Universidade Federal do Rio Grande do Sul
Instituto de Informática

INF01017
Aprendizado de Máquina

Trabalho Prático Final – Etapa 1

Predição de Consumo de Combustível

Luís Filipe Martini Gastmann (00276150)
Pedro Lubaszewski Lima (00341810)
Vinícius Boff Alves (00335551)

Turma U

9 de novembro de 2024

Sumário

1.1 Definição do Problema e Coleta de Dados

O objetivo deste trabalho é prever o consumo médio de combustível de um carro através de algumas das suas características e origens de fabricação. Alguns atributos das instâncias são a marca, a quantidade de cilíndros, o porte do veículo etc.

O conjunto de dados utilizado para desenvolver este trabalho foi obtido da seguinte página do Kaggle: Explore Car Performance: Fuel Efficiency Data. Essa tarefa contará com diversas técnicas de preparação dos dados para posteriormente iniciar a seleção e avaliação de modelos para essa tarefa.

2.1 Análise Exploratória e Pré-processamento dos Dados

2.1.1 Análise Exploratória dos Dados

Este *dataset* possui 550 instâncias, com 11 atributos preditores e 1 atributo alvo. Esse último torna a tarefa dos modelos em regressão, visto que o objetivo aqui é prever o consumo médio de combustível de carros. No conjunto de dados, esse atributo predito se chama *combination mpg*.

Dos atributos preditores, observa-se que há 5 atributos numéricos (*city mpg*, *cylinders*, *displacement*, *highway mpg* e *year*). Além deles, os 6 atributos restantes são categóricos: *class*, *drive*, *fuel type*, *make*, *model* e *transmission*. Com isso em mente, criou-se alguns gráficos para analisar correlações e distribuições dos dados. A seguir, o *violin plot* do atributo alvo pelo número de veículos:

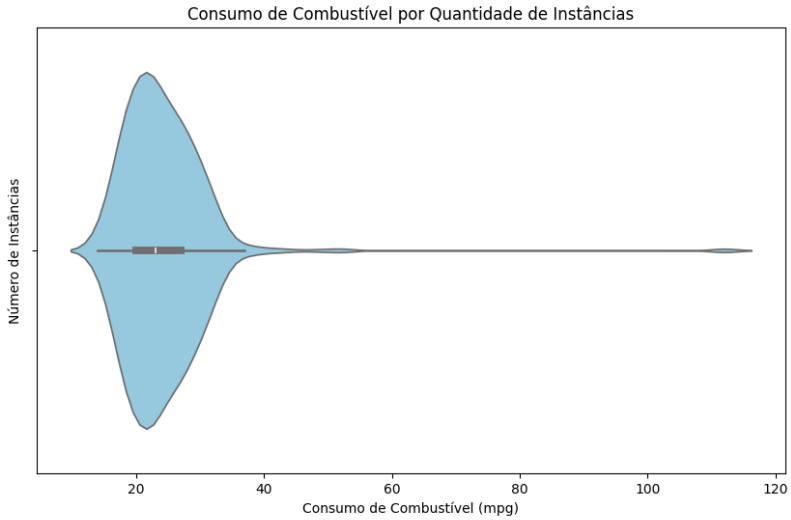

Figura 1: *Violin Plot* de Consumo Médio

Nesse ponto, já se observa que há instâncias problemáticas, claramente *outliers* que devem ser removidas do conjunto de dados. Ademais, a maior concentração dos veículos apresenta consumo médio entre 15 e 30mpg, algo que pode guiar bastante os modelos. Após isso, analisou-se o atributo de classes de veículos para ter uma ideia da sua distribuição em um gráfico de setores.

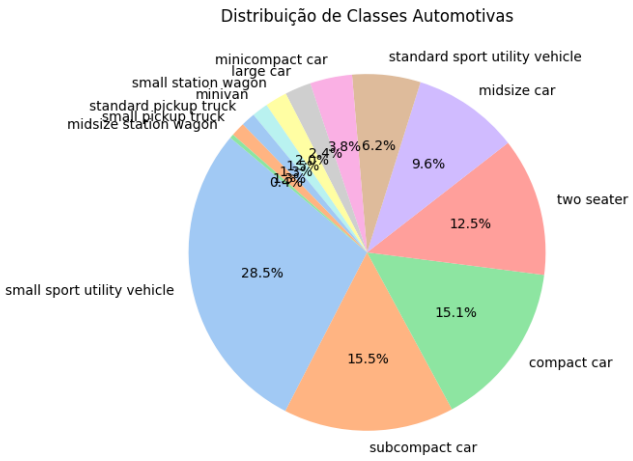

Figura 2: Distribuição de Classes de Veículos

Com ele, é perceptível que talvez seja necessário agrupar as classes de veículos e outros

atributos que contenham classes com muito poucos representantes, como *midsize station wagon*, por exemplo, em uma classe geral chamada *others*. Porém, isso será melhor analisado, se necessário, na segunda etapa do trabalho. No entanto, algo que é necessário é a codificação dos atributos categóricos em numéricos para padronizar a entrada numérica dos modelos.

Por fim, analisou-se a Correlação de Pearson entre os atributos numéricos através do *heatmap* abaixo:

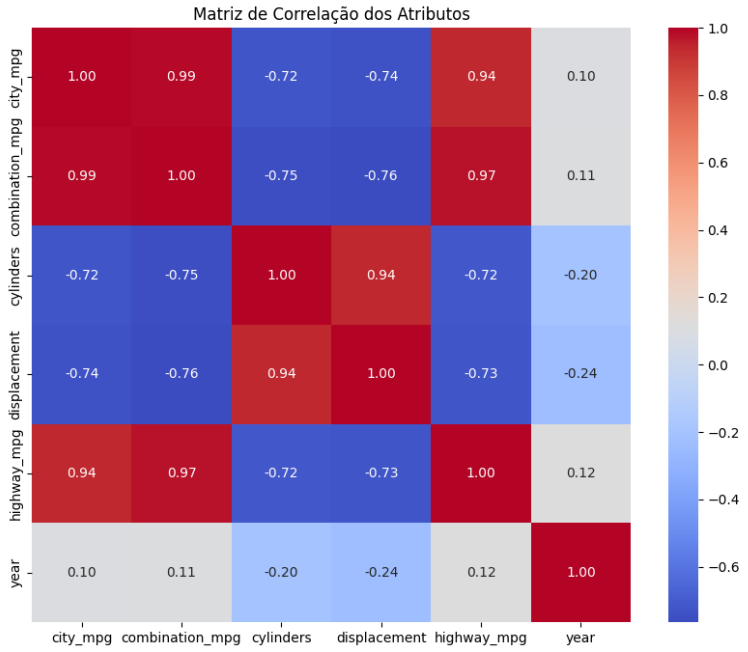

Figura 3: *Heatmap* de Correlações entre Atributos Numéricos

Olhando para o gráfico acima, percebe-se que há diversos atributos preditores que apresentam alto nível de correlação com a saída *combination mpg* e entre si. Ambos os atributos *highway mpg* e *city mpg* serão descartados do *dataset* por terem alta correlação entre si e com o valor da saída do modelo, facilitando demais a tarefa dos modelos. Além disso, *displacement* ou (exclusivo) *cylinders* será excluído do conjunto de dados por terem alta correlação entre si.

Isto não foi ilustrado pelos gráficos, porém os dados numéricos apresentam diversas escalas diferentes, exigindo técnicas de normalização após a separação de conjuntos de treinamento/validação e de testes.

2.1.2 Pré-processamento dos Dados

Como já discutido na seção acima e analisando alguns trabalhos realizados com esse conjunto de dados, esses sendo de Lunovian [3], da Ayşe [1] e do Priyanshu [5], implementou-se algumas estratégias de pré-processamento nos dados deste problema.

Primeiramente, realizou-se a remoção dos atributos discutidos anteriormentes. Esses sendo o *highway mpg*, *city mpg* e *displacement*, visto que apresentam alto níveis de correlação entre si e com o valor da saída do modelo. Após essa remoção, filtrou-se as instâncias com muitos atributos faltantes (em geral, carros elétricos que funcionam diferentemente dos carros que funcionam com combustíveis fósseis).

Depois disso, foi realizada a limpeza de instâncias cujo atributo alvo representava um *outlier* no gráfico de distribuição de consumo médio dos veículos. Essa estratégia é necessária para evitar que os modelos sejam treinados com instâncias que não representam bem a grande a maioria dos veículos. Utilizou-se a medida padrão de 1.5 vezes o IQR para identificar *outliers*.

Sobre a codificação dos atributos, o grupo utilizou a codificação padrão de categórico para numérico que é o *one-hot encoding*. A justificativa para isso é que os atributos categóricos

não apresentam nenhuma ordem específica que justifique utilizar uma codificação em números inteiros. Além disso, em relação à normalização, baseado no trabalho do Jason Brownlee [2] e do Matheus Vasconcelos [4], optou-se por utilizar a padronização por *z-score*, visto que os dados aproximam uma distribuição normal. Tanto a codificação, quanto a normalização foram aplicadas após a separação de conjuntos de treinamento/validação e de testes.

Um outro problema encontrado consiste na possibilidade de não ser visto algum modelo de carro (*model*) no treinamento. Isso pode acontecer pois há uma quantidade muito grande de modelos diferentes, alguns deles com poucas instâncias. Para solucionar esse problema, "forçou-se"a haver pelo menos uma instância de cada modelo de veículo no conjunto de treinamento, antes de realizar a separação de conjunto de dados para teste. Para realizar essa última separação, utilizou-se a função `train_test_split()` com um `test_size` de 15%.

3.1 Abordagem, Algoritmos e Estratégias de Avaliação

4.1 *Spot-checking* de Algoritmos

Bibliografia

[1] Ayşe Nur Durmaz. OruntuTanima-Perceptron. Website available: https://www.kaggle.com/code/ayenurdurmaz/oruntutanima-perceptron. 2024.

[2] Jason Brownlee. How to Use the ColumnTransformer for Data Preparation. Website available: https://machinelearningmastery.com/columntransformer-for-numerical-and-categorical-data/. 2020.

[3] Lunovian. Which Cars Contribute to a Greener Future? Website available: https://www.kaggle.com/code/anmatngu/which-cars-contribute-to-a-greener-future. 2024.

[4] Matheus Vasconcelos. Distribuição normal e a classe Standard Scaler da biblioteca Scikit-learn... Website available: https://medium.com/@mhvasconcelos/distribuio-normal-e-a-biblioteca-standard-scaler-em-python-f21c52070c6b. 2023.

[5] Priyanshu shukla. Analysis on Car performance dataset. Website available: https://www.kaggle.com/code/priyanshu594/analysis-on-car-performance-dataset. 2024.