**Curso de Engenharia de Computação**
**Disciplina: Inteligência Artificial**
**Professora: Alexandra Zimpeck**

# Descrição do Trabalho de Implementação

O aprendizado de máquina é o campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados (Arthur Samuel, 1959). O aprendizado pode ser classificado de três formas: supervisionado, não supervisionado e por reforço. Neste trabalho, os alunos devem escolher um problema de classificação, e criar três modelos aplicando diferentes algoritmos de aprendizado supervisionado. Para concluir com êxito a tarefa, siga os seguintes passos:

**1) Banco de dados**

Acesse os repositórios abaixo e escolha um banco de dados que aborde um tema de sua preferência e que represente um problema de classificação:

a) Kaggle: *Your Machine Learning and Data Science Community*

b) UCI: *Machine Learning Repository*

Considere 70% das amostras do banco de dados para treino e reserve 30% das amostras para a etapa de teste.

**2) Escolha dos algoritmos de aprendizado de máquina**

Para a geração dos modelos, escolha três algoritmos de aprendizagem supervisionada que podem ser aplicados em problemas de classificação:

- Árvores de decisão (*Decision Tree*)
- Naive Bayes (*Naive Bayes Classifier*)
- Suporte de Máquina de Vetores (SVM – *Support Vector Machine*)
- K-ésimo vizinho mais próximo (KNN – *K-Nearst Neighbors*)
- Floresta aleatória (*Random Forest*)
- Regressão logística (*Logistic Regression*)

Entenda o funcionamento e implemente os códigos em Python a fim de realizar o treinamento dos modelos utilizando os três algoritmos escolhidos.

**3) Avaliação dos resultados**

Informe o desempenho de cada um dos três modelos considerando como métrica a acurácia (valor em porcentagem). Após, responda à questão abaixo:

- Qual modelo apresentou maior e menor acurácia? Indique possíveis justificativas para este tipo de comportamento, considerando o princípio de funcionamento dos algoritmos de aprendizagem supervisionada utilizados.

**4) Aperfeiçoamento focado nos dados**

Realize o aperfeiçoamento em cada modelo de forma a adequar os dados do banco escolhido. Você deve escolher duas melhorias a serem feitas:

- Tratamento de dados faltantes.
- Tratamento dos *outliers*.
- Realização do balanceamento dos dados.
- Exclusão de alguns atributos.
- Redução da dimensionalidade.

Após inserir as melhorias no banco de dados, avalie novamente o desempenho de cada um dos modelos, também usando como métrica a acurácia. Após, responda as seguintes questões:

- O pré-processamento dos dados ajudou a melhorar a acurácia dos modelos? Justifique sua resposta.
- Houve alguma diferença na acurácia quando duas melhorias foram adicionadas ao mesmo tempo e quando cada uma delas foi feita separadamente? Justifique sua resposta.

## Relatório e apresentação

Devem conter:

a) Descrição completa do banco de dados escolhido (contextualização do tema, significado de cada um dos atributos e desfecho, fonte).
b) Explicação dos algoritmos de aprendizado de máquina escolhidos.
c) Explicação do conceito e justificativa dos aperfeiçoamentos escolhidos.
d) Explicação dos códigos em Python para criação dos modelos.
e) Discussão sobre o desempenho atingido pelos modelos (acurácia) com e sem o pré-processamento dos dados.
f) Respostas às perguntas feitas nos itens 3 e 4.
g) Conclusões finais.

## Informações gerais

1. O trabalho deve ser feito individualmente.
2. Entrega do relatório: até 30 de junho de 2021 às 23h59 (*Google Classroom*).
3. Data de apresentação: 23 de junho de 2021 às 19h15 (10 minutos por aluno).
4. Valor: 3 pontos na nota do segundo bimestre.