# Automobile Data Set

Seguros de Carros

# Informações do Dataset

- Nome: Automobile
- Descrição: Avaliação sobre confiabilidade para o fornecimento de seguro para carros.
- Instâncias: 205 com 26 atributos.

# Principais Colunas

- **Symboling**: Avaliação sobre o risco de se fornecer um seguro para o analisado. O valor de risco máximo 3 e o mínimo é -3.
- **Normalized-lossers**: Taxa de desvalorização do carro ao longo de um ano com base na desvalorização de outros carros.
- **Price**: Preço do carro.

# Objetivo principal:

- Descobrir qual a grau de confiabilidade na hora de fornecer um seguro para um carro com base em suas características técnicas, preço e desvalorização ao longo do ano.

# Tipos de colunas:

Numérico(*int*)

- symboling
- normalized-losses
- engine-type
- compression-ratio
- horsepower
- peak-rpm
- city-mpg
- highway-mpg
- price

Categórico(*boolean*)

- make
- fuel-type
- aspiration
- num-of-doors
- body-style
- drive-wheels
- engine-location
- fuel-system
- num-of-cylinders

Real(*float*)

- wheel-base
- length
- width
- height
- curb-weight
- engine-size
- bore
- stroke

# Campos nulos

- Os campos nulos neste dataset são representados pelo caractere ‘*?*’
- Quantidade de campos nulos encontrados nas seguintes colunas:
- normalized-losses : 41
- num-of-doors : 2
- bore : 4
- stroke : 4
- horsepower : 2
- peak-rpm : 2
- price : 4

## Campos novos

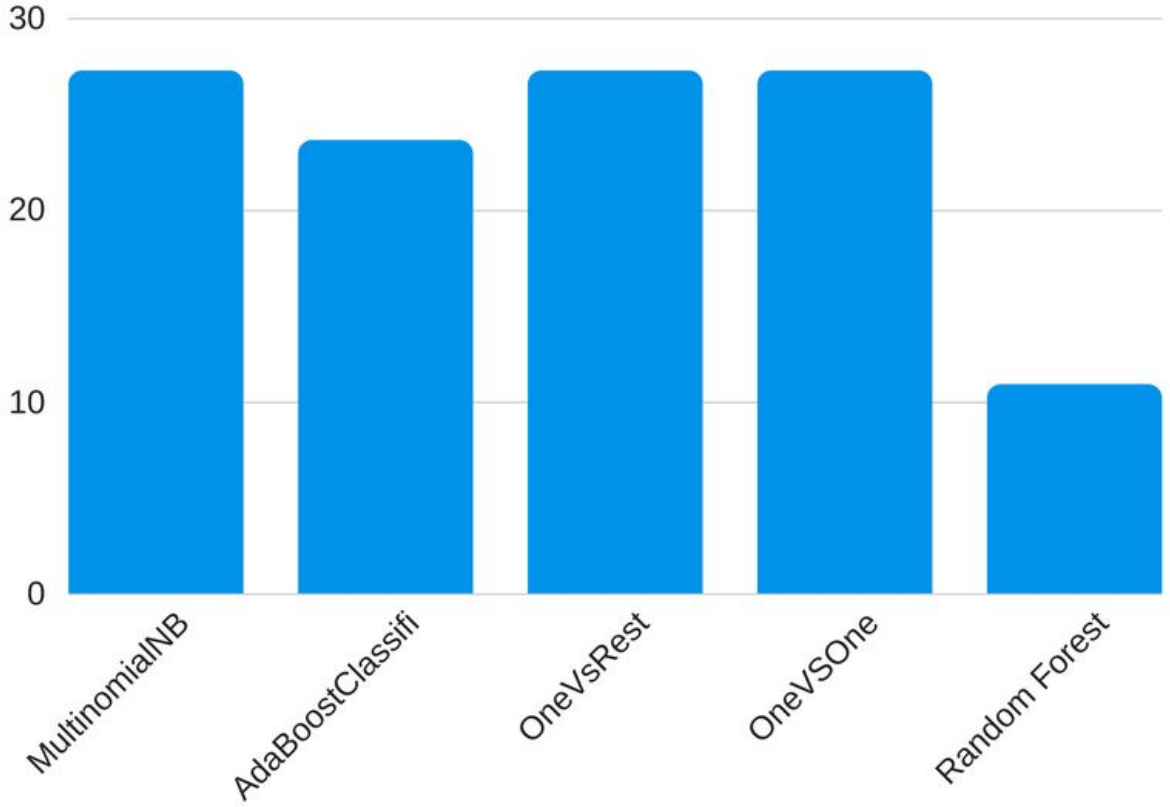
- Nenhum campo foi removido, apenas houveram campos de variáveis categóricas que foram convertidos para diversas colunas de campos de valor boolean pela biblioteca pandas.

## Campos Descartados

- make: Esse campo representa a fabricante do carro. Foi descartado porque algumas fabricantes possui diferentes tipos de carros como popular, esportivo

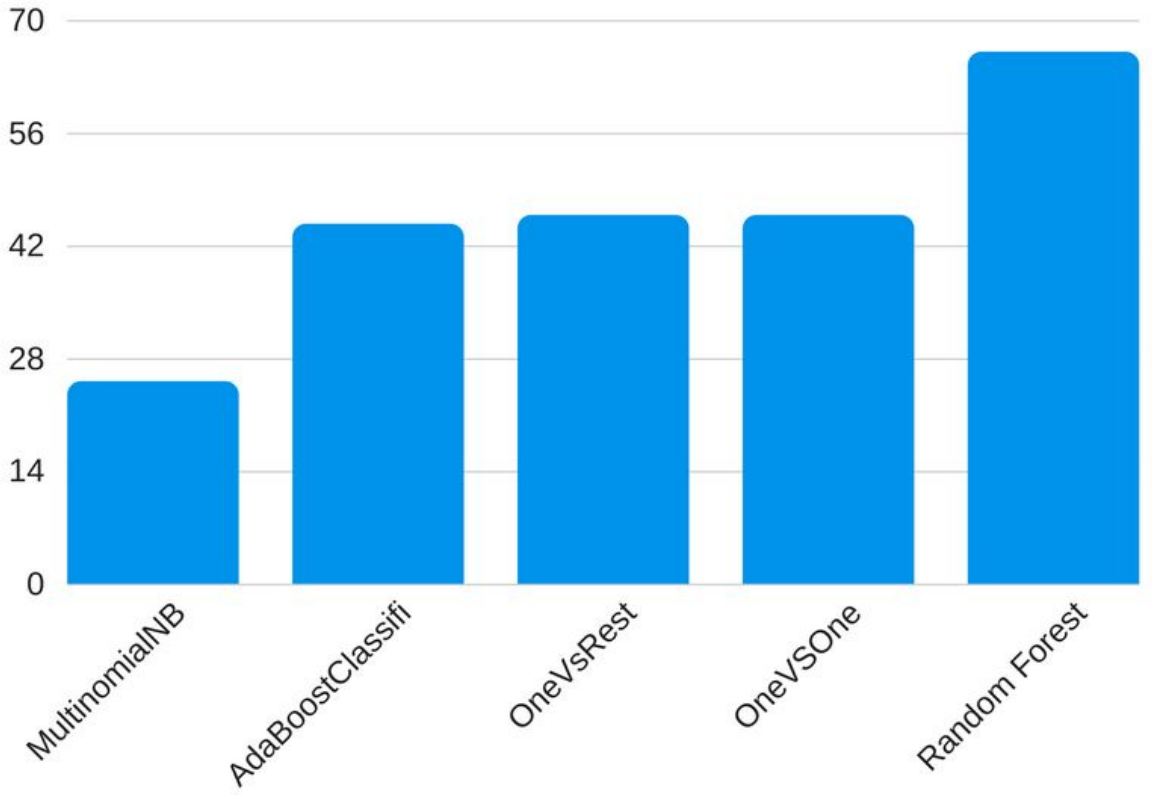
# Aplicando Algoritimos

Treino:150
Teste:56



# Utilizando K-Folding

Treino:150
Teste:56



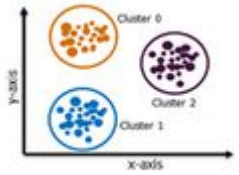
# Utilizando Algoritmo KMeans

- **Conceito**: O K-means é um algoritmo do tipo não supervisionado, ou seja, que não trabalha com dados rotulados.O objetivo desse algoritmo é encontrar similaridades entre os dados e agrupá-los conforme o número de cluster passado pelo argumento k.

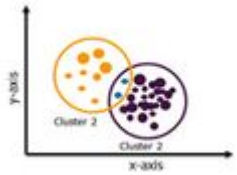
- **Funcionamento**: Dado um determinado registro em um conjunto de dados, é calculado uma distância entre os valores dos seus atributos com os demais. O objetivo de encontrar a similaridade entre esses registros.

# Tipos de grupos ou cluster

- Exclusive Cluster ou Grupo Exclusivo




- Overlapping Cluster ou Cluster Sobreposto




- Hierarchial Cluster ou Cluster hierárquico

# Processo KMeans

- **Inicialização**: gera de forma aleatória k centroids. Estes centroids são pontos de dados que serão utilizados, como o nome sugere, de pontos centrais dos clusters.
- **Atribuição ao Cluster**: Nesta etapa, é calculado a distância entre todos os pontos de dados e cada um dos centroids.Cada registro será atribuído ao centroid ou cluster que tem a menor distância.
- **Movimentação de Centroids**: calculada a média dos valores dos pontos de dados de cada cluster e o valor médio será o novo centróide.
- **Otimização do K-médias**: As fases Atribuição ao Cluster e Movimentação de Centroids são repetidas até o cluster se tornar estático ou algum critério de parada tenha sido atingido

# Ilustração do processo

# Execução do Algoritmo

- O algoritmo não apresenta resultados determinísticos.
- Foram utilizados 150 dados de treino e 56 de teste e validação
- Também foram utilizados 3 clusters com centroids em pontos aleatórios
- A precisão do algoritmo possui uma média de 18.18%

# Referências


- KMeans: http://minerandodados.com.br/index.php/2017/12/12/entenda-o-algoritmo-k-means/
- Dataset: https://archive.ics.uci.edu/ml/datasets/Automobile