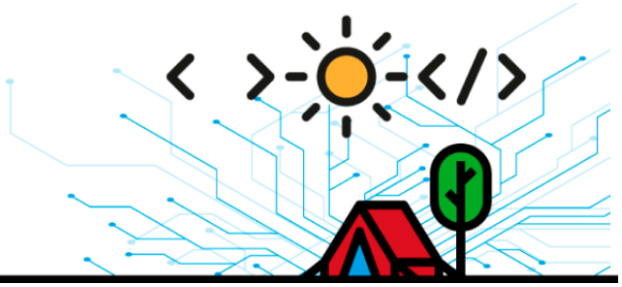
Enap

coding
BOOTCAMP



Bootcamp Machine Learning – Turma 1/21

Curso: Bootcamp Machine Learning
Docentes: Erick Muzart Fonseca dos Santos e Fernando Melo
Período:16 de novembro a 15 de dezembro
Horário: 8h – 12h, 14h30 - 17h30
Carga Horária: 130h

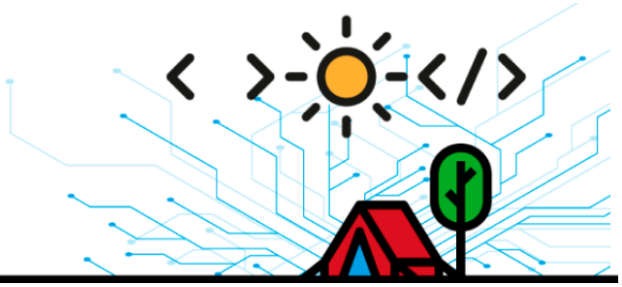
Objetivo / Competência:

O objetivo deste módulo é ensinar os fundamentos de *machine learning* (ML, ou aprendizado de máquina) em que, em vez do programador descrever explicitamente os procedimentos a serem realizados para se criar o resultado esperado, são fornecidos exemplos de resultados e o próprio algoritmo de aprendizado mapeia o padrão de relações entre os dados de entrada e o resultado esperado, realizando assim previsões para novos casos ainda não encontrados. Por exemplo, dispondo de dados de desempenho passado de alunos no Enem, e de metadados que descrevem esses alunos, é possível treinar um modelo para prever o desempenho esperado de futuros alunos, permitindo conceber intervenções personalizadas e suportar decisões apoiadas em dados.

O participante aprenderá a identificar oportunidades de uso e como aplicar técnicas de ML para descobrir padrões em seus próprios dados, construir modelos preditivos para estimar alguma variável de interesse em função dos demais dados disponíveis e melhorar a compreensão do fenômeno subjacente à geração dos dados, para apoio a decisão e otimização de resultados.

Ementa:

1. Diferenças entre programação tradicional e aprendizado de máquina (ML): O desafio de prever resultados de um fenômeno sem um modelo explícito de seu funcionamento.
2. Categorias de *machine learning* (ML): Supervisionado, não supervisionado e por reforço.



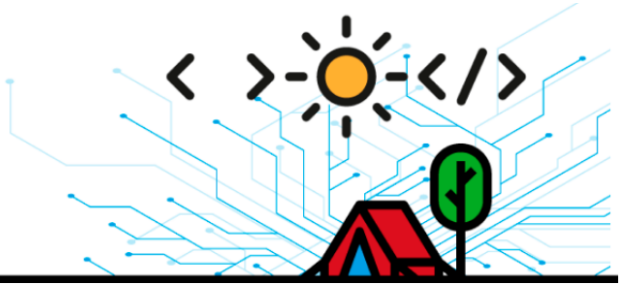
3. Diversidade de aplicações.
4. Desafio: estimativa de preço de imóveis em função de suas características.
5. Regressão linear:
    - Intuição, cenários de uso.
    - Preparação de dados para modelização.
    - Uso da biblioteca python scikit-learn.
6. Conceitos gerais de ML, aplicados ao caso básico de regressão linear:
    - particionamento dos dados em treinamento/teste/validação;
    - Over e under fitting;
    - Determinantes de desempenho: mais dados, controle de complexidade do modelo, regularização, data augmentation; otimização e gradiente descendente.
7. Regressão logística: Extensão da regressão linear como primeiro classificador.
8. Árvore de decisão:
    - Construção e interpretação.
    - Extensão para *Random Forest*, aplicação sistemática em tarefas de previsão.
    - Desafio: prever sobreviventes do naufrágio do Titanic.
9. Visão geral de redes neurais: conceito, modelos pré-treinados, aplicações em dados tabulares e processamento de linguagem natural (NLP), com vetorização de palavras e categorias.

Metodologia de Ensino:

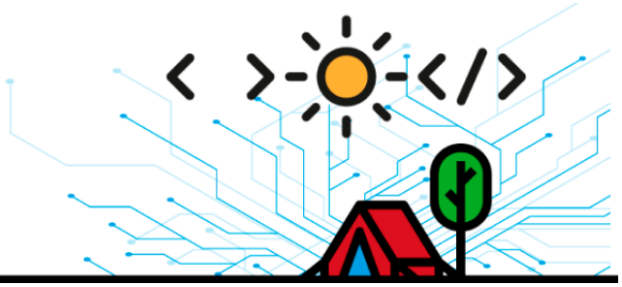
Alternância entre conteúdo expositivo curto, exercícios com codificação para consolidar o domínio das técnicas apresentadas e aplicações em novos conjuntos de dados, de forma guiada, para facilitar a experimentação das técnicas sobre dados reais e o ganho de autonomia do aluno.

Avaliação da Aprendizagem:

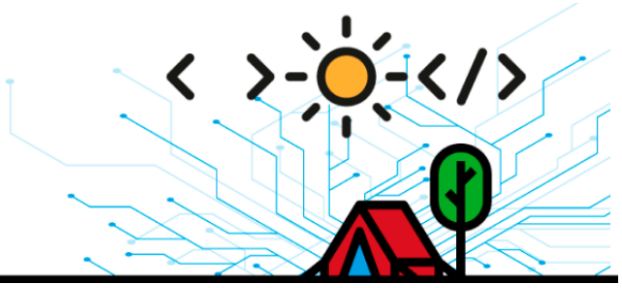
Avaliação exclusivamente pela qualidade do projeto apresentado ao final do curso e desenvolvido progressivamente ao longo do conteúdo, com suporte de tutoria.



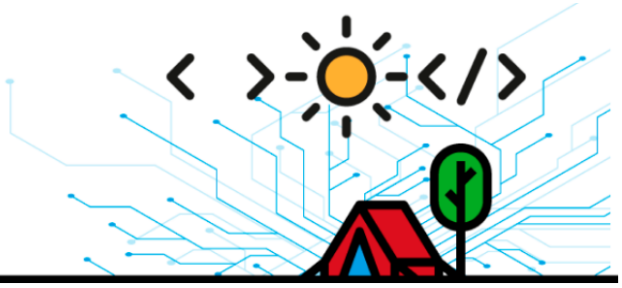
Aprovação com a demonstração de elementos mínimos do uso das técnicas apresentadas durante e curso, com pontuação extra para usos mais avançados, originalidade ou relevância da solução desenvolvida.



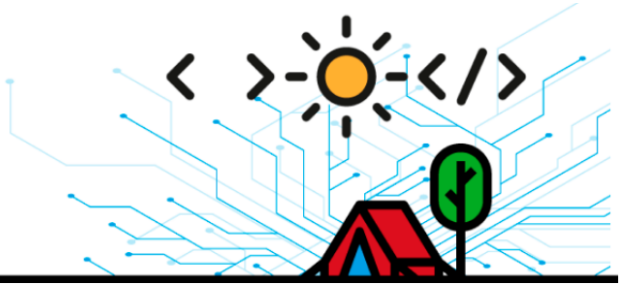
PLANO DE AULA:	
1º Período	
	Contextualização de ML com um desafio: estimativa de preço de imóveis em função de suas características. Entendendo um problema de modelagem preditiva. Possíveis abordagens e soluções. Amostra de código para modelagem (regressão linear e knn).
2º Período	
	Visão geral do que é machine learning (ML) e diferenças com programação explícita. Principais conceitos. Tipos de ML. Aplicações. Identificando oportunidades de uso de ML na Administração. Pré-requisitos. Disponibilidade de dados. Problemas bem formulados.
3º Período	
	Aprendizado não supervisionado. Clusterização com algoritmos simples: DBSCAN e K-Means Exemplos e exercícios. Aplicação para rotulagem de dados.
4º Período	
	Regressão linear: intuição, cenários de uso. Uso da biblioteca python scikit-learn. Diversas aplicações e exercícios. Preparação de dados para modelização.
5º Período	
	Conceitos gerais de ML, aplicados ao caso básico de regressão linear: particionamento dos dados em treinamento/teste/validação; Over e under fitting;
6º Período	
	Determinantes de desempenho: mais dados, controle de complexidade do modelo, regularização, *data augmentation*; otimização e gradiente descendente.
7º Período	



	Regressão logística: Extensão da regressão linear como primeiro classificador. Aplicando classificação para prever sobreviventes do Titanic.
8º Período	
	Codificação de dados categóricos: ordinal encoder e one-hot encoder. Trade offs e codificações mais complexas.
9º Período	
	Bases de Processamento de Linguagem Natural (NLP) Pré-processamento, tokenização, bag of words, tf-idf. Aplicações em classificação.
10º Período	
	Árvore de decisão: Construção e interpretação. Aplicação de árvore de decisão: Titanic Apresentação do conjunto de dados do Enem e modelagem..
11º Período	
	Extensão de árvore de decisão para *Random Forest*: aplicação sistemática em tarefas de previsão. Aplicação de Random Forest.
12º Período	
	Random Forest: interpretação de resultados e desempenho. Aplicação em datasets já explorados, regressão e classificação; comparação de desempenho; novo dataset.
13º Período	
	Ensemble de modelos, pipeline, otimização de hiperparâmetros. Interpretação de resultados de modelagem, comparação e seleção de melhores modelos em função de suas características.
14º Período	
	Visão geral de redes neurais: conceito, modelos pré-treinados, aplicações.



	Biblioteca fast.ai & pytorch para uso de redes neurais profundas, Deep Learning.
15º Período	
	Deep Learning para dados tabulares.
16º Período	
	Deep Learning em NLP: ULMFiT e BERTimbau
17º Período	
	Trade offs entre modelos simples e modelos complexos. Quantidade de dados de treinamento, complexidade da tarefa. Gradient Boosted Machine: XGBoost
18º Período	
	Avaliação de modelos e interpretação dos resultados. Importância das variáveis por permutação (*Permutation feature importance*). Explicabilidade de modelos.
19º Período	
	Operacionalização de modelos (*deploy*). Modelos em nuvem: Voila e heroku Manutenção de modelos em produção. MLOps. AutoML. Considerações éticas de ML: vieses, transparência, dados privados
20º Período	
	Como apresentar seus resultados de ciência de dados e modelagem preditiva: valorizar o conteúdo e resultados do projeto, construir confiança da qualidade dos resultados.
21º Período	
	Revisão e ajustes finais dos projetos
22º Período	



	Apresentação dos projetos, comentários e discussão com com a turma acerca das características de cada projeto; avaliação. Encerramento

Bibliografia Básica:

1. **Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2**, de Sebastian Raschka:
https://www.goodreads.com/book/show/47969269-python-machine-learning
Principal livro de referência, com código python e explicações intuitivas.

2. **Deep Learning for Coders with fastai and PyTorch: AI Applications Without a PhD**, de Jeremy Howard e Sylvain Gugger:
https://www.amazon.com.br/Deep-Learning-Coders-fastai-PyTorch-ebook/dp/B08C2KM7NR

3. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, de Aurélien Géron:
https://www.goodreads.com/en/book/show/40363665
Abordagem alternativa, cobrindo das bases de ML até Deep Learning.

Bibliografia Complementar:

- Módulos do curso de ciência de dados da plataforma Dataquest:
https://www.dataquest.io/path/data-scientist/
Há diversos módulos em python, voltados para desenvolvimento de código e aplicação prática, úteis para ganhar experiência nas bibliotecas mais utilizadas: Machine Learning Fundamentals, Calculus for Machine Learning, Linear Algebra for Machine Learning, Linear Regression for Machine Learning, Decision Trees, Deep Learning: Fundamentals, Natural Language Processing.

- Introduction to Machine Learning with Python: A Guide for Data Scientists, de Andreas C. Müller, Sarah
Guido: https://www.goodreads.com/book/show/24346909-introduction-to-machine-learning-with-python
Aprofunda o uso da biblioteca python scikit-learn.

- Data Science for Business: What you need to know about data mining and data-analytic thinking, de Foster Provost, Tom Fawcett:
https://www.goodreads.com/book/show/17912916-data-science-for-business

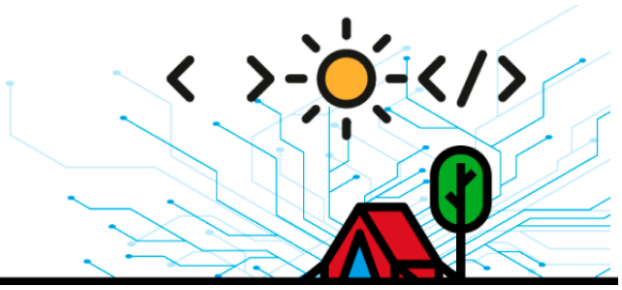
Livro mais conceitual, de aplicações de modelagem preditiva nos negócios e sobre a comunicação entre as áreas de negócio e as áreas técnicas de desenvolvimento. Também existe edição em português.

- Palestras e aulas de Deep Learning ministradas pelos instrutores:
https://www.youtube.com/channel/UC8ORwJ1BlleYth5uaW8TF2A/videos

Currículo resumido do docente (com foto):

ERICK MUZART FONSECA DOS SANTOS

Graduado em computação, especializado em análise de dados e Deep Learning (DL). Egresso do Deep Learning Summer School da Université de Montreal, Canadá, em 2017, onde teve contato com o estado da arte de DL e com alguns dos melhores pesquisadores mundiais da área. Auditor e cientista de dados no Tribunal de Contas da União (TCU), lotado no Centro de Pesquisa e Inovação onde atua no programa de capacitação dos auditores em análise de dados. Um dos fundadores do grupo de estudo em DL de Brasília, tendo sido instrutor em mais de uma dezena de cursos presenciais e à distância, no TCU, Enap e ESMPU.



FERNANDO LUIZ BRITO DE MELO

Cientista de dados do Senado Federal, é bacharel em Administração e possui especialização em Inteligência Artificial pela Johns Hopkins University.

Com experiência de mais de 20 anos em projetos de análise de dados, é co-organizador do grupo meetup Machine Learning Brasília com mais de 1.900 participantes.

Co-organizador do Grupo de Estudos Deep Learning Brasília, onde atua como professor voluntário com o objetivo de popularizar o uso da Inteligência Artificial e organizar cursos abertos à toda a comunidade de Brasília. Foi instrutor em pelo menos 8 cursos de machine learning e deep learning, ministrados presencialmente no Centro de Treinamento ISC-TCU e remotamente pela ENAP, MPU e TCU.