

<b>UNIVERSIDADE DE PERNAMBUCO</b>		
<b>UNIDADE – ESCOLA POLITÉCNICA DA UNIVERSIDADE DE PERNAMBUCO</b>		
<b>DISCIPLINA – APRENDIZADO DE MÁQUINA PARA CIENTISTAS</b>		
<b>CÓDIGO DA DISCIPLINA – FIS51</b>		
<b>CARGA HORÁRIA TOTAL – 60 HORAS TEÓRICAS</b>		
<b>EMENTA</b> <i>Introdução ao aprendizado de máquina, problemas que o aprendizado de máquina pode resolver, tarefas e dados, Python, Scikit-learn, bibliotecas e ferramentas essenciais, Jupyter Notebook, NumPy, SciPy, matplotlib, pandas, mglearn, aprendizado supervisionado, classificação e regressão, generalização, ajustes, modelos lineares, classificadores, aprendizado e pré-processamento não-supervisionados, dimensionamento, representando dados e recursos de engenharia, modelos lineares e árvores, avaliação e melhoria do modelo, validação cruzada, métricas, cadeias de algoritmos, trabalhando com dados de texto, modelagem de tópicos e aglomerados de documentos, abordando um problema de aprendizado de máquina.</i>		
<b>ÁREA/EIXO/NÚCLEO</b>  <i>CIÊNCIAS EXATAS FÍSICA DE MATERIAIS NÚCLEO PROFISSIONALIZANTE</i>	<b>COMPETÊNCIA(S)</b>  <i>1. Compreender os fundamentos da aprendizagem de máquina. 2. Compreender o conceito de aprendizado supervisionado. 3. Compreender os fundamentos do aprendizado e do pré-processamento não supervisionados. 4. Compreender processos de melhoria de modelos e suas métricas.</i>	<b>HABILIDADES</b>  <b>COMPETÊNCIA 1</b> <ul style="list-style-type: none"> <li>• Compreender os fundamentos do aprendizado de máquina.</li> <li>• Conhecer os problemas que o aprendizado de máquina pode resolver.</li> <li>• Compreender o conjunto de bibliotecas e ferramentas essenciais: Python, Scikit-learn, Jupyter Notebook, NumPy, SciPy, matplotlib, pandas, mglearn.</li> <li>• Realizar uma primeira aplicação de aprendizado de máquina na classificação de espécies de íris.</li> <li>• Compreender o conceito de medição de sucesso, treinamento e teste de dados.</li> <li>• Construir um primeiro modelo de k-vizinhos mais próximos e realizar sua avaliação.</li> </ul> <b>COMPETÊNCIA 2</b> <ul style="list-style-type: none"> <li>• Compreender o conceito de aprendizado supervisionado.</li> <li>• Entender os mecanismos de classificação, regressão, ajuste excessivo e ajuste insuficiente.</li> <li>• Compreender o funcionamento de algoritmos de aprendizado de máquina supervisionados e métodos de classificação e decisão.</li> </ul>



5. *Compreender os fundamentos da análise de dados em texto e abordar problemas em aprendizado de máquina.*

- Compreender o funcionamento de redes neurais e aprendizagem profunda e suas relações nas estimativas de incerteza dos classificadores.

### **COMPETÊNCIA 3**

- Compreender o aprendizado e o pré-processamento não supervisionados: tipos e desafios no aprendizado e dimensionamento, diferentes tipos de pré-processamento,
- Compreender os fundamentos do tratamento e das transformações de dados.
- Entender os conceitos de dimensionamento de dados de treinamento e teste e avaliação de algoritmos de aglomerados.

### **COMPETÊNCIA 4**

- Compreender avaliação e melhoria de modelos: validação cruzada, validação cruzada no Scikit-learn, benefícios da validação cruzada, validação cruzada estratificada de k-fold e outras estratégias,
- Compreender o funcionamento da pesquisa em grade, pesquisa em grade com validação cruzada.
- Entender os conceitos de métricas e pontuação de avaliação, métricas para classificação binária, métricas para classificação multiclasse, métricas de regressão, usando métricas de avaliação na seleção de modelos.
- Compreender a estrutura de cadeias de algoritmos e pipelines, seleção de parâmetros e pré-processamento.

### **COMPETÊNCIA 5**

- Compreender os fundamentos da análise de dados de texto.
- Aplicar o aprendizado de máquina para problemas em dados de texto: tokenização, stemização e lematização avançados, modelagem de tópicos e aglomerados de documentos, alocação de Dirichlet latente.
- Abordar um problema de aprendizado de máquina: do protótipo à produção, testando sistemas de produção, construindo seu próprio estimador, estruturas e pacotes, classificação, recomendação e outros tipos de aprendizado, modelagem probabilística, inferência e programação probabilística, redes neurais.

## **CONTEÚDO PROGRAMÁTICO**



1. *Introdução ao aprendizado de máquina: problemas que o aprendizado de máquina pode resolver, conhecendo sua tarefa e conhecendo seus dados, Phyton, instalando Scikit-learn, bibliotecas e ferramentas essenciais, Jupyter Notebook, NumPy, SciPy, matplotlib, pandas, mglearn, Python 2 versus Python 3, a primeira aplicação e a classificação de espécies de íris, medindo o sucesso, treinando e testando dados, construindo seu primeiro modelo: k-vizinhos mais próximos, fazendo previsões, avaliando o modelo.*
2. *Aprendizado supervisionado: classificação e regressão, generalização, ajuste excessivo e ajuste insuficiente, relação da complexidade do modelo para o tamanho de conjunto de dados, algoritmos de aprendizado de máquina supervisionados, alguns conjuntos de dados de amostra, k-vizinhos mais próximos, modelos lineares, classificadores de Bayes, árvores de decisão, conjuntos de árvores de decisão, máquinas de vetores de suporte kernelizado, redes neurais e aprendizagem profunda, estimativas de incerteza dos classificadores, função de decisão, prevendo probabilidades, incerteza na classificação multiclases.*
3. *Aprendizado e Pré-processamento não supervisionados: tipos de aprendizado não supervisionado, desafios no aprendizado não-supervisionado, pré-processamento e dimensionamento, diferentes tipos de pré-processamento, aplicação de transformações de dados, dimensionamento de dados de treinamento e teste, efeito do pré-processamento no aprendizado supervisionado, redução de dimensionalidade, extração de recursos e aprendizado múltiplo, análise de componentes principais (PCA), fatoração de matriz não-negativa (FMN), aprendendo com o t-SNE, aglomeração, k-means, agrupamento espacial de aplicativos com ruído baseado em densidade, comparando e avaliando algoritmos de aglomerados.*
4. *Representando dados e recursos de engenharia: variáveis categóricas, One-Hot-Encoding (variáveis dummy), números e codificação, armazenamento, discretização, modelos lineares e árvores, interações e polinômios, transformações não-lineares univariadas, seleção automática de recursos, estatísticas univariadas, seleção de recursos baseados em modelo, seleção de recursos iterativos, utilizando o conhecimento especializado.*
5. *Avaliação e Melhoria do Modelo: validação cruzada, validação cruzada no Scikit-learn, benefícios da validação cruzada, validação cruzada estratificada de k-fold e outras estratégias, pesquisa em grade, pesquisa simples em grade, o perigo de sobrescrever os parâmetros e o conjunto de validação, pesquisa em grade com validação cruzada, métricas e pontuação de avaliação, métricas para classificação binária, métricas para classificação multiclasse, métricas de regressão, usando métricas de avaliação na seleção de modelos.*
6. *Cadeias de algoritmos e pipelines: Seleção de parâmetros com pré-processamento, criando pipelines, usando pipelines em pesquisas em grade, a interface geral das pipelines, criação conveniente de pipeline com make\_pipeline, pesquisando em qual modelo usar.*
7. *Trabalhando com dados de texto: tipos de dados representados como sequências de caracteres, exemplo de aplicação na análise de sentimentos de críticas de filmes, representando dados de texto como um pacote de palavras, aplicando itens de palavras em um conjunto de dados de brinquedos, palavras-chave para revisões de filmes, palavras-chave, intervalando os dados com tf – idf, investigando coeficientes de modelo, palavras-chave com mais de uma palavra (n-grama), Tokenização, Stemming e Lematização avançados, modelagem de tópicos e aglomerados de documentos, alocação de Dirichlet latente.*
8. *Abordando um problema de aprendizado de máquina: humanos em loop, do protótipo à produção, testando sistemas de produção, construindo seu próprio estimador, outras estruturas e pacotes de aprendizado de máquina, classificação, sistemas de recomendação e outros tipos de aprendizado, modelagem probabilística, inferência e programação probabilística, redes neurais, dimensionamento para conjuntos de dados maiores.*

## BIBLIOGRAFIA



1. MÜLLER, A. C., AND GUIDO, S. **Introduction to Machine Learning with Python: A Guide for Data Scientists**. 1st Ed., O'Reilly Media, 2016.
2. WATT, J., and BORHANI, R., and KATSAGGELOS, A. K. **Machine Learning Refined: Foundations, Algorithms, and Applications**. 2<sup>nd</sup> Ed., Cambridge University Press, 2020.
3. JAMES, G., and WITTEN, D., and HASTIE, T., and TIBSHIRANI, R. **An Introduction to Statistical Learning: with Applications in R**. 7<sup>th</sup> Ed., Springer, 2017.
4. JACOBS, P. **Machine Learning with Python: Advanced Guide in Machine Learning with Python**. 1<sup>st</sup> Ed., Independently published, 2019.
5. DEISENROTH, M. P., and FAISAL, A. A., and ONG, C. S. **Mathematics for Machine Learning**. 1<sup>st</sup> Ed., Cambridge University Press, 2020.
6. BURKOV, A. **The Hundred-Page Machine Learning Book**. 1<sup>st</sup> Ed., Andriy Burkov, 2019.

