MBA
USP
ESALQ

INTRODUÇÃO AO R E INTRODUÇÃO AO MACHINE LEARNING

Prof. Dr. Wilson Tarantin Junior

*A responsabilidade pela idoneidade, originalidade e licitude dos conteúdos didáticos apresentados, é do professor.

Introdução ao *Machine Learning*

Machine learning

- Definição
  - Podem ser encontradas muitas definições para o conceito *machine learning*
  - Porém, de forma geral, pode-se entender como a utilização dos dados e de algoritmos (que contêm também métodos estatísticos) para a produção de informações que serão relevantes para a tomada de decisão
    - Por exemplo: criando modelos preditivos ou mesmo classificando dados

Banco de dados

- Definição e composição
  - O banco de dados é o objeto onde estão armazenadas as informações de interesse para a análise ou estudo em questão
  - Em muitos casos, o banco de dados contém uma amostra, que nada mais é do que um subconjunto extraído da população
  - O banco de dados é composto por variáveis e por observações
    - **Variáveis**: características/atributos observados, medidos ou categorizados
    - **Observações**: as unidades que têm suas características e atributos medidos

Banco de dados

- Estrutura para uso
  - Normalmente, o banco de dados é estruturado com as variáveis em colunas e as observações em linhas

	Idade	Altura	Cidade	Profissão	Renda	...
Pessoa 1						
Pessoa 2						
Pessoa 3						
Pessoa 4						
Pessoa 5						
Pessoa 6						
Pessoa 7						
Pessoa 8						
Pessoa 9						
Pessoa 10						
...						

MBA USP ESALQ

Tipos de variáveis

- As variáveis podem ser divididas em
  - **Métricas**: são as variáveis **quantitativas**, isto é, apresentam características que podem ser mensuradas ou contadas
  - **Não métricas**: são as variáveis **qualitativas**, sendo que indicam características que não podem ser medidas. Tais variáveis contêm categorias, por isto, muitas vezes, são chamadas de variáveis **categóricas**
  - **A identificação do tipo de variável é fundamental para a escolha do método estatístico que será utilizado na análise dos dados**

Variáveis qualitativas

- Características principais
  - As variáveis qualitativas têm sua representação feita por meio de tabelas de distribuição de frequências ou gráficos
  - Também é possível calcular a **moda** de uma variável qualitativa, definida como a categoria que mais repete na distribuição da variável
    - Não é possível calcular outras medidas de resumo como média ou desvio padrão para variáveis qualitativas
  - Exemplos: escalas likert, nacionalidade, cor do veículo, profissão...

Variáveis quantitativas

- Características principais
  - As variáveis quantitativas podem ser representadas por diversas ferramentas, como gráficos, medidas de posição ou localização, dispersão e de forma. A seguir, alguns exemplos:
    - Medidas de posição ou localização: média, mediana, percentis
    - Medidas de dispersão: desvio padrão
    - Medidas de forma: assimetria e curtose
  - Exemplos: idade, renda em Reais, número de habitantes no município, distância em metros, rentabilidade percentual de uma empresa....

Detalhando as variáveis

- Outras características relevantes
  - Variáveis qualitativas: dicotômica ou policotômica; nominal ou ordinal
    - Dicotômica: duas categorias (binária); Policotômica: mais de duas categorias
    - Nominal: não estabelece relação de grandeza/ordem; Ordinal: estabelece uma ordem
  - Variáveis quantitativas: discretas ou contínuas
    - Discretas: possuem conjunto finito e numerável de valores, em geral, são obtidas a partir de dados de contagem (0, 1, 2, 3, 4, 5...)
    - Contínuas: assumem valores pertencentes ao intervalo de números reais

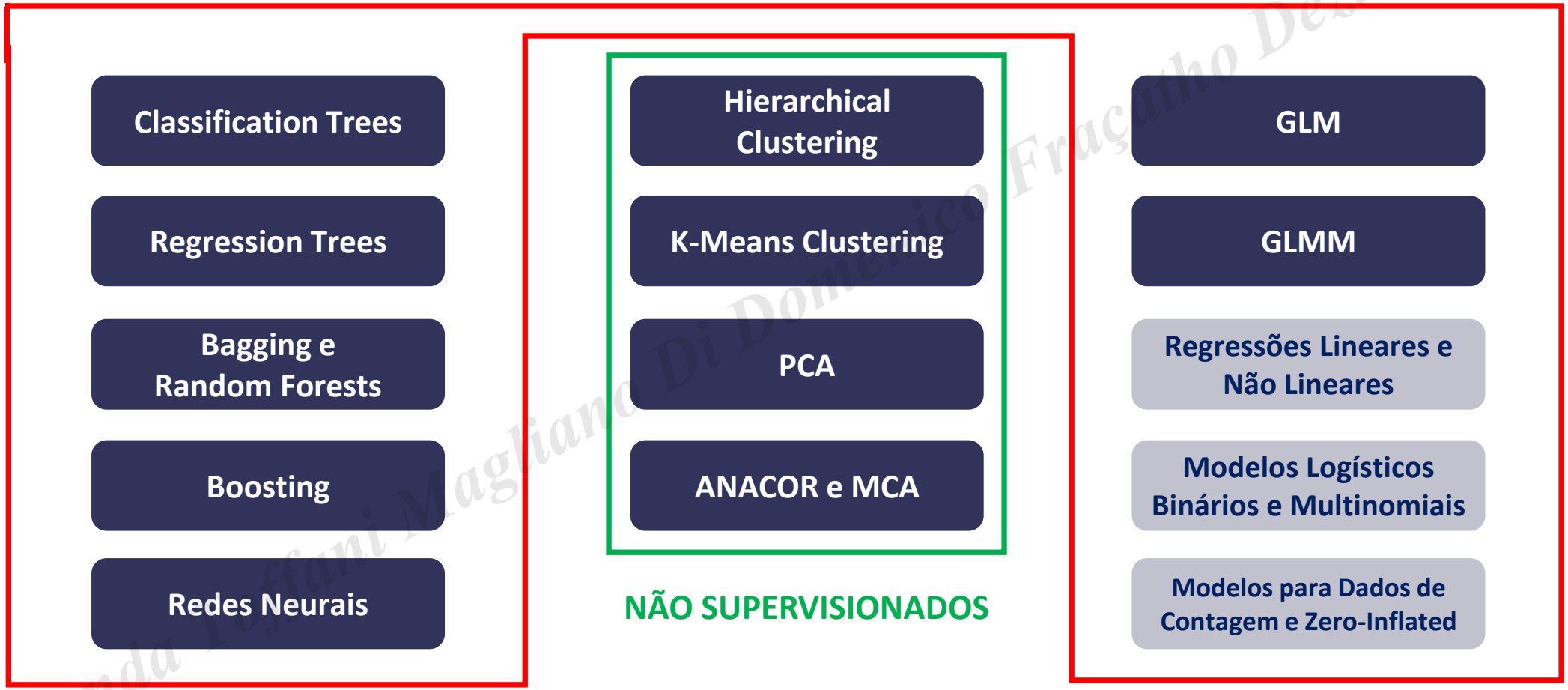
Variáveis e técnicas

- Como identificar a técnica adequada
  - Modelos **não supervisionados** de *machine learning*
    - Também são conhecidos como modelos **exploratórios** ou técnicas de interdependência
    - Estudo da relação entre variáveis, mas sem a intenção de criar modelos confirmatórios
    - Não há inferência dos resultados encontrados para observações fora da amostra
    - Objetivos: redução dos dados, classificação ou agrupamento de observações e variáveis, correlação ou associação entre variáveis

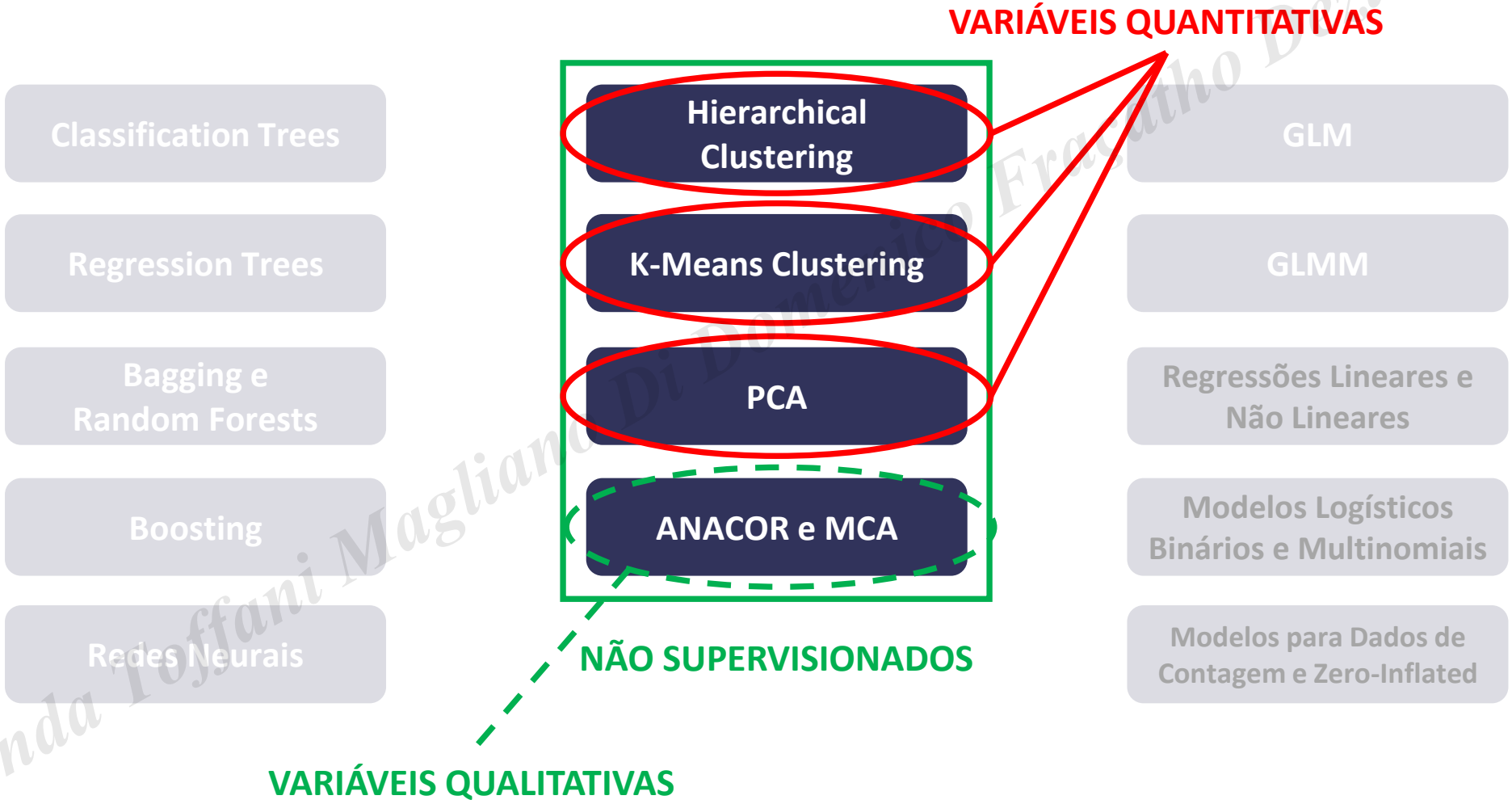
Variáveis e técnicas

- Como identificar a técnica adequada
  - Modelos **supervisionados** de *machine learning*
    - Conhecidos como modelos **confirmatórios** ou técnicas de dependência
    - O objetivo é estimar modelos, equações, com o intuito de **elaborar previsões**
    - Portanto, há **inferência** dos resultados para outras observações fora da amostra

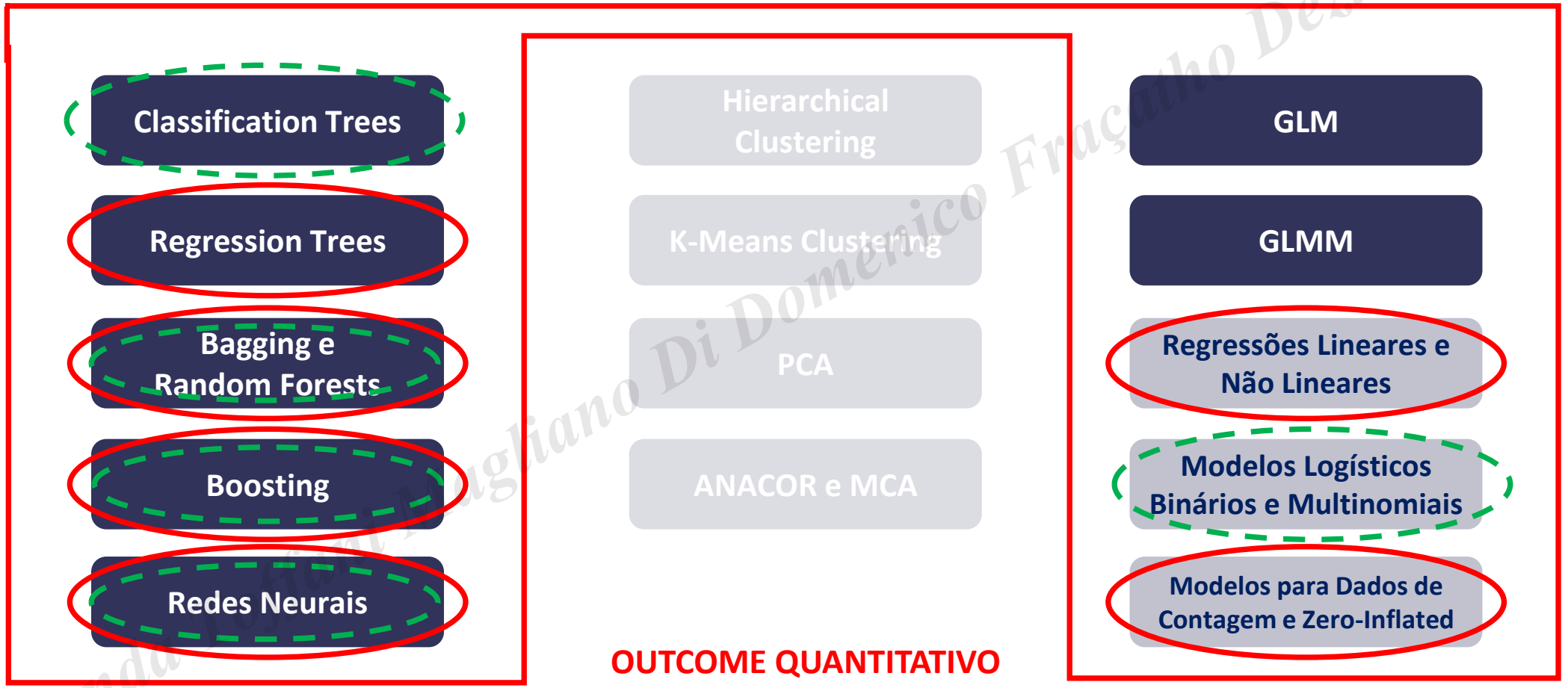
Classificação de algoritmos



Classificação de algoritmos



Classificação de algoritmos



Referências

Fávero, Luiz Paulo; Belfiore, Patrícia. (2017). Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®. Rio de Janeiro: Elsevier

Introdução ao *Software* R

Software R e RStudio

- Apresentação
  - Vamos direto ao RStudio e ao script de aula para conhecermos e aplicarmos os conceitos fundamentais sobre o R!
  - A introdução e apresentação constam no próprio script
  - **Lembrando: primeiramente, sempre vamos descompactar o arquivo (caso esteja compactado) e sempre vamos abrir por meio do arquivo de R Project**
  - **Nos materiais complementares, há um tutorial de instalação dos programas**

Indicações

- Leituras para desenvolvimento no R
  - Hands-On Programming with R (Grolemund, 2014)
  - R for Data Science (Wickham & Grolemund, 2016)
  - Ciência de Dados com R (Guerra, Oliveira, McDonnell & Gonzaga, 2020)

OBRIGADO!

linkedin.com/in/wilson-tarantin-junior-359476190