Instituto Nacional de Telecomunicações - Inatel

AG002 – Engenharia de Computação

Prof. Me. Marcelo Vinícius Cysneiros Aragão
Prof. Me. Renzo Paranaíba Mesquita

1 Introdução

Neste semestre a AG002 acontecerá na forma de um trabalho prático. Você deverá utilizar seus conhecimentos de Programação, Bancos de Dados e Inteligência Artificial para, a partir do conjunto de dados proposto, treinar, avaliar e disponibilizar um modelo de aprendizado de máquina para classificar dados relacionados ao câncer de mama.

2 Conjunto de Dados

O câncer de mama é um dos tipos mais frequentes em mulheres e uma das principais causas de morte por câncer. Portanto, estabelece-se a necessidade de remover o tumor precocemente para reduzir a recorrência da doença. Uma vez que a recorrência dentro de 5 anos após o diagnóstico está correlacionada com a chance de óbito, compreender e prever a susceptibilidade à recorrência são tarefas críticas. [1]

O conjunto de dados apresenta 286 amostras com dados referentes a pacientes que foram submetidas a cirurgia para remoção de tecido canceroso, no Instituto de Oncologia do Centro Médico Universitário, Liubliana, Eslovênia (então Iugoslávia). São 9 atributos que podem ser utilizados para indicar (ou não) a recorrência deste tipo de doença.

Neste trabalho será utilizada uma versão modificada do conjunto originalmente concebido pelos médicos Zwitter e Soklic [2] em 1988. Os dados originais foram obtidos do UCI Machine Learning Repository, que foram codificados de acordo com uma *codetable* por meio de um *script*.

3 Etapas para Realização

1. Instalar o banco de dados MySQL.
2. Baixar e executar o *script* para criação do *schema* e importação dos dados.
3. Fazer a leitura dos dados utilizando Pandas ou JDBC, por exemplo.
4. Escolher um dos modelos de classificação a seguir:
   - Decision Tree: Wikipedia, KDnuggets e scikit-learn.
   - k-Nearest Neighbors: Wikipedia, Towards Data Science e scikit-learn.
   - Multilayer Perceptron: Wikipedia, KDnuggets e scikit-learn.
   - Naïve Bayes: Wikipedia, Towards Data Science e scikit-learn.
   - Perceptron: Wikipedia, Towards Data Science e scikit-learn.
5. Separar o conjunto de dados em duas partes: 80% para treinamento e 20% para testes.
   - Treinar o modelo escolhido usando 80% dos dados.
   - Avaliar o modelo escolhido usando os 20% restantes.
6. Exibir métricas de avaliação no terminal.
7. Criar uma opção que permita ao usuário inserir dados arbitrários que devem ser classificados pelo modelo. O modelo deverá imprimir se, com base no conhecimento adquirido com os dados do conjunto, os dados inseridos constituem chance de recorrência da doença ("sim" ou "não"). Dica: método `predict`.

4 Orientações Adicionais

- O trabalho deverá ser feito em dupla;
- Qualquer linguagem de programação pode ser utilizada;
- A entrega deverá ser feita por meio de um arquivo zip com todo o conteúdo do projeto, ou o link de um repositório privado do GitHub;
- Para apresentação, o aluno deverá gravar um vídeo de no máximo 7min de duração, explicando em detalhes as etapas do projeto desenvolvido;
- O vídeo poderá ser feito gravando a própria tela do computador enquanto o aluno explica ou até mesmo ser usado o *smartphone*, desde que as explicações das etapas estejam nítidas;
- A entrega deve ser feita até o dia 25/06/2023. Disponibilize vídeo e arquivo zip (se for usar) no OneDrive, com permissão de acesso para renzo@inatel.br. Se usar GitHub (em vez de arquivo zip), disponibilize o link também com permissão de acesso.

Referências

[1] Clara Sorensen e Yujue Wu. "Predictors for Breast Cancer Recurrence". Em: *Undergraduate Statistics Class Project (USCLAP)* (2018).

[2] Matjaz Zwitter e Milan Soklic. *Breast Cancer Data Set*. 1988.