

Nome: Aimê Gomes da Nobrega  
Nro Cruzeiro do Sul: 21862222  
Disciplina: Projeto e Aplicação de Mineração de Dados

# Projeto de Mineração de Dados

## Descrição









Fazer um experimento de Machine Learning que preveja o preço dos carros a partir de dados como fabricante, modelo, horsepower, etc.

## Regressão Linear

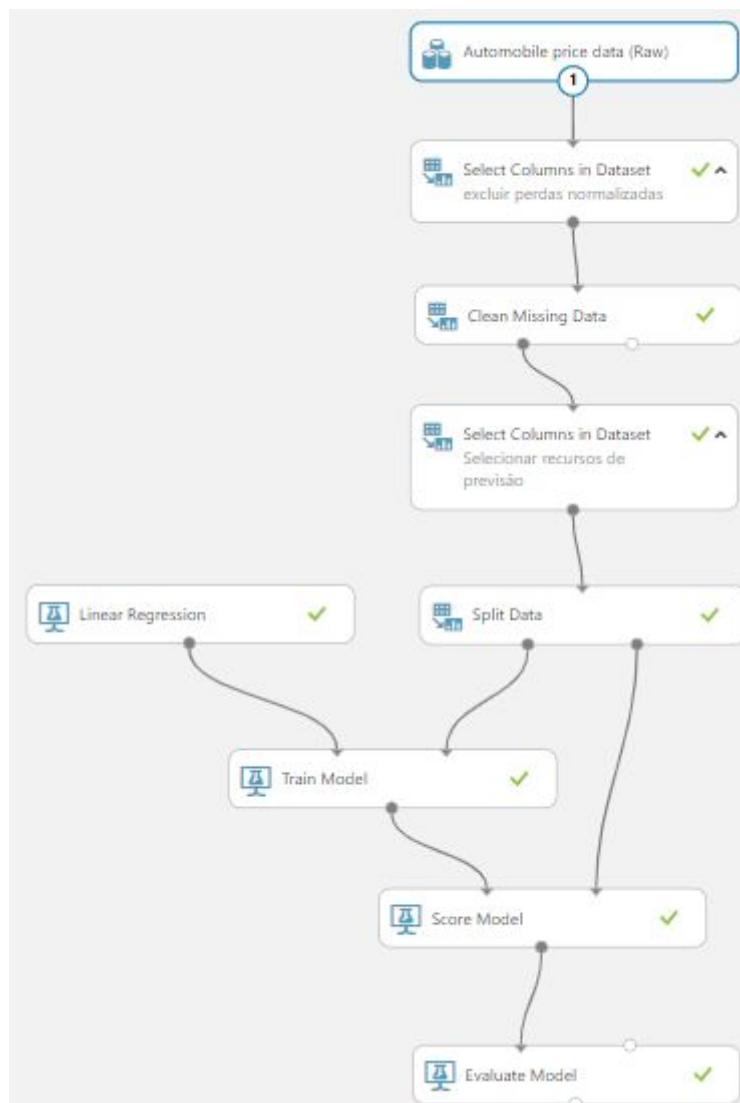
## Dados

Será utilizado um dataset disponível no workset como Sample chamado Automobile price data (Raw). Esse dataset inclui informações de automóveis, como fabricante, modelo e preço.

Previsão de preço de automóvel > Automobile price data (Raw) > dataset

rows		columns						
205		26						
g	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location
								
		alfa-romero	gas	std	two	convertible	rwd	front
		alfa-romero	gas	std	two	convertible	rwd	front
		alfa-romero	gas	std	two	hatchback	rwd	front
164		audi	gas	std	four	sedan	fwd	front
164		audi	gas	std	four	sedan	4wd	front
		audi	gas	std	two	sedan	fwd	front
158		audi	gas	std	four	sedan	fwd	front

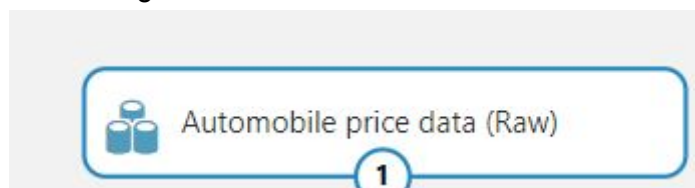
## Modelo



### Pré-processamento dos dados

É necessário fazer um pré-processamento dos dados, para limpar e obter um dataset que possa ser analisado.

- Pegar os dados



- Limpar os dados

Primeiramente nota-se que a coluna Normalized-losses possui muitos valores faltando. Ela será excluída para melhor manejo dos dados.

#### Statistics

Mean	122
Median	115
Min	65
Max	256
Standard Deviation	35.4422
Unique Values	51
Missing Values	41
Feature Type	Numeric Feature

#### Visualizations

normalized-losses

#### Select Columns in Dataset

Select columns

**Selected columns:**  
**All columns**  
**Exclude column**  
**names:** normalized-  
losses

A seguir, será excluída linhas faltando valores. Nesse caso, será excluída a linha inteira caso encontre um dado faltando.


Properties Project

#### Clean Missing Data

Columns to be cleaned

**Selected columns:**  
**All columns**

Launch column selector

Minimum missing valu... 

0

Maximum missing valu... 

1

Cleaning mode

Remove entire row 

Assim, obtém-se no final um dataset limpo, ou seja, um dataset todo preenchido.

- Definir parâmetros  
Aqui será selecionado as colunas que contém as informações relevantes para prever o valor do carro.

Properties Project

#### ▲ Select Columns in Dataset

Select columns

**Selected columns:**  
**Column names:**  
make,body-style,wheel-  
base,engine-  
size,horsepower,peak-  
rpm,highway-mpg,price

Launch column selector

## Treinamento

Será usada 75% dos dados para treinamento e 25% para teste.

#### ▲ Split Data

Splitting mode

Split Rows

Fraction of rows in the ...

0.75

☒ Randomized split

Treinaremos o modelo usando a coluna "Price".



#### ▲ Train Model

Label column

**Selected columns:**  
**Column names:** price

Porque queremos prever o preço, que é um número, usaremos regressão linear.

## Score e Evaluation Results

price	Scored Labels	Visualizando o Score Model nós podemos ver os valores previstos de preços previstos pelo algoritmo. "Price" é o valor original, e "Scored Labels" o valor gerado pelo modelo.
		
11259	10286.204819	
6669	5446.847864	
6229	6344.800711	
6855	5528.302953	
16500	13498.476233	
16515	16097.608038	
11048	8315.257218	
5572	6630.154608	
16430	19913.408695	
5389	5732.201761	
41315	30548.819502	

Evaluation Results mostra a acurácia do algoritmo. Seu coeficiente de determinação foi de 91%, mostrando-se funcional e com bom desempenho.

### Metrics

Mean Absolute Error	1656.147651
Root Mean Squared Error	2456.983209
Relative Absolute Error	0.276606
Relative Squared Error	0.089608
Coefficient of Determination	0.910392

Esse algoritmo pode ser usado no mundo real para prever o preço de novos modelos. Assim, é útil para, por exemplo, revendedoras de carros.