Universidade Federal do Rio Grande do Sul
Instituto de Informática



INF01017
Aprendizado de Máquina

# Atividade Prática 06

## Descrição dos Dados e Análise de Necessidades de Pré-processamento

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

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 essa observação, já é previsto que será necessário fazer a codificação dos atributos categóricos e a normalização dos atributos numéricos. Para converter os atributos categóricos em númericos, como são todos categoriais sem ordem estabelecida, será utilizada a codificação *One-hot Encoding*. É claro que isso aumenta a dimensionalidade do problema. Por conta disso, ainda será avaliado se serão necessárias medidas, como PCA, para reduzir a quantidade de atributos. Além disso, antes dessas conversões, o grupo decidiu abandonar os atributos de *city mpg* e *highway mpg* por facilitarem demais o trabalho do preditor, visto que o atributo alvo, *combination mpg*, pode ser calculado diretamente com esses dois atributos, algo que tornaria o modelo menos relevante. Se não for possível atingir uma métrica de qualidade satisfatória, alguma dessas características pode ser reintroduzida no conjunto de dados. Portanto, o conjunto de dados acabou com 3 atributos numéricos normalizados e 6 atributos categóricos codificados.

Em relação a outras características das *features* do problema, criou-se alguns gráficos para analisar correlações e distribuições dos dados. A seguir, o histograma do atributo alvo pelo número de veículos:

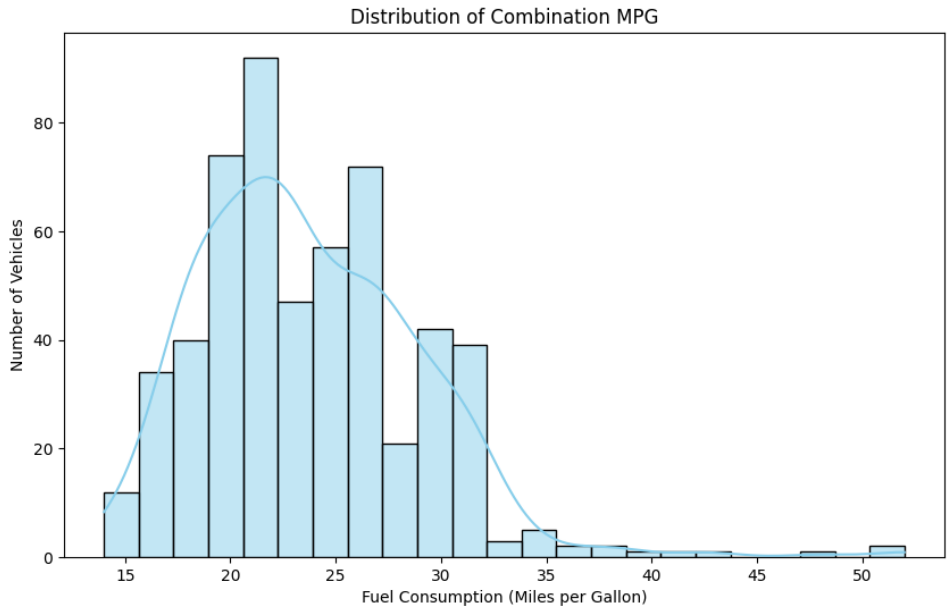


Figura 1: Histograma de Consumo Médio

Nesse gráfico, percebe-se que a maioria dos modelos tem consumo médio de combustível entre 20 e 30 mpg. Portanto, isso pode ser útil caso o grupo necessite realizar a remoção de *outliers* e também é problemático pois não há muitas instâncias com consumo acima de 35 mpg, tornando os modelos mais fracos nas predições com essa grandeza de consumo.

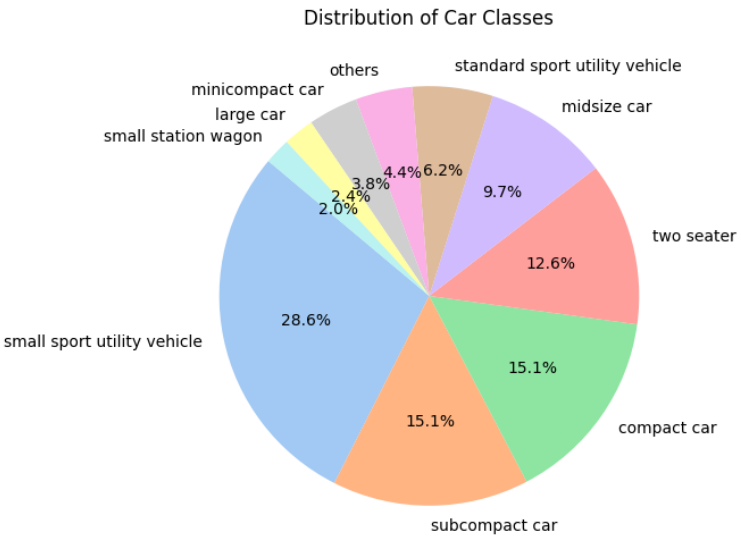


Figura 2: Distribuição de Classes de Veículos

Há diversas classes de veículos diferentes, com as mais diversas proporções. Por conta disso, adotar-se-á a política de agrupar essas classes e outros atributos categoriais em uma classe chamada *others* quando a quantidade de instâncias dela no conjunto de treinamento for insignificante para melhorar a generalização do modelo.

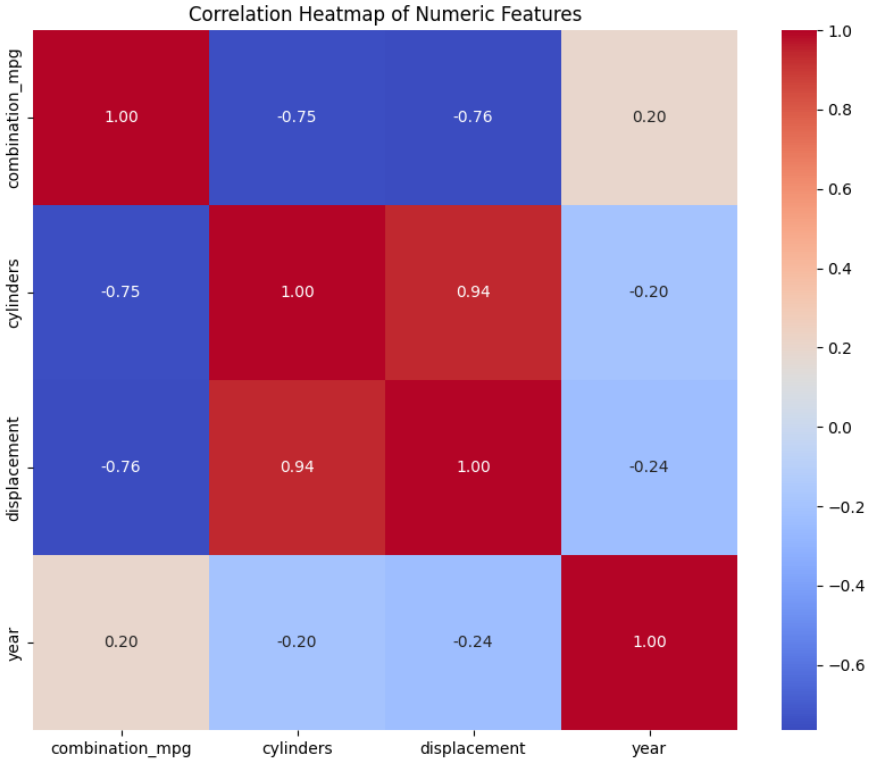


Figura 3: *Heatmap* de Correlações entre Atributos Numéricos

No diagrama acima, são observadas as correlações dos atributos numéricos e do atributo alvo. Esse mapa foi feito através da Correlação de Pearson de cada um dos atributos. Nele, é perceptível que bem possivelmente possa ser removido o atributo *cylinders* ou (exclusivo) o atributo *displacement*, visto que eles apresentam alto índice de correlação, tornando a utilização dos dois redundante para os modelos.

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 com poucas instâncias. Para solucionar esse problema, "forçou-se"a haver pelo menos uma instância de cada modelo de veículo no conjunto de treinamento.