

DISCIPLINA: Introdução à Análise Exploratória de Dados

Carga horária: 60 horas/aula

Número de Créditos: 4 créditos

Número mínimo de alunos: 5

Número máximo de alunos: 25

OBJETIVOS: - A **Ciência de Dados** é um campo interdisciplinar que utiliza métodos, processos, algoritmos e sistemas científicos para extrair conhecimento e informações valiosas de dados estruturados e não estruturados. Para transformar dados em conhecimento significativo e útil são necessárias ferramentas e habilidades especializadas. Desse modo *ciência de dados* é um conceito amplo, que busca unificar: *técnicas estatísticas, análise de dados, aprendizado de máquina, conhecimento de domínio e ciência da informação* com o intuito de entender e analisar “*fenômenos reais*” com os dados. Um dos primeiros passos para *extrair conhecimento* e informações valiosas de dados (*insight*) a **Análise Exploratória de Dados (EDA – Exploratory Data Analysis)** é uma abordagem estatística para analisar conjuntos de dados (*datasets*) com o objetivo de resumir suas principais características, utilizando para isso métodos de *visualização de dados*. Com todas estas questões em mente, esta disciplina busca oferecer aos alunos dos cursos de Bacharelado em Sistemas de Informação e Computação uma introdução sobre **processo de ciência de dados** e da *abordagem de análise exploratória de dados (AED)* com o intuito de maximizar a percepção sobre um conjunto de dados e resumir suas principais características. No curso também será oferecida uma introdução **ao software de código aberto R** e o **ambiente de desenvolvimento integrado RStudio** com o intuito de aprender a utilizá-los para **transformar dados em conhecimento aplicável**.

METODOLOGIA DE ENSINO: - Aulas teóricas; estudo e análise de textos, livros e artigos sobre o assunto; aulas práticas sobre o uso das ferramentas de software livre **R** e **RStudio** e o uso de dados abertos (*Open Data*).

CRITÉRIOS DE AVALIAÇÃO DE APRENDIZAGEM: - Trabalhos práticos realizados como atividades didáticas durante as aulas e um trabalho final escrito em linguagem “R” para verificação do aprendizado.

CONTEÚDO PROGRAMÁTICO: -

1. Introdução à Ciência de Dados
 - 1.1. Um histórico sobre “Data Science”
2. O que é Ciência de Dados
 - 2.1. Definição de Ciência de Dados
 - 2.2. Áreas de Conhecimento
 - 2.3. Dados e Tomada de Decisão
 - 2.4. O que é “Data Science, afinal?”

- 3. Ciência de Dados e Big Data
 - 3.1. Conceitos de dados e *datasets*
 - 3.2. Big Data
 - 3.3. Ciência de Dados e Big Data
- 4. Ferramentas para Ciência de Dados
 - 4.1 Ambiente de Desenvolvimento Integrado (IDE)
 - 4.1.1. Linguagem R e IDE RStudio
 - 4.1.2 Linguagem Python e GUI Anaconda Navigator
 - 4.2 Sistema de Controle de Versão: Git, GitHub e GitLab
- 5. Ciência de Dados e Estatística
 - 5.1. Revisitando a Estatística: Conceitos básicos
 - 5.2. Ciência de Dados e Estatística
 - 6.3. Ciência de Dados é uma Arte
 - 5.4. Aprendizado de Estatística
- 6. Processo de Ciência de Dados
 - 6.1. Quadro geral do processo de ciência de dados
 - 6.2. Dados, Software Livre e Governos “abertos”
- 7. A Pesquisa do tipo *Survey*
 - 7.1 Características científicas da pesquisa survey
 - 7.2 Finalidades da pesquisa e unidades de análise
 - 7.3 Desenhos básicos de survey
- 8. Introdução e Exemplo de Análise Exploratória de Dados
- 9. Pesquisa Reprodutível
 - 9.1. O que é e o que significa pesquisa reprodutível
 - 9.2. Dados, Software e Governos “abertos”

BIBLIOGRAFIA:

- AMARAL, Fernando. Introdução à Ciência de Dados, Mineração de Dados e Big Data, e-book, Alta Books, 2018.
- FLACH, Peter. Machine Learning: The Art and Science of Algorithms that Make Sense of Data, Cambridge University Press, 2012.
- GANDRUD, Christopher. Reproducible Research with R and RStudio, Third Edition, Chapman and Hall/CRC, 2020.
- GRUS, Joel. Data Science do Zero. Primeiras Regras com o Python, Alta Books, 2018.
- LANTZ, Brett. Machine Learning with R, 3ª Edition, Packt Publishing 2019.
- MAYER-SCHONBERGER, Viktor; CUKIER, Kenneth. Big Data: Como Extrair Volume, Variedade, Velocidade e Valor da Avalanche de Informação Cotidiana, Campus - Grupo Elsevier, 2013.
- PATHAK, Manas A. Beginning Data Science with R, Springer, 2014.

PROVOST, Foster; FAWCETT, Tom. Data Science Para Negócios: O que Você Precisa Saber Sobre Mineração de Dados e Pensamento Analítico de Dados, Alta Books, 2018.

SASHINSKAYA, Maria. Open Data: All You Want To Know About Open Data, CreateSpace Independent Publishing Platform, 2017.

SCHMULLER, Joseph; BATISTA, Samantha. Análise Estatística com R Para Leigos, Editora Alta Books, 2019.

SILVA, Leandro Augusto. Introdução a Mineração de Dados: Com Aplicações em R, Elsevier, 2016.

SINGH, Julian. Open Data 101: The latest trends, challenges and research in government open data, Cooee Press, 2017.

STODDEN, Victoria; LEISCH, Friedrich; PENG, Roger D. Implementing Reproducible Research, Chapman and Hall/CRC, 2014.

WICKHAM, Hadley; GROLEMUND, Garrett; BATISTA, Samantha. R Para Data Science, Editora Alta Books, 2019.

WILLIAMS, Graham. Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery (Use R!), Springer, 2011.

ZUMEL, Nina; MOUNT, John. Practical Data Science with R, 2ª Edition, Manning Publications, 2019.

EMENTA: - Nesta disciplina o aluno deverá obter uma introdução abrangente sobre o que é Ciência de Dados e os possíveis usos desta metodologia de forma geral, e em particular sobre **Análise Exploratória de Dados** (AED) com o intuito de maximizar a percepção sobre um conjunto de dados e resumir suas principais características. Desta forma, a disciplina pretende oferecer aos estudantes, de forma abrangente e introdutória, ferramentas e habilidades especializadas extrair conhecimento e informações dos dados de forma eficaz em projetos de pesquisas e também em suas profissões.