



UNIVERSIDADE ESTADUAL DE CAMPINAS

PROGRAMAS E BIBLIOGRAFIAS

2º período letivo de 2023



Disciplina:

HZ133 A - Tópicos Especiais em Demografia I: Introdução à Ciência de Dados

Ementa:

Introduzir os estudantes ao arcabouço conceitual, exemplos e a algumas técnicas utilizadas na área de ciências de dados para que possam aplicar as novas técnicas e ferramentas estudadas em problemas reais, utilizando para isso a linguagem de programação R e suas respectivas bibliotecas mais comumente utilizadas para este tipo de análise.

Programa:

Introdução do curso
Conceito de Big Data, Data Mining, Data Science e aplicações
Tipos e formatos de banco de dados.
Importando banco de dados, pré-processamento, manipulação de conjuntos de dados, transformação e harmonização de dados
Mineração de dados
Visualizações de dados com o ggplot2
Introdução aos modelos de Machine Learning
Construção de modelos de classificação (Random Forest, XGBoost e Logistic Regression)

Bibliografia:

ALCOFORADO, L.F. Utilizando a Linguagem R: conceitos, manipulação, visualização, modelagem e elaboração de relatórios. Rio de Janeiro: Alta Books, 2021. 384p.

BURGER, S.V. Introduction to Machine Learning with R: Rigorous Mathematical Analysis. Beijing: O'Reilly, 2018.

BUSSAB, W.O. e MORETTIN, P.A. Estatística Básica. 5ª Edição. São Paulo: Atual, 2003.

IZBICKI, R. E.; SANTOS, T. M. DOS. Aprendizado de máquina: uma abordagem estatística. 1ª edição. 2020. 272 páginas. ISBN: 978-65-00-02410-4.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. An Introduction to Statistical Learning with Applications in R. New York, NY: Springer, 2021.



KUHN, M.; JOHNSON, K. Applied predictive modeling. New York, NY: Springer, 2013.

KUHN, M. Building predictive models in R using caret package. Journal of Statistical Software, v. 28, n. 5, p. 1-26, 2008.

***LEVIN, J. e FOX, J.A. Estatística para Ciências Humanas. 9ª Edição. Editora Pearson Prentice-Hall, 2004.**

R4DS, WICKHAM, H. (2017): <https://r4ds.had.co.nz/>

R for grad students, Y. Wendy Huynh (2019): https://bookdown.org/yih_huynh/Guideto-R-Book/

<https://clauswilke.com/dataviz/>

<https://socviz.co/>

Materiais adicionais serão disponibilizados no GitHub da professora e inseridos no Moodle ao longo do curso.

Observações:

É desejável o conhecimento da estatística básica: medidas descritivas, distribuições e teste de hipótese. Os conhecimentos da Estatística serão utilizados ao longo de todo o curso. O curso será realizado em R e é desejável o conhecimento básico em R, assim como uma motivação para aprender novas habilidades em R. Todo o curso será desenvolvido no formato teórico-prático.

Os alunos deverão ter disponibilidade para estudar os conteúdos antes das aulas.

AVALIAÇÃO

A avaliação consiste de:

- 1) Exercícios semanais, com entregas na aula subsequente;
- 2) Apresentação de um artigo que utiliza alguma ferramenta aprendida no curso em um problema da área de ciências sociais;
- 3) Um trabalho/projeto de aplicação em um problema real da sociedade.

Distribuição das Notas

Exercícios: 60 pontos

Apresentação de um artigo: 10 pontos

Apresentação do trabalho/projeto: 30 pontos

A nota final será obtida pelo somatório de todas as notas dividida por 10. O ponto de corte para



aprovação é 5 pontos.

Será considerado aprovado o estudante que obtiver nota e frequência de pelo menos 75% do curso. Exceções serão avaliadas de acordo com o artigo 72 do Regimento Geral dos Cursos de Graduação.

Professora:

Luciana Correia Alves (lcalves@unicamp.br). O atendimento individual deverá ser marcado com o professor por email.

Demais informações:

- Os exercícios serão realizados individualmente.
- Usaremos a plataforma Moodle como repositório do material
- As aulas semanais terão uma parte teórica e prática, mas no geral o curso será predominantemente prático
- Os exercícios deverão ser entregues na data estabelecida por meio do Moodle. Um relatório de entrega é gerado para o professor.
- A cada aula os alunos terão uma série de atividades/exercícios para serem desenvolvidas em aula e ao longo da semana, com entrega na aula subsequente. Serão exercícios a partir de banco de dados que cada um irá escolher. São bancos de aula prática. Se o estudante tiver o seu próprio banco de dados também poderá utilizar.
- O objetivo do curso será aprender praticando. Essas atividades tem a finalidade de ensinar o estudante a fixar o conteúdo e a praticar as análises no software de forma simultânea. O curso será dado integralmente no RStudio.
- No trabalho final os estudantes deverão construir uma proposta/projeto de aplicação de um dos conteúdos aprendidos em problemas reais. A apresentação será no último dia de aula.