

Trabalho Final – Aprendizagem supervisionada

Componente Curricular: GEX618 – Inteligência Artificial

Créditos: 4

Ano/semestre: 2024/1

Professor: Felipe Grando

1. Descrição

Trabalho individual ou em duplas cujo objetivo é o estudo e a aplicação de métodos de aprendizagem supervisionada por meio de uma aplicação prática com a técnica Naive Bayes, Árvores de Decisão ou Redes Neurais Artificiais com arquitetura multicamada totalmente conectada para frente (MLPs – *Multi-layer Perceptron*).

O trabalho deve seguir as seguintes etapas:

1. (1 ponto) Selecione um conjunto de dados rotulados disponível na plataforma Kaggle (<https://www.kaggle.com/datasets>).
 - a. O conjunto de dados deve possuir categorias como rótulos, caracterizando um problema de classificação binária ou multiclasse.
 - b. Evitar conjuntos de dados que possuem amostras não independentes (textos, áudios, vídeos e séries temporais, por exemplo).
 - c. Evitar conjuntos de dados compostos por imagens.
 - d. O conjunto de dados deve possuir no mínimo 100 amostras e indica-se que não seja escolhido um conjunto de dados maior que 1GB em tamanho.
2. (2 pontos) Trate os dados conforme características apresentadas.
 - a. Limpeza e seleção: retire todos os dados (atributos e amostras) que não são relevantes para o problema.
 - b. Dados ausentes: use métricas de tendência central para completar dados de amostras incompletas (lembre-se de aplicar a métrica adequada para cada tipo de atributo).
 - i. <https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html>
 - c. Quando o conjunto de dados tiver atributos contínuos e também atributos discretos, discretize os dados contínuos ao transformá-los em categorias representadas por intervalos.
 - i. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.KBinsDiscretizer.html>
 - d. Quando o conjunto de dados tiver atributos discretos representados por texto, codifique o texto em valores numéricos.
 - i. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html>
 - e. Quando o conjunto de dados tiver apenas atributos contínuos, padronize os dados contínuos ao modificar a escala dos dados para uma escala com média igual a 0 e desvio padrão igual a 1.
 - i. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>
3. (2 pontos) Selecione e treine um modelo classificador usando uma técnica de aprendizagem supervisionada dentre: Naive Bayes, Árvores de Decisão ou Redes Neurais Artificiais com arquitetura multicamada totalmente conectada para frente (MLPs).

- a. Usar GaussianNB quando os atributos das amostras forem valores contínuos.
 - i. https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html
 - b. Usar MultinomialNB ou DecisionTreeClassifier quando os atributos das amostras forem valores discretos.
 - i. https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html
 - ii. <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>
 - c. A técnica MLPClassifier pode ser utilizada independente do tipo dos atributos das amostras, desde que todos sejam representados por valores numéricos.
 - i. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html
 - d. Use apenas 70% dos dados para o treinamento e reserve 30% dos dados para os testes.
 - i. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
4. (2 pontos) Avalie o modelo classificador com as previsões do modelo sobre o conjunto de testes.
- a. Compute a acurácia.
 - i. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html
 - b. Compute a matriz de confusão.
 - i. <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.ConfusionMatrixDisplay.html>
5. (1 ponto) Organize e documente o código implementado (leitura e tratamento dos dados, treino e teste do modelo).
- a. Entregue, no SIGAA, documento (PDF) com nome dos integrantes do grupo e link para o projeto no GitHub (deixe o projeto público) ou arquivo compactado (RAR ou ZIP) com os arquivos do projeto implementado (identificar os integrantes do grupo em arquivo texto ou comentário dentro do código).
6. (2 pontos) Apresente (slides ou PDF) e discuta o trabalho com a turma.
- a. Explique o conjunto de dados selecionado, descrevendo suas principais características.
 - b. Descreva os tratamentos utilizados no conjunto de dados.
 - c. Informe a técnica de IA utilizada para o modelo classificador.
 - d. Mostre e analise os resultados do modelo apresentado nos testes realizados.
 - e. O tempo máximo de apresentação é de 20min por grupo.
 - f. Não é necessário explicar ou mostrar o código implementado.

2. Entrega e prazos

A entrega do trabalho deverá ser realizada no SIGAA até o dia indicado na tarefa **Trabalho Final** que foi cadastrada no SIGAA. Serão desconsideradas as entregas do trabalho por outros meios de comunicação ou entregas realizadas fora do prazo.

3. Método de Avaliação

A avaliação do trabalho considerará a adequação aos requisitos solicitados na descrição do trabalho e atribuirá pontos conforme foi descrito em cada requisito, totalizando 10 pontos.

Será atribuída nota 0 (zero) para todos os estudantes que não entregarem o trabalho dentro do prazo estipulado, que plagiarem o trabalho em sua totalidade ou em considerável parte de um colega ou de fonte externa, ou que não realizarem a apresentação para a turma.