FGV EMAp
ESCOLA DE MATEMÁTICA APLICADA

CURSO: Ciência de Dados e Inteligência Artificial – 1º semestre de 2023
DISCIPLINA: **Técnicas e Algoritmos em Ciência de Dados**
PROFESSOR(ES): Alberto Paccanaro
CARGA HORÁRIA: 60h
PRÉ-REQUISITO: Linguagens de Programação e Análise Exploratória de Dados e Visualização

# PLANO DE ENSINO

## 1. Ementa

Introdução: IA, Aprendizado de Máquina, Ciência de Dados; Problemas de classificação e regressão; Métodos para avaliar o desempenho de generalização; Regressão linear; Modelos lineares regularizados; Regressão logística; Redes neurais; Árvores de decisão; Ensemble learning e florestas aleatórias; Técnicas de aprendizagem não supervisionadas, Clustering; Técnicas para redução de dimensionalidade.
Ferramentas para Computação Científica e Ciência de Dados: numpy, pandas, scikit-learn. Em geral, implementação dos algoritmos acima em python.

## 2. Objetivos da disciplina

Esta disciplina tem como objetivo geral fornecer uma visão geral das principais ideias e técnicas de Aprendizado de Máquina. Ele irá abranger os fundamentos matemáticos dos algoritmos, bem como sua implementação prática e aplicação a conjuntos de dados do mundo real. Especificamente, ao final do curso, os alunos devem ser capazes de:

- desenvolver, validar, avaliar e usar efetivamente modelos de aprendizado de máquina
- aplicar métodos e técnicas como Regressão linear, Regressão logística, Redes neurais, Árvores de decisão, e florestas aleatórias para dados do mundo real

## 3. Procedimentos de ensino (metodologia)

O curso tem duas aulas por semana.
A cada semana, durante a primeira aula, irei abordar novos algoritmos e seus fundamentos teóricos e matemáticos. Fornecerei aos alunos os slides que usarei em sala de aula e os indicarei os capítulos de livros relevantes e o material onde o conteúdo pode ser encontrado.
A segunda aula de cada semana será uma aula de laboratório, onde os alunos aprenderão a programar esses algoritmos em python e aplicá-los a conjuntos de dados do mundo real. Estas atividades de programação terão como objetivo consolidar os conhecimentos apresentados na aula teórica.
Todo o material mencionado acima será disponibilizado na plataforma eClass.

## 4. Conteúdo programático detalhado

*Observação: este horário é provisório e o conteúdo das aulas pode precisar ser modificado durante o curso.*

| DATA | | TÓPICO | ATIVIDADE |
|---|---|---|---|
| Terça-feira | 14-Feb | Introdução ao curso. Por que aprendizado de máquina? | aula |
| Quinta-feira | 16-Feb | Python Test; Palestra: Aprendizado de máquina para Biologia, Medicina e Farmacologia | lab |
| Terça-feira | 21-Feb | Feriado | |
| Quinta-feira | 23-Feb | Feriado | |
| Terça-feira | 28-Feb | Taxonomia da IA. | aula |
| Quinta-feira | 02-Mar | Tutorial: numpy; Revisão e aplicação prática do numpy | lab |
| Terça-feira | 07-Mar | Introdução aos conceitos gerais. Avaliação de desempenho. (parte 1) | aula |
| Quinta-feira | 09-Mar | Tutorial: pandas; Revisão e aplicação prática do pandas e data parsing | lab |
| Terça-feira | 14-Mar | Introdução aos conceitos gerais. Avaliação de desempenho. (parte 2), o algoritmo KNN | aula |
| Quinta-feira | 16-Mar | Tutorial: Introdução ao sklearn; aplicação prática do sklearn | lab |
| Terça-feira | 21-Mar | Regressão linear, gradiente descendente | aula |
| Quinta-feira | 23-Mar | Implementação de Regressão linear, gradiente descendente | lab |
| Terça-feira | 28-Mar | Classificação linear, discriminante linear de Fisher, perceptron | aula |
| Quinta-feira | 30-Mar | Exercícios de implementação sobre problemas de Classificação linear, discr. Lin. de Fisher, percept | lab |
| Terça-feira | 04-Apr | Regressão logistica | aula |
| Quinta-feira | 06-Apr | Feriado | |
| Terça-feira | 11-Apr | Revisão | aula |
| Quinta-feira | 13-Apr | Implementação e aplicação de regressão logistica | lab |
| Terça-feira | 18-Apr | A1 | |
| Quinta-feira | 20-Apr | A1 | |
| Terça-feira | 25-Apr | Redes neurais (parte 1) | aula |
| Quinta-feira | 27-Apr | Implementação e aplicação de redes neurais (backpropagation, forward pass) | lab |
| Terça-feira | 02-May | Redes neurais (parte 2) | aula |
| Quinta-feira | 04-May | Implementação e aplicação de redes neurais (backpropagation, backward pass) | lab |
| Terça-feira | 09-May | Árvores de decisão | aula |
| Quinta-feira | 11-May | Implementação e aplicação de árvores de decisão | lab |
| Terça-feira | 16-May | Ensemble methods: Bagging, Boosting, Random Forests | aula |
| Quinta-feira | 18-May | Implementação e aplicação de Bagging, Boosting, Random Forests | lab |
| Terça-feira | 23-May | Clustering (part 1) | aula |
| Quinta-feira | 25-May | Implementação e aplicação de clustering | lab |
| Terça-feira | 30-May | Clustering (part 2) | aula |
| Quinta-feira | 01-Jun | Implementação e aplicação de clustering | lab |
| Terça-feira | 06-Jun | Técnicas para redução de dimensionalidade | aula |
| Quinta-feira | 08-Jun | Feriado | |
| Terça-feira | 13-Jun | Revisão | aula |
| Quinta-feira | 15-Jun | TUTORIAL: autodiff backprop | lab |
| Terça-feira | 20-Jun | A2 | |
| Quinta-feira | 22-Jun | A2 | |

## 5. Procedimentos de avaliação

O conhecimento dos fundamentos teóricos e matemáticos dos algoritmos de aprendizagem de máquina será avaliado através de dois exames, um durante A1 e outro durante A2. Cada exame valerá 30% da nota final.

A capacidade dos alunos de implementar e aplicar os algoritmos em problemas do mundo real será avaliada por meio de quatro trabalhos individuais que exigirão que os alunos enviem código python que implementa todo o fluxo de trabalho de aprendizado de máquina (carregamento de dados, pré-processamento, treinamento, avaliação, etc.). Cada trabalho individual valerá 10% da nota final. Estas são as datas principais para os trabalhos individuais:

| | ENTREGA A ALUNOS | PRAZO | FEEDBACK FORNECIDO EM |
|---|---|---|---|
| **trabalho individual 1** | 16/3 | 23/3 | 30/3 |
| **trabalho individual 2** | 30/3 | 6/4 | 11/4 |
| **trabalho individual 3** | 4/5 | 11/5 | 18/5 |
| **trabalho individual 4** | 1/5 | 9/6 | 16/6 |

## 6. Bibliografia Obrigatória

Aurelien Geron
Hands–On Machine Learning with Scikit–Learn and TensorFlow
Concepts, Tools, and Techniques to Build Intelligent Systems
2ª edição, O'Reilly, outubro de 2019

Christopher Bishop
Pattern Recognition and Machine Learning
Springer, 2006
Disponível em: https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

## 7. Bibliografia Complementar

Stormy Attaway
MATLAB: A Practical Introduction to Programming and Problem Solving
5th Edition, Elsevier, 2018

James, G., Witten, D., Hastie, T., Tibshirani, R
An Introduction to Statistical Learning with Applications in R
Disponível em: https://web.stanford.edu/~hastie/Papers/ESLII.pdf

Wes McKinney. Python for Data Analysis O'Reilly Media.

The Scipy community. Numpy Manual. 2017
https://docs.scipy.org/doc/numpy-1.13.0/contents.html

Dan Saber. A Dramatic Tour through Python's Data Visualization Landscape (including ggplot and Altair)
https://dsaber.com/2016/10/02/a-dramatic-tour-through-pythons-data-visualization-landscape-including-ggplot-and-altair/

Jupyter Team. The Jupyter notebook User Documentation
https://jupyter-notebook.readthedocs.io/en/stable/

An introduction to machine learning with scikit-learn — scikit-learn 0.24.1 documentation.
https://scikit-learn.org/stable/tutorial/basic/tutorial.html.

User guide and tutorial — seaborn 0.11.1 documentation.
https://seaborn.pydata.org/tutorial.html.

## 8. Minicurrículo do(s) Professor(s)

**Alberto Paccanaro** Sou Professor Titular da Escola de Matemática Aplicada (EMAp) da FGV do Rio de Janeiro, onde ingressei em 2020. Obtive meu doutorado em Ciência da Computação em 2002 pela Universidade de Toronto, com especialização em Aprendizado de Máquina sob orientação de Geoffrey Hinton . Entre 2002 e 2006, fiz pós-doutorado em Biologia Computacional, primeiro no laboratório de Mansoor Saqi na Queen Mary University of London, e depois no laboratório de Mark Gerstein na Yale University. Tornei-me PI em 2006, obtendo o cargo de professor na Royal Holloway University of London, onde comecei meu laboratório (www.paccanarolab.org). Em 2014 tornei-me Professor Titular de Aprendizagem de Máquina e Biologia Computacional e Diretor do Centro de Sistemas e Biologia Sintética, da mesma Universidade. Sou professor visitante da Universidade Católica de Assunção (Paraguai), onde conduzo um posto avançado de meu laboratório. Também fui professor / membro visitante em

Cornell, Yale e na Universidade de Veneza. Sou responsável por várias colaborações internacionais na área de Aprendizagem de Máquina aplicada à Biologia e Medicina. Eu co-dirijo bolsas de pesquisa junto com acadêmicos da Yale University, Cornell University, University of Tennessee e da Catholic University of Asuncion. Vários de meus algoritmos de aprendizado de máquina foram publicados em revistas como Nature, Nature Methods, Nature Communications, Cell, PNAS.

**Rubén Jiménez** (monitor do curso) Possui graduação em Ingeniería en Informática pela Universidad Católica Nuestra Señora de la Asunción (2017) e doutorado em Computer Science (Bioinformatics) pela Royal Holloway, University of London (2022). Tem experiência na área de biotecnologia e aprendizado de máquina, com especialização em aprendizado profundo e bioinformatica.

## 9. Link para o Currículo Lattes

Prof. Alberto Paccanaro: http://lattes.cnpq.br/9819989502690120