

CP5 – Modelo de Classificação com IA

Trabalho Individual

Objetivo Geral

Desenvolver um modelo de classificação supervisionado de Machine Learning utilizando um dataset de livre escolha. O estudante deverá aplicar um pipeline completo de ciência de dados, incluindo exploração, preparação dos dados, treinamento, avaliação do modelo e interpretação dos resultados.

A entrega é individual e deve demonstrar aplicação prática, raciocínio crítico e clareza metodológica.

Contexto do Entregável

A inteligência artificial aplicada à classificação é um dos pilares de sistemas modernos como detecção de fraudes, diagnósticos médicos, análise de sentimentos e recomendação de conteúdo. Nesta atividade, cada aluno deverá:

- Escolher um problema real de classificação.
- Selecionar um dataset confiável.
- Definir uma estratégia metodológica clara.
- Construir e avaliar modelos de IA.
- Apresentar os resultados de forma objetiva e técnica.
- Avaliar criticamente um trabalho de um colega (avaliação por pares).

Fontes para Encontrar Datasets

É obrigatório usar dados públicos e éticos (sem informações pessoais sensíveis). Sugestões:

Plataforma	Link
Kaggle	https://www.kaggle.com/datasets
UCI Machine Learning Repository	https://archive.ics.uci.edu
Hugging Face Datasets	https://huggingface.co/datasets
Google Dataset Search	https://datasetsearch.research.google.com
OpenML	https://www.openml.org
Dados públicos do governo brasileiro	https://dados.gov.br

Requisitos do dataset: mínimo 300 registros e pelo menos 5 variáveis explicativas.

Pipeline Obrigatório do Projeto

O trabalho deve seguir, no mínimo, as seguintes etapas:

1. Definição do Problema

- Objetivo de classificação (binária ou multiclasse).
- Justificativa da escolha do tema.

2. Descrição do Dataset

- Origem e fonte do dado.
- Quantidade de linhas/colunas.

- Descrição das variáveis.

3. Pré-processamento dos Dados

- Checagem e tratamento de valores ausentes.
- Tratamento de outliers (se necessário).
- Encoding de variáveis categóricas.
- Normalização/Padronização (quando aplicável).

4. Modelagem

- Testar pelo menos 3 modelos de classificação:
 - Regressão Logística
 - Análise de Discriminante Linear (LDA)
 - Gaussian Naive Bayes
 - k-vizinhos mais próximos (KNN)
 - Máquina de Vetor Suporte (SVM)
 - Árvore de Decisão
 - RandomForest
- Utilização de `train_test_split`.

5. Avaliação do Modelo

- Usar no mínimo 3 métricas de avaliação:
 - Acurácia
 - Precisão
 - Recall
 - F1-Score
 - Matriz de Confusão

6. Interpretação dos Resultados

- análise comparativa entre os modelos.
- justificar o modelo final escolhido.

7. Conclusão

- principais aprendizados.
- possíveis melhorias futuras.

Ferramentas Recomendadas

- Linguagem: Python
- Bibliotecas: pandas, numpy, scikit-learn, matplotlib, seaborn
- Ambiente sugerido: Google Colab ou Jupyter Notebook

Bibliografia e Referências de Apoio

- Géron, Aurélien – Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow
- Bishop, Christopher – Pattern Recognition and Machine Learning
- Hastie, Friedman & Tibshirani – The Elements of Statistical Learning
- Mitchell, Tom – Machine Learning
- McKinney, Wes – Python for Data Analysis
- Documentação Scikit-Learn: <https://scikit-learn.org>

Entregáveis

- Repositório no GitHub contendo o dataset, o notebook jupyter (.ipynb) - que pode ser feito no google colab - com código e comentários explicando a metodologia e resultados e o README com as instruções.

Avaliação por Pares

Cada estudante fará a avaliação do trabalho de um colega de forma anônima, o qual será encaminhado pelo docente. A avaliação seguirá critérios técnicos e será parte da nota final.

Rubrica de Avaliação

Critério	Peso	Descrição
Escolha do problema e dataset	10%	Relevância, clareza e justificativa
Organização e clareza do código	15%	Estrutura, comentários e legibilidade
Pré-processamento de dados	20%	Tratamento correto e boas práticas
Modelagem e experimentação	20%	Diversidade e consistência
Métricas de avaliação	15%	Correção e interpretação
Conclusões e análise crítica	10%	Coerência e profundidade
Avaliação por pares	10%	Qualidade do feedback dado e recebido

Avaliação por pares: até 2 dias após a entrega dos projetos.