vários insights extraídos, destacam-se :

	frequencia	Porcentagem (%)
unacc	1210	0.700231
acc	384	0.222222
good	69	0.039931
vgood	65	0.037616

Figura 6 - Class Grouped

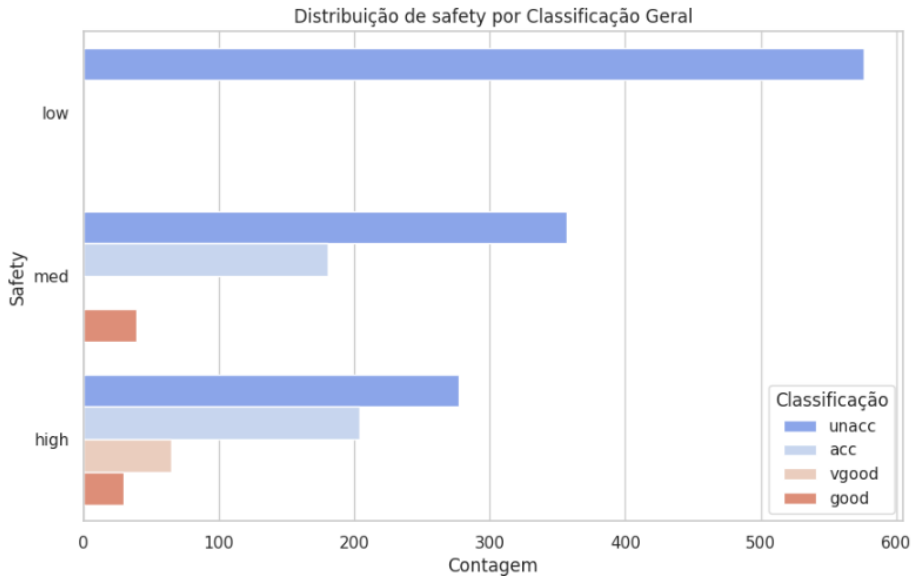


Figura 7 - Safety vs Class

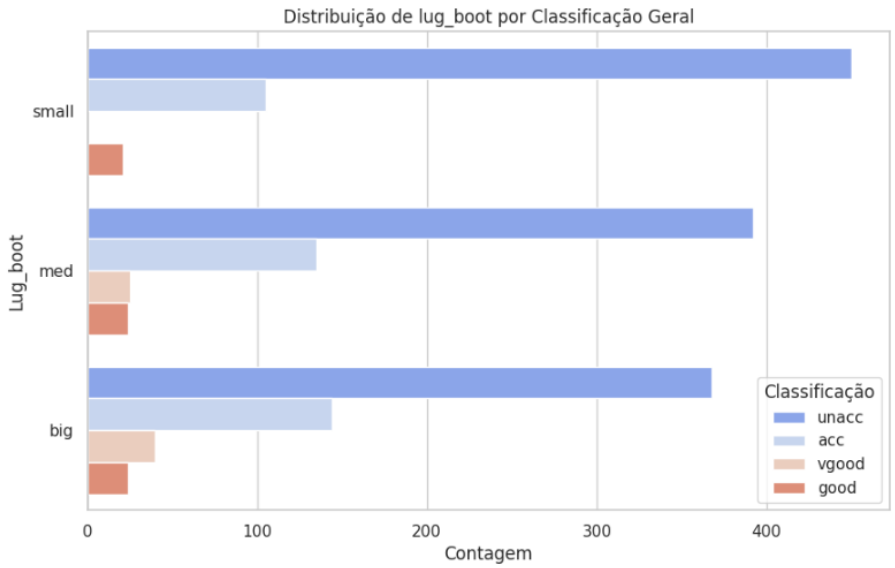


Figura 8 - Porta Mala vs Class

- Categoria 'Unacceptable' é comum: Cerca de 70% (Figura 6) dos carros foram classificados como 'Unacceptable'. Isso mostra uma tendência clara, indicando que muitos carros não atendem a certos critérios desejados.
- Número de portas e avaliações: Foi notado que carros com 5 ou mais portas tendem a receber avaliações um pouco melhores. Isso sugere que ter mais portas pode ser visto de forma positiva, embora a diferença não seja muito grande.
- Segurança e Classificação: Carros classificados com baixa segurança foram todos considerados 'Unacceptable'(Figura 7). Isso ressalta a importância da segurança nos critérios de avaliação.
- Espaço do porta-malas: Veículos com um porta-malas grande (Figura 8) receberam avaliações melhores. Isso indica que um espaço maior de armazenamento é valorizado.
- Carros caros nem sempre são bem avaliados: Carros com preços mais altos não necessariamente receberam melhores avaliações, o que pode sugerir que os consumidores buscam um bom equilíbrio entre custo e benefício.

Com base nesses insights, a recomendação para os fabricantes de carros seria focar na segurança e no custo-benefício, além de considerar a importância do espaço.

Essas são áreas que parecem influenciar as expectativas do mercado e, se melhoradas, podem levar a uma maior satisfação do cliente.

B) Given Logistic Regression, Random Forest Classifier and Decision Tree, which model performs better when predicting car class? Justify your answer with data.

Durante a análise comparativa dos modelos de aprendizado de máquina, observa-se que o modelo Random Forest teve um desempenho melhor em comparação com os modelos Decision Tree e Logistic Regression. Embora os modelos anteriores tenham alcançado taxas de acurácia de 81.9% e 75.72%, respectivamente, suas pontuações de f1-score macro foram de 39.98% e 35.14%, o que pode ser atribuído à classe "unacc", que impacta negativamente a taxa de acerto do f1-score macro.

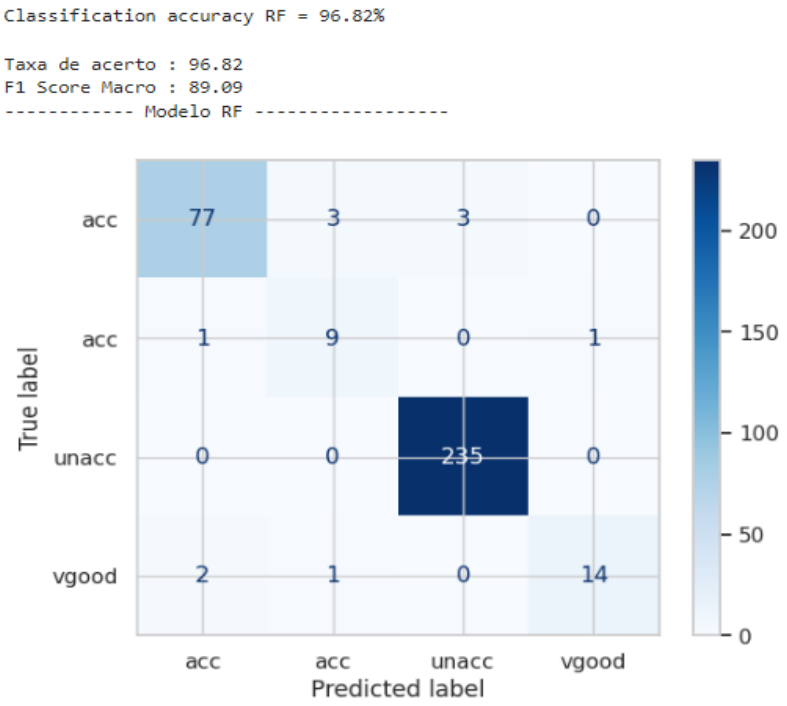


Figura 9 - RF

Por outro lado, o modelo Random Forest teve uma taxa de acerto de 96.82%, (Figura 9) acompanhada por um f1-score macro de 89%. Este desempenho notável reflete a capacidade do modelo em fazer previsões precisas para a classe "unacc", elevando assim sua pontuação de f1-score macro.

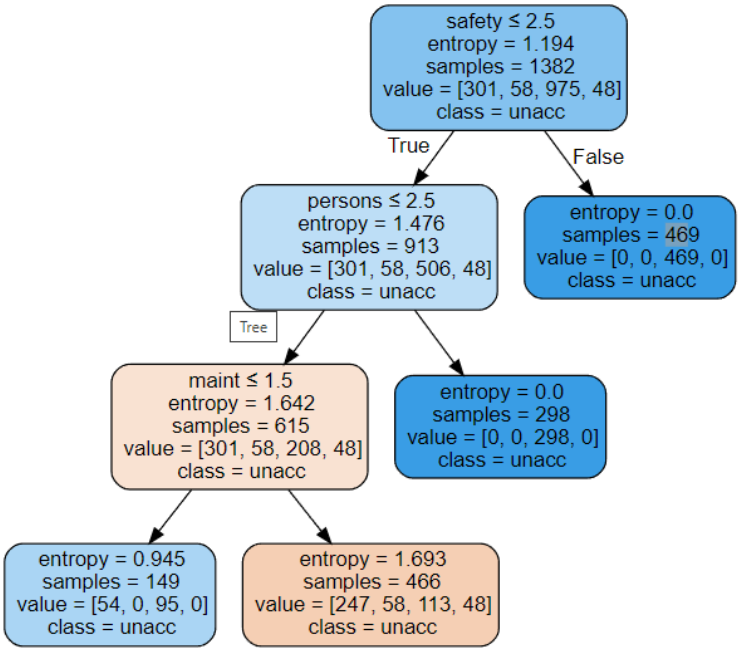
C) Rank feature importance with respect to Random Forest Model and share your insights.

A capacidade de um veículo em acomodar um certo número de passageiros não é apenas uma questão de espaço ou conveniência; ela se revela um critério chave na definição de sua classificação como por exemplo: carro familiar, esportivo ou compacto. Esse fator influencia diretamente em aspectos como segurança, conforto e funcionalidade, o que depende da quantidade de indivíduos que o carro pode transportar com conforto.

Na análise de dados veiculares, o modelo RandomForest analisa a importância das variáveis ao observar como as divisões feitas nas árvores modelos reduzem a impureza dos nós. As variáveis que, ao dividir os dados, resultam em uma maior redução de impureza ou erro são vistas como mais importantes. Portanto, 'persons' (pessoas) e 'buying' (compra) são variáveis que se destacam por ajudar significativamente na diferenciação e classificação correta dos dados de treino.

É importante notar que a importância das variáveis depende do conjunto de dados específico e de como o modelo foi treinado. Mudanças nos dados ou no treinamento podem alterar quais variáveis são consideradas mais importantes. Além disso, a importância de uma variável não indica causalidade, mas sim que ela é útil para o modelo fazer suas previsões ou classificações com os dados que tem.

D) Present a visualization of the Decision Tree and share your insights.



Ao analisar a árvore de decisão apresentada na figura acima, observa-se que os critérios decisivos para a classificação de automóveis. A primeira e mais notável divisão ocorre no atributo de segurança: veículos com índices de segurança inferiores ou iguais a 2.5 são prontamente classificados como 'unacceptable' (unacc). Este critério reflete a preeminência da segurança enquanto fator de avaliação na percepção da qualidade do carro.

Seguindo a estrutura da árvore, o segundo fator decisivo revela: a capacidade do carro em termos de número de passageiros. Veículos com a capacidade de acomodar 2.5 pessoas ou menos recebem, igualmente, a classificação 'unacc'.

Este modelo estratégico parece dar prioridade à segurança e capacidade do veículo, e subsequente atenção à manutenção, para afinar a categorização dos carros.

Além disso, o modelo sinaliza a manutenção como outro fator decisivo. Carros com custos de manutenção menores ou iguais a 1.5 continuam na categoria 'acc', o que sugere uma correlação entre menores custos de manutenção e percepções

negativas sobre o veículo. Ao proceder através dos ramos da árvore, o modelo meticulosamente minimiza a entropia, isto é, a desordem ou incerteza nas decisões, conduzindo a uma categorização cada vez mais acurada e evidente.

O objetivo claro é reduzir a entropia - uma medida de incerteza ou impureza - facilitando assim uma classificação mais direta e menos ambígua. É evidente que os fatores escolhidos são cruciais para esta análise, mostrando um processo de decisão metódico para chegar a uma classificação consistente.