PLANO DE ENSINO

IDENTIFICAÇÃO

HABILITAÇÃO:	Profissionalizante em Ciência de Dados	**C. H. TOTAL:** 400 horas
PROFESSOR (A):	Gerson do Nascimento Silva (PhD Student, UnB)	

PLANEJAMENTO

BASES TECNOLÓGICAS

Ementa: Desenvolvimento de código-fonte (*script)* utilizando linguagem *python* para implementação, mostrando como:

1. **Uso da biblioteca "scikit-learn" (sklearn) e "pandas" :**
   - *a)* Carregar dados de habitação, via sklearn (Boston);
   - *b)* Criar dados simulados para regressão;
   - *c)* Criar dados simulados para classificação;
   - *d)* Criar dados simulados para armazenamento em *cluster;*
   - *e)* Preparar um fluxo de trabalho de aprendizado de máquina;
   - *f)* Converter recursos (características) categóricos em recursos numéricos;
   - *g)* Imputar rótulos de classes ausentes;
   - *h)* Imputar rótulos de classes ausentes usando método "vizinho próximo";
   - *i)* Excluir instâncias com valores ausentes;
   - *j)* Como fazer operações numéricas;
   - *k)* Como encontrar *outliers*;
   - *l)* Codificar recursos categóricos ordinais;
   - *m)* Lidar com classes de desequilíbrio com redução da resolução;
   - *n)* Como lidar com classes de desbalanceadas;
   - *o)* Como lidar com *outliers*;
   - *p)* Imputar valores ausentes com médias;
   - *q)* Codificação com vários rótulos;
   - *r)* Codificação com recursos nominais categóricos;
   - *s)* Processar recursos categóricos;
   - *t)* Redimensionar recursos;
   - *u)* Padronizar recursos;
   - *v)* Padronizar dados "IRIS";
   - *w)* Dividir dados *DateTime* ("features") para criar vários recursos;
   - *x)* Calcular a diferença entre datas;
   - *y)* Codificar os dias da semana;
   - *z)* Tratar valores ausentes em uma série temporal;
   - *aa)* Como introduzir o tempo "LAG" (lagged time-series), tempo de latência são muito usadas em analises econômicas;
   - *bb)* Como lidar com "Janelas de Tempo";
   - *cc)* Selecionar *DateTime* dentro de um intervalo;
   - *dd)* Selecionar *DateTime [formato (PM) ou (AM)]* dentro de um intervalo
   - *ee)* Como trabalhar itens em uma lista;

2. **Uso da biblioteca "Seaborn", PCA, ANOVA, Curva ROC, Hiper-parâmetros e Modelos de Regressão**
   - a) Reindexar *dataframes*;
   - b) Substituir múltiplos valores usando biblioteca computacional *Pandas*;
   - c) Plotar *dataframes* com biblioteca computacional *Seaborn*;
   - d) Limpeza, estruturação e enriquecimento de dados brutos – "*Data Wrangling*";
   - e) Plotar gráfico de barras usando *dataframes Pandas*;
   - f) Utilizar séries temporais em *Pandas*;
   - g) Gerar séries temporais com *Pandas* e *Seaborn*;
   - h) Gerar "*grouped bar*";
   - i) Determinar e plotar coeficientes de correlação;
   - j) Reduzir dimensionalidade de matriz esparsa;
   - k) Usar PCA (*Principal Component Analysis*) para reduzir dimensionalidade;
   - l) Extrair características usando PCA;
   - m) Usar "*ANOVA F-values*" para selecionar características;
   - n) Usar "*Chi squared*" para selecionar características (*features*);
   - o) Remover características altamente correlacionadas;
   - p) Verificar acurácia de um modelo com validação cruzada;
   - q) Verificar "pontuação AUC" de um modelo;
   - r) Plotar curva de aprendizado de máquina;
   - s) Plotar "curva ROC";

t) Usar o algoritmo "*Random Forest*";
u) Sintonizar hiper-parâmetros com algoritmo "*GridSearchCV*" e "*Random Search*";
v) Otimizar hiper-parâmetros de um modelo de regressão logística;
w) Otimizar hiper-parâmetros de um modelo de árvore de decisão;
x) Criar e otimizar "baseline" de um modelo de regressão linear;
y) Criar e otimizar "baseline" de um modelo de regressão ridge;
z) Criar e otimizar "baseline" de um modelo de regressão lasso e "ElasticNet".

3. Classificação, Clusterização e Regressão

a) Criar e otimizar modelo para regressão e classificação;
b) Utilizar "*nearest neighbours*" para regressão e classificação;
c) Fazer agrupamento aglomerativo (*Agglomerative Clustering*);
d) Fazer clusterização com o "*Kmeans*";
e) Fazer clusterização baseado em afinidade;
f) Utilizar "*DBSCAN Clustering*";
g) Utilizar a abordagem do deslocamento médio (*MinShift*);
h) Utilizar a arvore de classificação e regressão;
i) Utilizar "*AdaBoost*";
j) Utilizar "*RandomForest*";
k) Utilizar "*GradientBoosting*";
l) Utilizar classificador e regressor multicamadas – "*MLP - Multi Layer Perceptron*";
m) Utilizar classificador e regressor de reforço gradual "*XgBoost*";
n) Utilizar classificador e regressor "*CatBoost*";
o) Utilizar classificador e regressor "*LightGBM*";
p) Utilizar classificador e regressor "*SVM*";
q) Classificar com modelos lineares – "*Multiclass Classification*";
r) Classificar com modelos lineares – "*Naive Bayes*";
s) Classificar com modelos lineares – "*Nearest Neighbors"*;
t) Classificar com modelos lineares – "*LDA* e *QDA*";
u) Classificar com modelos lineares – "*Tree Model*";
v) Classificar com modelos lineares – "*Ensemble Bagging Model*";
w) Classificar com modelos lineares – "*Ensemble Boosting Model*";
x) Utilizar métrica de classificação e regressão;
y) Comparar algoritmos de classificação;
z) Implementar "*Ensemble Model*";
aa) Salvar modelos treinados;
bb) Avaliar modelos com curvas de aprendizagem;
cc) Paralelizar execução e validação cruzada no "*XGBoost*";
dd) Otimizar número de árvores no "*XGBoost*".

4. Visualização de dados: Plotagem de gráficos

a) Autocorrelação (ACF) e Autocorrelação Parcial (PACF);
b) Pizza com destacamento;
c) Plotagem de textos;
d) Divergências de escala;
e) Densidade;
f) Series temporais múltiplas com escalas;
g) Boxplot;
h) Correlograma – "Correllogram";
i) Curvas de densidade – "Cross Correlation";
j) Decomposição de serie temporal;
k) Dispersão com linha de regressão linear de melhor ajuste;
l) Área;
m) Barras;
n) Bolha – "Bubble";
o) Cascata- "Waterfall";
p) Lotes em par – "Pairwise";
q) Histograma empilhado;
r) Variáveis continuas (Histograma);
s) Histogramas marginais – "Marginal Boxplot";
t) Quadro de marcadores – "Lollipop";
u) Piramide populacional;
v) Diagramas de dispersão – "Scatter";
w) Anotações de picos e vales em series temporais;
x) Cilindros – "Cylinder";
y) Feixes sólidos – "Joy";
z) Dado sazonal.

TEORIA	PRÁTICA
	- Pré-processamento: (representação e a qualidade dos dados**)**

-video aula: (*playlist* do PhD Francisco A. Rodrigues – USP) • **"**Estatísticas**"**; • **"**Probabilidades**"**; • **"**Redes Complexas**"**; • **"**Processos Estocásticos**"**; • **"**Ciência de Dados**"**; **-video aula: (*playlist* do Dr. Alexandre L. M. Levada – USP)** • **"**Introdução a Teoria dos Grafos**";** • **"**Reconhecimento de Padrões**"**; **-video aula: (*playlist* da Dra. Cibele Russo – USP)** • **"**Visualização e Exploração de Dados**";** • **"**Análise Multivariada e Aprendizado Não-Supervisionado**"**.	passo importante no processo de tratar dados. A frase "*garbage in, garbage out*" é particularmente aplicável aos projetos da trilha de IA (data mining, data science e machine learning). Os métodos de coleta de dados geralmente são poucamente controlados, resultando em valores de intervalo *out-of-range* (por exemplo, renda: −100), combinações de dados impossíveis (por exemplo, sexo: masculino, grávidas: sim), *missing values*, etc. As análise de dados que não foram cuidadosamente selecionados para tais problemas pode produzir resultados enganosos. Assim, a representação e a qualidade dos dados são antes de tudo uma análise. **- Análise de dados:** é um processo de inspeção, limpeza, transformação e modelagem de dados com o objetivo de descobrir informações úteis, informar conclusões e apoiar a tomada de decisões. A análise de dados tem múltiplas facetas e abordagens, abrangendo diversas técnicas sob uma variedade de nomes, e é usada em diferentes domínios dos negócios, ciências e ciências sociais. Desempenha um papel tornando a tomada de decisões mais científicas e ajudando no processo de operar com mais eficácia. **- Classificação:** baseia-se em prever a categoria de uma observação dada. Procura-se estimar um "classificador" que gere como saída a classificação qualitativa de um dado não observado com base em dados de entrada (que abrangem observações com classificações já definidas). Exemplo: um classificador que utilize dados não observados de um paciente e classifique-o como doente ou não-doente. **- Regressão:** de forma similar a classificação, utiliza dados de entrada (preditores) já observados para prever uma resposta. A grande diferença é que, neste caso, procura-se estimar um valor numérico e não uma classificação de uma observação. Exemplo: estimar um modelo que utilize a idade e os anos de escolaridade de um indivíduo não-observado anteriormente para tentar prever seu salário. Utiliza-se como base desse modelo: idades, anos de escolaridades e salários de diversos indivíduos já observados anteriormente. **- Agrupamento:** também conhecido como "*Clustering*", tem como objetivo agrupar observações em grupos conhecidos como "*clusters*". Essas observações apresentam similaridades dentro de seu *cluster* e diferenças em relação aos demais *clusters* formados. Diferente da Classificação, não é realizada a rotulação dos *clusters*, fazendo com que não exista uma clusterização errada ou certa. A clusterização utilizada resulta em diferentes tipos de *clusters*, e a escolha dessas técnicas deve ser previamente analisada pelo pesquisador. Exemplo: agrupar fotos de animais similares em *clusters*, sem ter o conhecimento prévio de qual animal está sendo apresentado. **- Visualização:** Tornar a apresentação dos dados atraente e de fácil entendimento; Identificar tendências; Perceber situações atípicas em um conjunto de dados; Contar uma história encontrada nos dados; Reforçar um argumento ou opinião; Destacar um ponto importante em um conjunto de dados.

ANÁLISE DA REALIDADE
Pré-Requisito
1. Lógica de Programação;

2. Programação Orientada a Objetos;
3. Análise de Sistemas Orientada a Objeto;
4. Fundamentos e Modelagem de banco de Dados.

Necessidade da turma:

- Programação intermediário em *python* ou outra linguagem de programação que traga subsídios para entendimento dos conceitos computacionais.

PROJEÇÃO DE FINALIDADES

Objetivos Gerais (Competências)

Compreender os conceitos de IA e seus recursos e capacidade para implementar código-fonte reproduzível.

Objetivos Específicos (Habilidades)

-Capacidade para implementar código-fonte (uso de *python* ou outras tecnologias com mesma finalidade);
-Capacidade para implementar API'S para o pré-processamento de dados;
-Capacidade para implementar API'S para análise de dados;
-Integração do uso de API`S em manipulação de banco de dados;
-Entender como aplicar frameworsks python focados em – Classificação;
-Entender como aplicar frameworsks python focados em – Clusterização;
-Entender como aplicar frameworsks python focados em – Regressão;
-Ser capaz de implementar código-fonte reproduzível utilizando conceitos anteriores.

FORMAS DE MEDIAÇÃO

Procedimentos Metodológicos

-Elaborar modelos de implementação de código-fonte que valide as teorias, com uso de exemplos;
-Propor trabalhos práticos sobre o assunto;
-Corrigir os trabalhos práticos;
-Elaborar avaliação do conhecimento sobre o assunto ministrado;
-Uso de computadores para implementar exemplos;

AVALIAÇÃO*

Instrumento(s)	Data
• Avaliação única por meio do instrumento *Quiz;* • Criação de um projeto final **"***End to End***"** implementado em código-fonte reproduzível python.	Ao fim do módulo estudado.

***** No Exame Quiz não é permitido qualquer espécie de acréscimo de nota (trabalhos, participações e outros) estranhos à avaliação.

***** Quiz é um jogo mental no qual os jogadores tentam responder corretamente a questões que lhes são colocadas. A palavra também é utilizada como sinônimo de avaliação de aquisição de conhecimentos ou capacidades em ambientes de aprendizagem.

BIBLIOGRAFIA

BIBLIOGRAFIA BÁSICA (Títulos, periódicos, etc)

Título/Periódico	Autor	Edição	Editora	Ano	Livro Texto
Álgebra Linear	José luiz Boldrini [et al.]	3ª edição	Harbra Ltda	1980	-
Trilhas Python	Eduardo Pereira	1ª edição	Casa do Código	2018	-
Python para Desenvolvedores	Luiz Eduardo Borges	3ª edição	Novatec	2014	Sim

BIBLIOGRAFIA COMPLEMENTAR (Títulos, periódicos, etc)

Título/Periódico	Autor	Edição	Editora	Ano
Curso Intensivo de Python	Eric Matthes	1ª edição	Novatec	2017
Data Science from Scratch	Joel Grus	1ª edição	O'Reilly	2015
Introduction to Applied Linear Algebra	Stephen Boyd	1ª edição	Cambridge University Press	2018

Outros:	**Essa é uma coleção de livros da área de IA (sugestão p/ aquisição):**

	• 1. Beginning Python Visualization: Crafting Visual Transformation Scripts (Books for Professionals by Professionals) : https://amzn.to/2BwQqNM • 2. Data Analysis with Open Source Tools: A Hands-On Guide for Programmers and Data Scientists : https://amzn.to/2S6p34d • 3. Mining The Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub and More : https://amzn.to/2S3armh • 4. Python: Advanced Predictive Analytics : https://amzn.to/2SapMl2