Instituto Nacional de Telecomunicações - Inatel

AG002 – Engenharia de Computação

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

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 a crédito.

2 Conjunto de Dados

O conjunto de dados apresenta 1000 amostras, datadas de 1973 a 1975, com dados referentes à análise de crédito de um banco regional do sul da Alemanha. São 20 atributos que podem ser utilizados para classificar bons e maus candidatos a empréstimo. Neste trabalho será utilizada uma versão corrigida [1] do conjunto originalmente doado pelo professor Hans Hofmann (Universidade de Hamburgo) para o projeto europeu Statlog em 1994 [2].

- O conjunto de dados foi obtido do UCI Machine Learning Repository.
- Os atributos estão em alemão, e os dados estão codificados de acordo com uma *codetable*.

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, para que possa ser verificada a acurácia do modelo.
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 risco de crédito "bom" ou "ruim". Dica: utilize a função `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 pela Plataforma Teams (pela tarefa criada) até o dia 19/11/2022. Disponibilize vídeo (ou link de acesso ao vídeo) e arquivo zip com o código (se for usar). Se usar GitHub (no lugar de arquivo zip), disponibilize link também com permissão de acesso para guilherme@inatel.br.

Referências

[1] U Groemping. South german credit data: Correcting a widely used data set. *Rep. Math., Phys. Chem., Berlin, Germany, Tech. Rep*, 4:2019, 2019.

[2] Hans-Joachim Hofmann. Die anwendung des cart-verfahrens zur statistischen bonitätsanalyse von konsumentenkrediten. *Zeitschrift fur Betriebswirtschaft*, 60:941–962, 1990.