



*Universidade Federal do Piauí*

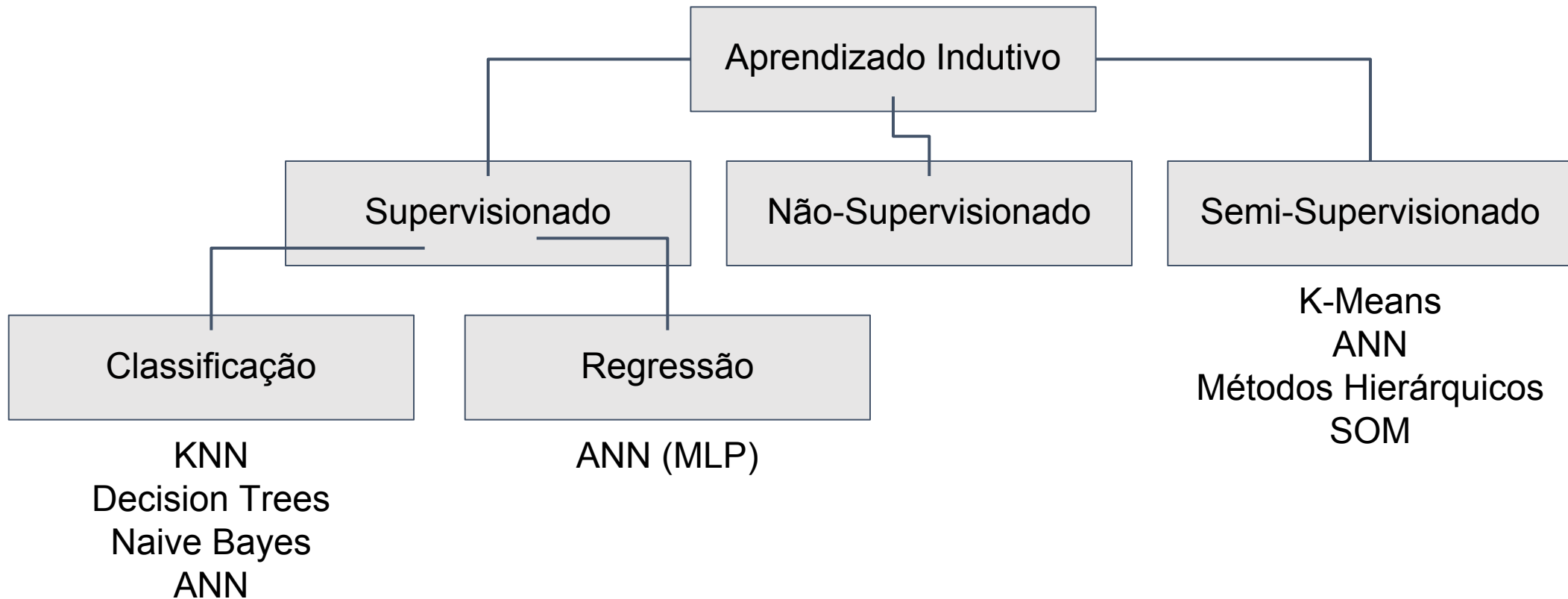
# Introdução ao *Machine Learning* com R

Diego Fernando de Sousa Lima  
Édson Damasceno

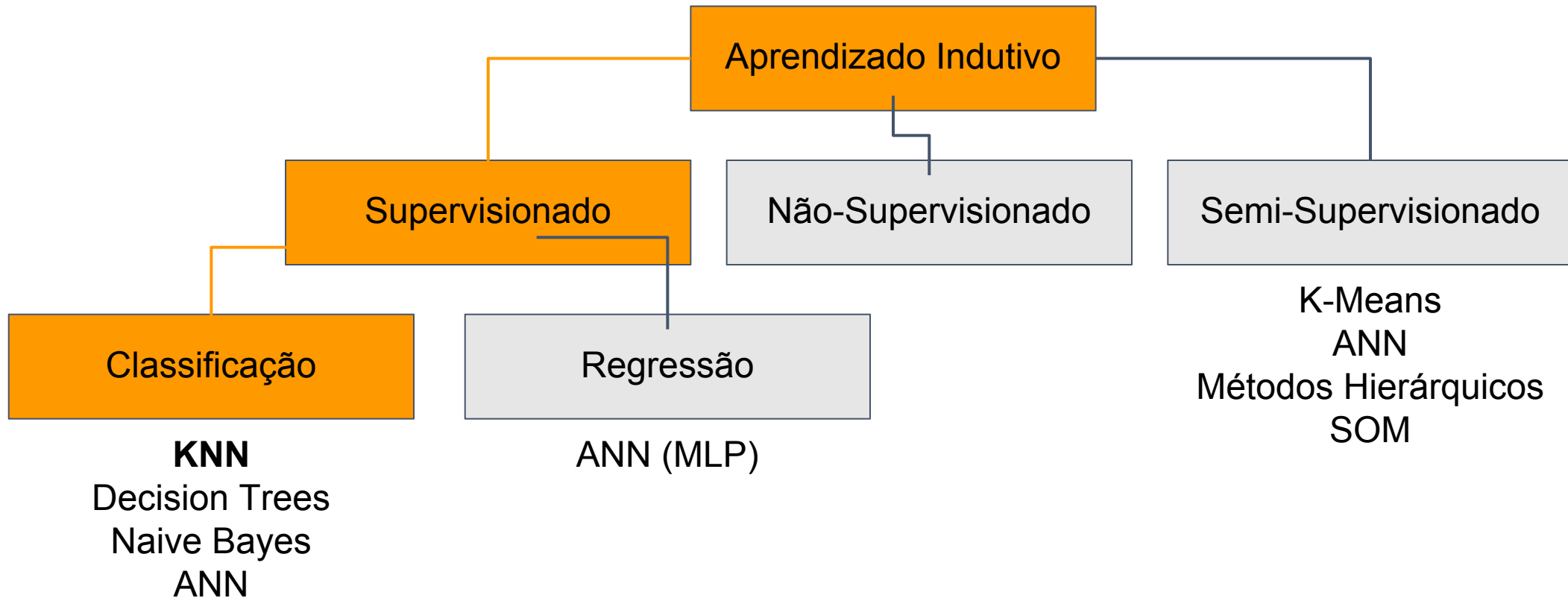
# Introdução

- *O que é Machine Learning*
  - Desenvolvimento de programas que melhoram com a experiência;
  - Resolvem problemas os quais não são solucionáveis por métodos computacionais convencionais ou técnicas baseadas em cálculo;
  - Quando as relações entre todas as variáveis dos sistemas são completamente compreendidas Aprendizagem de máquina não é necessária.

# Tipos de Aprendizado



# Tipos de Aprendizado



# Linguagem R



- R é uma linguagem e também um ambiente de desenvolvimento integrado para cálculos estatísticos e gráficos.
- Foi criada originalmente por Ross Ihaka e por Robert Gentleman no departamento de Estatística da universidade de Auckland, Nova Zelândia, e foi desenvolvido em um esforço colaborativo de pessoas em vários locais do mundo.

# Linguagem R



- Instalação
  - <https://github.com/diegofsousa/RLanguageInstallation>
- Sintaxe
  - *print("hello world")*
  - [http://ecologia.ib.usp.br/bie5782/doku.php?id=bie5782:02\\_tutoriais:tutorial1:start](http://ecologia.ib.usp.br/bie5782/doku.php?id=bie5782:02_tutoriais:tutorial1:start)
  - <https://cran.r-project.org/doc/contrib/Landeiro-Introducao.pdf>

# Pacote *caret*

- O pacote *caret* (*Classification And REgression Training*) é um conjunto de funções que tentam agilizar o processo de criação de modelos preditivos. O pacote contém ferramentas para:
  - *Data splitting* (divisão dos dados);
  - *Pré-processing* (pré-processamento);
  - *Feature Selection* (seleção de características);
  - *Model tuning using resampling* (*model tuning* usando reamostragem);
  - *Variable importance estimation* (Estimativa de importância de variável).

# Pacote *caret*

- Instalação
  - Abra o interpretador R e execute a linha de comando:
    - `install.packages("caret")`
- Documentação
  - <https://topepo.github.io/caret/>



# O que é Predição?

- O dogma da previsão é quando se tem uma situação em que temos um conjunto de dados e queremos classificar cada dado de acordo a extração de características.
- O elemento que tem propriedade para fazer a predição de cada nova amostra é a **Função de Previsão**.

# Etapas da Predição

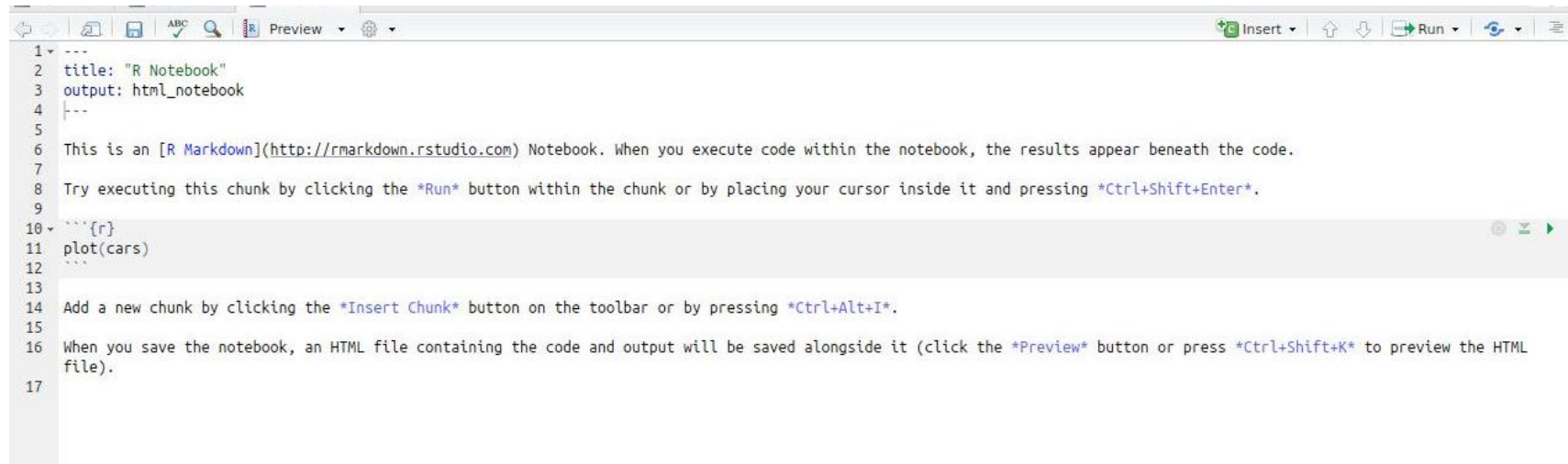
- Questão;
- Entrada de dados;
- Características;
- Algoritmos;
- Parâmetros;
- Avaliação.

---

Hands on!

# Preparação do ambiente

1. Abra o *RStudio*;
2. *New File > R Notebook*;
  - a. Algo como a figura abaixo:



The screenshot shows the RStudio interface with a new R Notebook file open. The notebook contains a code chunk with the following content:

```
1 ---
2 title: "R Notebook"
3 output: html_notebook
4 |---
5
6 This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.
7
8 Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.
9
10 ```{r}
11 plot(cars)
12 ```
13
14 Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.
15
16 When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).
17
```

# Preparação do ambiente

- Alguns atalhos:
  - Ctrl + Alt + I: Insere novo bloco de código;
  - Ctrl + Shift + Enter: Executa bloco de código atual;
  - Ctrl + Shift + K: Pré-visualização HTML.

# Datasets utilizados

- *Statlog (Vehicle Silhouettes) dataset*
  - [https://archive.ics.uci.edu/ml/datasets/Statlog+\(Vehicle+Silhouettes\)](https://archive.ics.uci.edu/ml/datasets/Statlog+(Vehicle+Silhouettes))
  - Descrição: *3D objects within a 2D image by application of an ensemble of shape feature extractors to the 2D silhouettes of the objects.*
  - Classificação: OPEL, SAAB, BUS ou VAN .

# Statlog dataset

- Importação do *dataset*

```
```{r}
dataurl <- "https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/vehicle/xaa.dat"
download.file(url = dataurl, destfile = "wine.data")
vehicles_df <- read.csv("vehicle.data", header = FALSE, sep='') #load data to wine_df dataframe
|
str(vehicles_df) #structure of our data frame
```
```

# Statlog dataset

- Particionamento dos dados

```
```{r}
set.seed(3033)
intrain <- createDataPartition(y = vehicles_df$V1, p= 0.7, list = FALSE)
training <- vehicles_df[intrain,]
testing <- vehicles_df[-intrain,]
cat("Dimensão do dataframe: ", dim(vehicles_df), "\n")
cat("Dimensão do treinamento: ", dim(training), "\n")
cat("Dimensão do teste: ", dim(testing), "\n")
```
```

```
Dimensão do dataframe:  94 19
Dimensão do treinamento: 68 19
Dimensão do teste:  26 19
```



# Statlog dataset

- Resumo da obra com *summary()*

```
```{r}
summary(vehicles_df)
```
```

| V1             | V2            | V3             | V4            | V5             | V6             | V7            | V8            | V9            | V10           |
|----------------|---------------|----------------|---------------|----------------|----------------|---------------|---------------|---------------|---------------|
| Min. : 73.00   | Min. :34.00   | Min. : 51.00   | Min. :105.0   | Min. : 50.00   | Min. : 5.000   | Min. :118.0   | Min. :26.00   | Min. :17.00   | Min. :118.0   |
| 1st Qu.: 87.00 | 1st Qu.:39.25 | 1st Qu.: 68.00 | 1st Qu.:140.2 | 1st Qu.: 55.25 | 1st Qu.: 6.000 | 1st Qu.:146.0 | 1st Qu.:34.00 | 1st Qu.:19.00 | 1st Qu.:135.2 |
| Median : 92.00 | Median :44.00 | Median : 81.50 | Median :166.5 | Median : 62.00 | Median : 8.000 | Median :154.0 | Median :43.00 | Median :19.00 | Median :146.0 |
| Mean : 92.56   | Mean :44.65   | Mean : 81.17   | Mean :168.1   | Mean : 62.59   | Mean : 8.894   | Mean :166.6   | Mean :41.45   | Mean :20.43   | Mean :147.4   |
| 3rd Qu.: 98.00 | 3rd Qu.:49.00 | 3rd Qu.: 95.50 | 3rd Qu.:197.0 | 3rd Qu.: 66.00 | 3rd Qu.:10.000 | 3rd Qu.:192.8 | 3rd Qu.:46.00 | 3rd Qu.:22.00 | 3rd Qu.:159.8 |
| Max. :119.00   | Max. :59.00   | Max. :108.00   | Max. :306.0   | Max. :126.00   | Max. :52.000   | Max. :261.0   | Max. :57.00   | Max. :28.00   | Max. :186.0   |

| V11           | V12           | V13           | V14            | V15            | V16           | V17           | V18           | V19     |
|---------------|---------------|---------------|----------------|----------------|---------------|---------------|---------------|---------|
| Min. :135.0   | Min. :206.0   | Min. :112.0   | Min. : 62.00   | Min. : 0.000   | Min. : 0.00   | Min. :176.0   | Min. :182.0   | bus :26 |
| 1st Qu.:168.0 | 1st Qu.:311.8 | 1st Qu.:148.8 | 1st Qu.: 68.25 | 1st Qu.: 2.000 | 1st Qu.: 5.00 | 1st Qu.:183.0 | 1st Qu.:189.2 | opel:20 |
| Median :175.5 | Median :354.0 | Median :173.5 | Median : 72.00 | Median : 5.000 | Median : 9.00 | Median :188.0 | Median :195.0 | saab:20 |
| Mean :188.2   | Mean :430.4   | Mean :175.5   | Mean : 73.89   | Mean : 5.372   | Mean :10.34   | Mean :188.3   | Mean :194.8   | van :28 |
| 3rd Qu.:217.8 | 3rd Qu.:569.2 | 3rd Qu.:203.5 | 3rd Qu.: 76.00 | 3rd Qu.: 7.750 | 3rd Qu.:14.00 | 3rd Qu.:192.8 | 3rd Qu.:199.8 |         |
| Max. :280.0   | Max. :998.0   | Max. :264.0   | Max. :127.00   | Max. :20.000   | Max. :38.00   | Max. :202.0   | Max. :209.0   |         |

## *Statlog dataset*

- Tornando a coluna 19 como fator (classificador)

```
```{r}
training[["V19"]] = factor(training[["V19"]])
```
```

# Statlog dataset

- Treino

```
#### {r}
trctrl <- trainControl(method = "boot")
set.seed(3333)
knn_fit <- train(V19 ~., data = training, method = "knn",
  trControl=trctrl,
  preProcess = c("center", "scale"),
  tuneLength = 10)
knn_fit
```

# Statlog dataset

- Treino
  - *trainControl*
    - *method = "boot"*
    - <https://topepo.github.io/caret/model-training-and-tuning.html#control>
  - *train*
    - *method = "KNN"*
    - Mais modelos:
      - <https://rdrr.io/cran/caret/man/models.html>

---

## *Statlog dataset*

- KNN?
- k-nearest neighbors ou k-vizinhos próximos.

---

## *Statlog dataset*

- Visualização de resultados
  - `knn_fit`
  - `plot(knn_fit)`

# Statlog dataset

- Usar predição para o conjunto de testes

```
```{r}
test_pred <- predict(knn_fit, newdata = testing)
test_pred
```
```

```
[1] van van bus opel van bus van saab saab van opel van van van van
bus van bus opel saab bus bus opel van van opel
Levels: bus opel saab van
```

---

## *Statlog dataset*

- Estatísticas para o conjunto de testes
  - `confusionMatrix(test_pred, testing$V19)`



# *Statlog dataset*

- Interpretando o *output*
  - Matriz de Confusão
  - Acurácia
  - *Kappa*
  - Sensibilidade
  - Especificidade

---

Conclusão...

---

# Referências

Documentação caret package. Disponível em: <<https://topepo.github.io/caret/>>

RUSSELL, S.; NORVIG, P. INTELIGÊNCIA ARTIFICIAL, 3E.[Sl: sn]. [S.l.], 2013.

Site oficial da Linguagem R. Disponível em: <<https://www.r-project.org/>>