



**PUC Minas**

DIRETORIA DE  
EDUCAÇÃO CONTINUADA

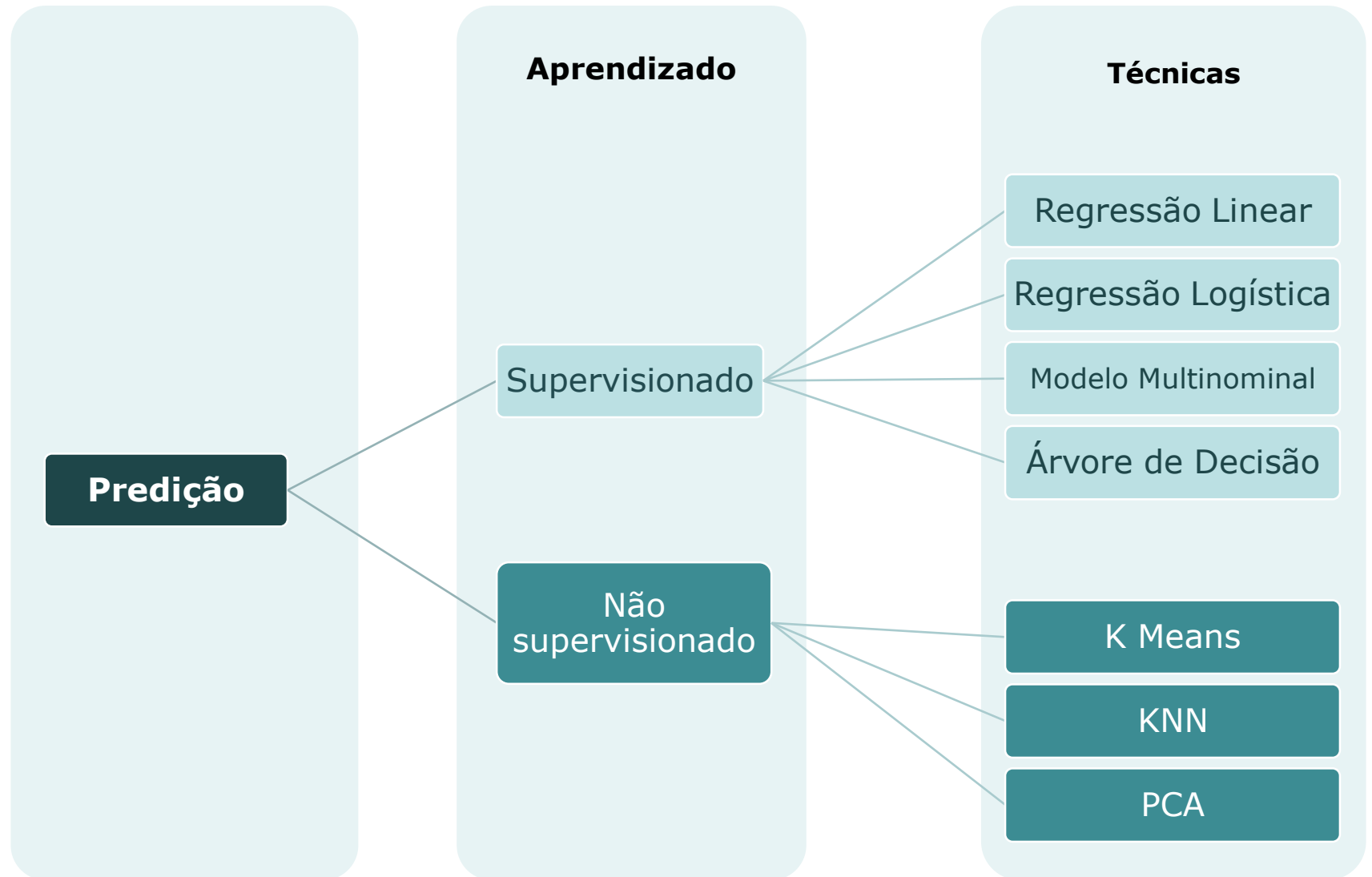
**Pós Graduação *Lato Sensu***

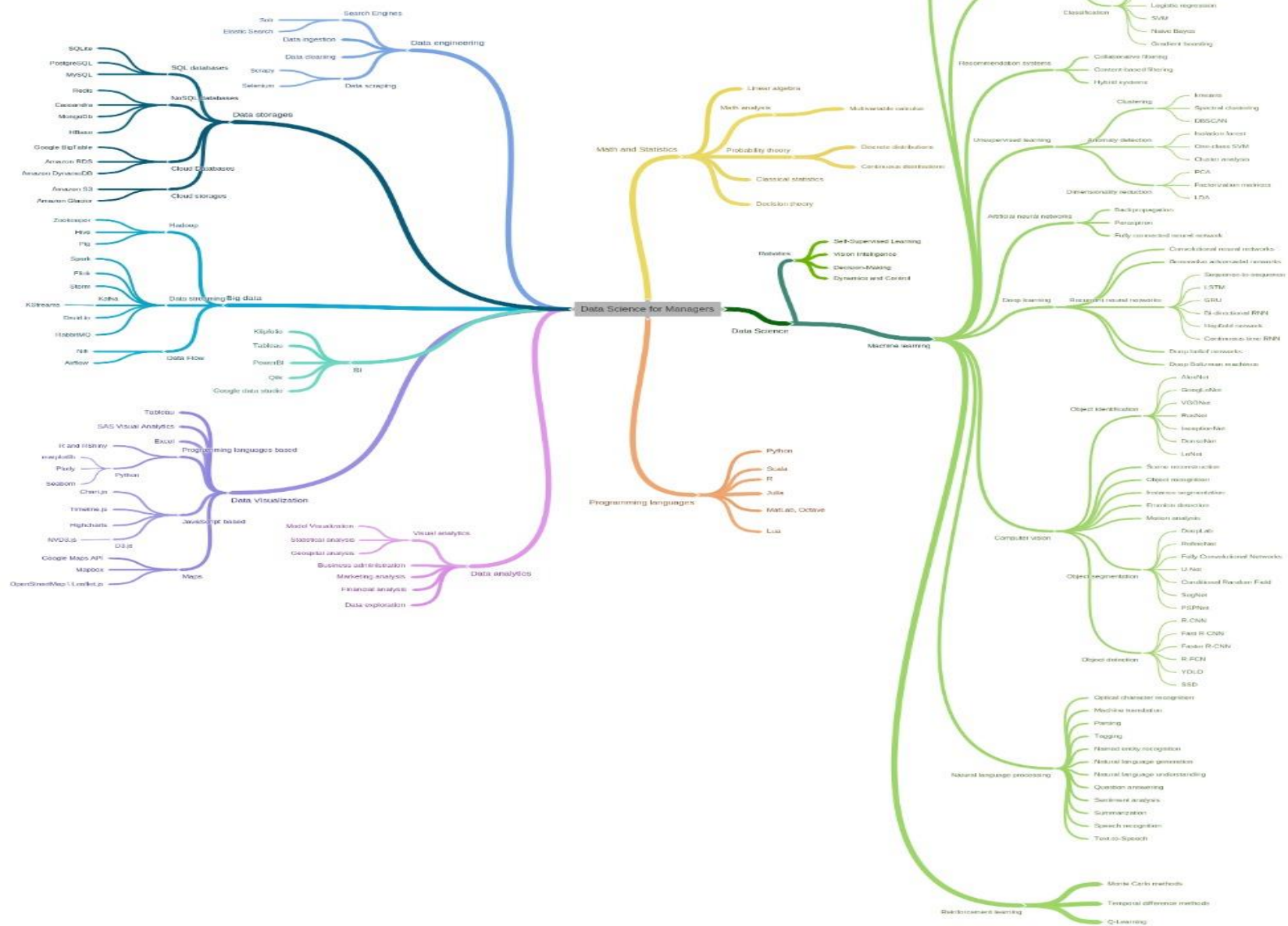
**CIÊNCIA DE DADOS E  
BIG DATA**

**Técnicas Estatísticas de Predição**

## ***Programa***

Calendário	Conteúdo	Referencial Teórico
05/02/2019	Regressão Linear (Simples e Multipla)	An Introduction to Statistical Learning with Applications in R, 2013
12/02/2019	Regressão Logística (Simples e Multipla)	
19/02/2019	Modelo Multinomial	Categorical Data Analysis, 2013
26/02/2019	Árvores de decisão e Florestas Aleatórias	An Introduction to Statistical Learning with Applications in R, 2013
12/03/2019	KNN, PCA e K Means Clustering	
19/03/2019	Avaliação	-



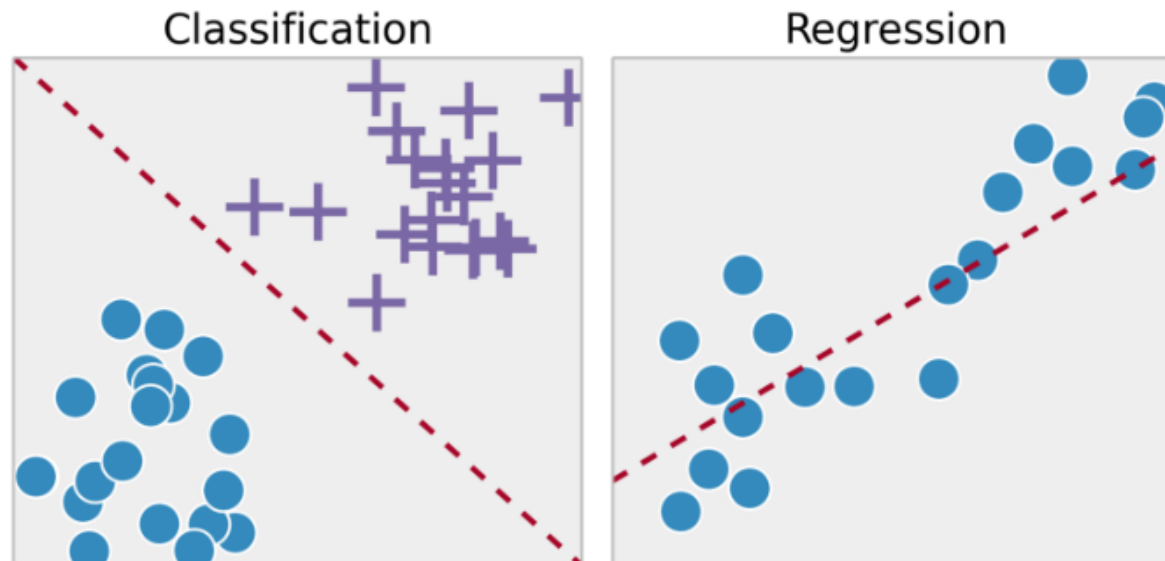


## ***Vantagens em realizar predição***

- Decisões estratégicas
- Identificação de clientes
- Diminuição de custos
- Aumento do retorno sobre o investimento
- Auxiliar na gestão de riscos

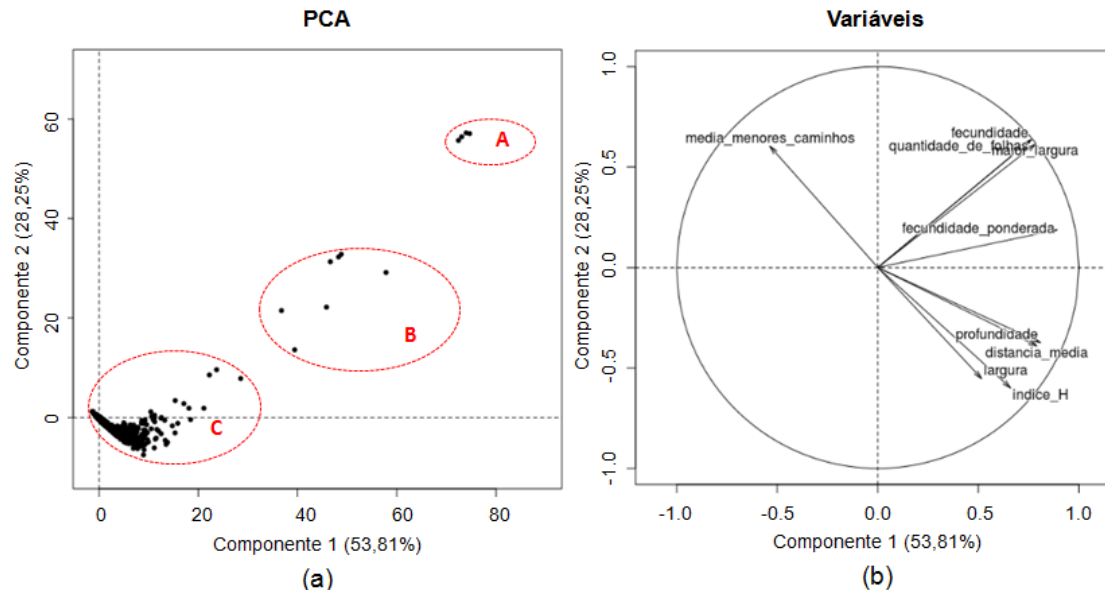
## Aprendizado Supervisionado

Problemas de aprendizagem supervisionados são classificados em problemas de “regressão” e “classificação”.



## Aprendizado Não Supervisionado

O aprendizado não supervisionado, por outro lado, nos permite abordar problemas com pouca ou nenhuma ideia do que como os resultados se apresentarão. Podemos derivar estrutura de dados onde nós não necessariamente conhecemos o efeito das variáveis. Deste modo o **método nos dá a importância das variáveis (features)**.



## ***Aprendizado***

Existe uma ordem para uso do tipo de aprendizado em DS?



## **META**

- Participação em uma competição de **Machine Learning** no *Kaggle*
- Aplicar uma das técnicas de predição utilizadas no curso
- Escrever um artigo sobre modelagem preditiva no *Linkedin*
  - *Causalidade*
  - *Risco*
  - *Identificação*
  - *etc*

## Exercício

**Problema:** Prever preços de casas para regiões nos EUA.

**Dados:** Informações sobre um monte de casas em regiões dos Estados Unidos. tudo está contido no arquivo: USA\_Housing.csv.

Os dados contém as seguintes colunas:

- '**Avg. Area Income**': Média da renda dos residentes de onde a casa está localizada.
- '**Avg. Area House Age**': Média de idade das casas da mesma cidade.
- '**Avg. Area Number of Rooms**': Número médio de quartos para casas na mesma cidade.
- '**Avg. Area Number of Bedrooms**': Número médio de quartos para casas na mesma cidade.
- '**Area Population**': A população da cidade onde a casa está localizada.
- '**Price**': Preço de venda da casa.
- '**Address**': Endereço da casa.

## Exercício

**Instalar Scikit-learn:** biblioteca chave para uso de ML

```
conda install scikit-learn
```

### Importar as bibliotecas

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

### Carregando os dados

```
USAhousing = pd.read_csv('USA_Housing.csv')
```

```
USAhousing.head()
```

### Explorando os dados

```
USAhousing.info()
```

```
USAhousing.describe()
```

### Criando plots

```
sns.pairplot(USAhousing)
```

```
sns.distplot(USAhousing['Price'])
```

```
sns.heatmap(USAhousing.corr())
```



## ***Exercício***

### **Jupyter Notebook**