PROGRAMA DO CURSO
CODING BOOTCAMP CIÊNCIA DE DADOS

Entidade Promotora: Escola Nacional de Administração Pública - Enap
Endereço: Asa Sul SPO área especial 2-A CEP: 70.610-900 Brasília - DF
CNPJ: nº 00.627.612/0001-09

Contato: bootcamp@enap.gov.br
site: https://www.enap.gov.br/pt/cursos/coding-bootcamp

Ínicio das aulas: 26 setembro de 2022
Término das aulas: 24 novembro de 2022

Horário: das 09h às 17h30 - aulas diárias (com intervalo entre 12h às 14h)
Carga horária: 40h prepwork (preparatório para o bootcamp, aulas assíncronas)
360 horas (aulas síncronas, exercícios e projetos)

Nível de formação: capacitação formato bootcamp de média duração

Objetivo / Competências a serem desenvolvidas:

Ao final do curso, o participante será capaz de:

- Realizar programação aplicada à Data Science em Python;
- Fazer desenho de bases de dados relacionais e construção de consultas avançadas com SQL;
- Compreender os conceitos matemáticos classificação e regressão por trás da ciência dos dados: estatística, probabilidade e álgebra linear;
- Conduzir análises avançadas com Jupyter notebook, Pandas e Statsmodels;
- Implementar modelos supervisionados e não supervisionados de Machine Learning com o scikit-learn;
- Aprender as melhores práticas de Machine Learning (pré-processamento, treinamento e testes, métricas de desempenho, etc.);
- Implementar modelos de Machine Learning na produção com a plataforma Google Cloud;
- Construir e treinar redes neurais profundas totalmente conectadas para resolver problemas de classificação e regressão;
- Usar redes neurais para detecção de objetos e processamento de linguagem natural;
- Usar as melhores práticas ao trabalhar em um projeto de ciência de dados dentro de uma equipe técnica.

Ementa do Curso:

Prepwork ~40 horas de trabalho online

1. Kit de ferramentas de Data Science: 80h

2. Decision Science: 40h

3. Machine Learning: 80h

4. Deep Learning: 40h

5. Data Engineering: 40h

6. Projetos finais: 80h

Conteúdo detalhado

PREPWORK - 40 HORAS

Antes de iniciar o bootcamp você terá que completar um trabalho de preparação online. Este trabalho leva cerca de 40 horas e cobre os conceitos básicos de Python, a linguagem pré-requisito do curso e alguns tópicos matemáticos usados todos os dias pelos cientistas de dados.

Conteúdo:
- Python Programming Basics
- SQL básico
- Matemática (estatística, probabilidade, álgebra linear)

1. Kit de ferramentas de Data Science - 80 horas de trabalho

Python para Data Science
Aprenda programação em Python, como trabalhar com Jupyter Notebook e como usar as poderosas bibliotecas Python, como Pandas e NumPy, para explorar e analisar grandes conjuntos de dados. Colete dados de várias fontes, incluindo arquivos CSV, consultas SQL em bancos de dados relacionais, Google Big Query, APIs e Web scraping.

Habilidades Aprendidas:

- Usar o Jupyter Notebook
- Carregar e explorar um conjunto de dados
- Extrair dados de diferentes fontes
- Pandas and NumPy
- Google Big Query
- Web scraping

Base de dados relacionais & SQL

Aprenda como formular uma boa pergunta e como respondê-la, construindo a consulta SQL correta. Este módulo cobrirá a arquitetura do esquema e depois mergulhará profundamente na manipulação avançada do SELECT para extrair informações úteis de um banco de dados independente ou usando um software cliente SQL como o DBeaver.

Habilidades Aprendidas:

- Arquitetura do esquema do banco de dados
- Traduzir uma pergunta comercial em uma consulta SQL
- Manipulações avançadas do SELECT
- Software cliente SQL como DBeaver ou Metabase

Visualização de dados

Torne sua análise de dados mais visual e compreensível, incluindo visualizações de dados em seu Notebook. Saiba como plotar seus quadros de dados usando bibliotecas Python como matplotlib e seaborn e transforme seus dados em insights acionáveis.

Habilidades Aprendidas

- Transforme seus dados em insights com visualizações de dados
- Diferentes categorias de gráficos
- Matplotlib e seaborn

Estatísticas, Probabilidade, Álgebra Linear

Entenda a matemática subjacente a todas as bibliotecas e modelos utilizados no bootcamp. Fique confortável com os conceitos básicos de estatística e probabilidades (média, variância, variável aleatória, Teorema de Bayes, etc.) e com o cálculo matricial, no centro das operações numéricas em bibliotecas como Pandas e Numpy.

Habilidades Aprendidas

- Estatísticas (variável aleatória, distribuição, variância, etc.)
- Probabilidade (teorema do limite central, teorema de Bayes)

● Álgebra Linear (cálculo matricial, derivados)

2. Decision Science - 40 horas de trabalho

Inferências estatísticas

Em seu primeiro mini-projeto de uma semana, você aprenderá como usar ferramentas estatísticas e análise de regressão multivariada para responder a uma questão real de negócios como um verdadeiro analista de dados.
Você aprenderá como sobreviver à fase de preparação de dados de um vasto conjunto de dados, como estruturar um repositório Python com programação orientada a objetos, a fim de limpar seu código e torná-lo reutilizável, e como encontrar, interpretar e apresentar insights significativos que dificilmente poderiam ser feitos com um software de planilha eletrônica convencional.

Habilidades Aprendidas

● Estruturar uma pasta do projeto Python
● Análise de dados
● Teste de Hipótese (A/B)
● Ferramentas estatísticas (statsmodels)
● Análise de regressão multivariada

Comunicação

Os analistas de dados têm o objetivo de comunicar suas descobertas a todos aqueles que não possuem conhecimentos técnicos! Você aprenderá a criar impacto explicando seus conhecimentos técnicos e transformando-os em decisões de negócio, usando análise de custo/benefício. Você será capaz de compartilhar seu progresso, apresentar e comparar seus resultados com seus colegas de equipe.

3. Machine Learning - 80 horas de trabalho

Neste módulo você entenderá as diferentes classes de modelos de machine learning e suas aplicações. Você mergulhará profundamente na biblioteca mais usada em Machine Learning: scikit-learn. Você começará com a **aprendizagem supervisionada** e métodos clássicos como regressões lineares e logísticas para resolver tarefas de previsão. Você então passará para a **aprendizagem não supervisionada** e implementará métodos como o PCA para redução da dimensionalidade ou agrupamento para descobrir grupos em um conjunto de dados. Além disso, iremos ensinar-lhe como **identificar o overfitting** e as diferentes

técnicas disponíveis para evitá-lo. Finalmente, você aprenderá como afinar e avaliar diferentes modelos para alcançar o melhor desempenho usando métodos como **validação cruzada e ajuste de hiperparâmetro**s. Ao longo do caminho, você irá implementar todos os algoritmos essenciais de aprendizagem, tais como **KNN, Máquinas Vetoriais de Suporte** e Métodos de Ensemble como **Random Forests ou Gradient Boosting**.

Fundamentos da aprendizagem supervisionada

Aprenda a limpar e prepare seu conjunto de dados através de técnicas de pré-processamento como vetorização e escalonamento. Familiarize-se com a incrível biblioteca scikit-learn e aprenda como treinar e avaliar o desempenho dos seus primeiros modelos de aprendizagem supervisionada (regressão linear, regressão logística e KNN), tanto para tarefas de regressão como de classificação. Implemente fases de treinamento e testes para garantir que seu modelo possa ser generalizado para dados não vistos e implantado na produção com precisão previsível.

Habilidades Aprendidas

- Biblioteca de scikit-learn para aprendizagem supervisionada
- Técnicas de pré-processamento (vetorização, seleção de características)
- Regressões lineares e logísticas, K-nearest neighbors
- Técnicas de avaliação de desempenho (validação cruzada, holdout)
- Métricas de desempenho (MSE, precisão, matriz de confusão...)

Generalização e sobreajustamento

Dê um mergulho profundo no treinamento de modelos para entender como os algoritmos de Machine Learning são implementados em seu núcleo. Aprenda a afinar os seus modelos para que eles generalizem melhor os dados invisíveis. E descubra os poderosos algoritmos SVM.

Habilidades Aprendidas

- Como são treinados os modelos? (minimização da função de perda por descida de gradiente, etc.)
- Generalização de modelos (evitar sobreajustamento através da regularização de perda-função)
- Afinação de hiperparâmetros (grid-search, etc.)
- Support Vector Machines (SVMs)

Aprendizagem sem Supervisão

Passar à aprendizagem não supervisionada e implementar métodos como o PCA para redução da dimensionalidade ou agrupamento para descoberta de grupos num conjunto de dados.

Habilidades Aprendidas

- scikit-learn para aprendizagem não supervisionada
- PCA (Análise de Componentes Principais)
- Modelos de clustering (K-means, etc.)

Métodos de montagem e Tópicos Especiais

Depois de aprender como canalizar todos os seus passos de modelagem juntos e combinar vários modelos em poderosos "Ensemble Models" como o Random Forests, você aplicará suas habilidades em aplicações de casos reais e participará de competições de Kaggle. Você também aprenderá modelos específicos para análise de Séries Temporais, bem como Processamento de Linguagem Natural.

Habilidades Aprendidas

- Pré-processamento e modelação de condutas em conjunto
- Participar de Concurso de Kaggle
- Ensemble Methods (Random Forest, Gradient Boosting...)
- Série cronológica (modelos SARIMA)
- Processamento de Linguagem Natural (Naive Bayes Classifiers, NLTK...)

4. Deep learning - 40 horas de trabalho

Revelar a magia por detrás do Deep Learning através da compreensão da arquitetura das redes neurais (neurônios, camadas, pilhas) e seus parâmetros (ativações, perdas, otimizadores). Adquira autonomia para construir suas próprias redes, especialmente para trabalhar com imagens, tempos e textos, enquanto aprende as técnicas e truques que fazem o Deep Learning funcionar. Este módulo é baseado em problemas da vida real que irão desafiá-lo a otimizar as suas funcionalidades e arquitetura de forma a obter o melhor desempenho.

Deep Learning facilitado

Descubra a biblioteca Keras Deep Learning que permite fazer protótipos facilmente enquanto tem a flexibilidade de afinar precisamente a sua rede neural. Além disso, o Google Colab irá acelerar muito o tempo computacional graças às GPUs dedicadas.

Habilidades Aprendidas

- Arquitetura de Rede Neural Densa
- Avaliação de desempenho e sobreajuste
- Biblioteca Tensorflow Keras
- Google Colab

Visão Computacional

Vá mais longe na visão computacional com Convolutional Neural Networks, arquiteturas projetadas para tirar o máximo proveito das imagens. Melhore a generalização do seu modelo graças a técnicas de aumento de dados e implemente métodos avançados para se beneficiar de arquiteturas de última geração, graças aos métodos de aprendizagem Transfer.

Habilidades Aprendidas

- Convolutional Neural Networks (CNN)
- Pré-processamento de imagens e carregamento de dados em lote
- Aumento de dados
- Transfer Learning (VGG16, etc.)
- Auto-encoders

Times-Series & Text data

Fique à vontade para gerenciar dados seqüenciais e textos (sequência de palavras) transformando-os em entradas apropriadas. Aproveite o poder das Redes Neurais Recorrentes para prever valores futuros e realizar processamento de linguagem natural valioso.

Habilidades Aprendidas

- Recurrent Neural Networks (RNN, LSTM, etc.)
- Previsões de séries temporais de múltiplas saídas
- Incorporação de palavras
- Análise dos sentimentos

5. Engenharia de Dados - 40 horas de trabalho

Em uma semana, aprenda todas as melhores práticas para resolver grandes problemas de ML apenas no seu computador e disponibilize a sua previsão para o mundo através de uma API! Primeiro, vamos ensinar-lhe a ser mais produtivo na construção de um modelo de machine-learning, usando o fluxo de trabalho mais adequado. Em seguida, vamos aproveitar uma biblioteca chamada MLflow para registrar suas múltiplas experimentações, iterações e ajustes. Terceiro, mostraremos a você como treinar em escala usando o poder da computação em nuvem com a Plataforma AI do Google Cloud. Finalmente, você

aprenderá Docker e, com isso, poderá implementar seu código e modelo para produção e torná-lo disponível para todo o mundo, usando o Cloud Run ou o Kubernetes Engine.

Machine Learning como Produto

Saiba como montar um projeto de machine learning da forma correta: Você vai passar do Jupyter Notebook para um editor de código da maneira certa para iterar rápida e confiantemente com um pipeline robusto e escalável. Você aprenderá python packaging, versionamento com Git e integração contínua com GitHub Actions. Você também vai aproveitar uma biblioteca chamada MLflow para registrar suas múltiplas experimentações, iteração e ajuste.

Habilidades Aprendidas

- Do Jupyter Notebook ao código
- Configurando um bom projeto ML (pastas, arquivos, etc.)
- Integração Contínua (Ações GitHub)
- MLflow para o desempenho do modelo MLflow
- Usando Sklearn-Pipeline (Encoders, Transformers)

Trem em escala com a plataforma Google Cloud

Mostraremos a você como aproveitar a capacidade de armazenamento na nuvem, bem como o poder computacional para treinar modelos em conjuntos de dados pesados usando a plataforma Google Cloud AI Platform.

Habilidades Aprendidas

- Plataforma Google Cloud (CGP)
- Usar armazenamento baseado em nuvem (Cloud Storage)
- Use máquinas virtuais para treinar modelos com GPUs (Plataforma de IA do Google, Cadernos de IA, Colab)

Implementar modelo na produção

Implante o seu modelo treinado para a produção com Docker ou mesmo Kubernetes para torná-lo disponível para o mundo. Sirva os seus resultados para uma interface baseada na web (Streamlit) e consulte o seu modelo através de chamadas API (FastAPI).

Habilidades Aprendidas

- API (Fast API)
- Virtualização (Compute Engine)
- Docker (Cloud Run)
- Kubernetes (Motor Kubernetes)

● Aplicação front-end baseada na Web (Streamlit)

6. Projetos Finais - 80 horas de trabalho

O objetivo deste módulo é reunir todos os componentes que você aprendeu até agora e trabalhar com uma equipe em problemas reais.

Projetos Estudantis

Você vai passar as últimas duas semanas do bootcamp em algum projeto de grupo trabalhando em um super problema de data science que você e seu grupo queiram resolver! Você usará uma mistura das suas próprias combinações de dados (se você tiver alguma da sua empresa) e repositórios de dados abertos (iniciativas governamentais, Kaggle, etc.). Será uma ótima maneira de praticar todas as ferramentas, técnicas e metodologias abordadas no Curso de Ciência de Dados e fará você perceber o quão autônomo você se tornou.

Habilidades Aprendidas

● Formule um problema em qualquer combinação de dados
● Faça um projeto de Ciência de Dados Ponta a Ponta
● Colabore como uma equipe de dados
● Comunique os seus resultados
● Apresente resultados com análise de custo/benefício
● Confira as demonstrações em https://lew.ag/data-demodays