TÍTULO: CLASSIFICAÇÃO E AGRUPAMENTO DE DADOS UTILIZANDO CONCEITOS DOS ALGORITMOS K-NN E K-MEANS

Instituição: UEMS – Universidade Estadual de Mato Grosso do Sul

Área temática: Ciências Exatas e da Terra

HANAOKA, Mario Massao[1] (mario.science@tutanota.com); **RUBIO-MERCEDES,** Cosme Eustaquio[2] (cosme@uems.br).

[1] – Aluno de graduação em Engenharia Física;
[2] – Orientador/Docente do curso de Engenharia Física.

Introdução: A área de aprendizado de máquina (*machine learning*) pode ser considerada um subconjunto da Inteligência Artificial. Técnicas baseadas em aprendizado de máquina complementam as técnicas estatísticas tradicionais e tem sido aplicadas com sucesso nas mais diversas áreas com a finalidade de analisar dados, detectar padrões e elaborar previsões. Alguns exemplos são: visão computacional, engenharias, finanças, entretenimento, biologia computacional e medicina. Avanços em computação, coleta e armazenamento de dados possibilitaram a aplicação de métodos de aprendizagem de máquina para a extração de informações de grandes quantidades de dados – também conhecidos como *Big Data*. Anteriormente, essa tarefa era considerada impraticável ou impossível. O aprendizado de máquina permite ao computador aprender através da experiência, isto é, sem a necessidade de ser explicitamente programado pelo usuário para cada nova tarefa. **Objetivos:** Na primeira parte do projeto, estudar fundamentos de programação em python, bases de dados e estatística com aplicações no aprendizado de máquina. Na segunda etapa, aprofundar os estudos em linguagem python, juntamente com a implementação dos algoritmos k-Nearest Neighbors (k-NN) e k-Means para as tarefas de classificação e agrupamento de dados, respectivamente. **Metodologia:** A linguagem de programação escolhida para desenvolver o projeto foi python e o ambiente de notebook escolhido foi o Google Colab. Para a tarefa de classificação utilizando o método k-NN, o conjunto de dados escolhido foi o Breast Cancer Wisconsin, que consiste em dados reais retirados de imagens e exames de câncer de mama, fornecido pela University Of California Irvine (UCI). Para a tarefa de agrupamento utilizando o método k-Means, o dataset sintético *make_blobs*, da biblioteca *scikit-learn*, foi escolhido. **Resultados:** Ao aplicar o método k-NN sobre o conjunto de dados *breast_cancer*, o modelo de classificação foi treinado, resultando em uma ácurácia de treino de 93,0% e acurácia de teste de 93,7%. A obtenção de uma acurácia de teste maior que a de treino sugere a ocorrência de um leve sobreajuste nos dados (*overfitting*). Além disso, foram utilizadas outras métricas para avaliar o modelo, tais como precisão, sensitividade e pontuação-F1. Os valores dessas métricas estiveram entre 90% e 98%, considerados excelentes. Como análise complementar, foi desenhada a curva ROC, que esteve conformidade com os demais resultados. Isso significa que, na maioria das vezes, o modelo é capaz de predizer corretamente o diagnóstico de câncer de mama para novos pacientes. Ao aplicar o método k-Means sobre os dados de *make_blobs*, a maioria das instâncias foi associada ao cluster apropriado. No entanto, há poucas instâncias que foram provavelmente rotuladas erroneamente – especialmente próximo à fronteira entre o cluster do topo esquerdo e o cluster central. Esse resultado era esperado, pois o k-Means não se comporta bem, de maneira geral, quando os clusters tem diâmetros muito distintos. O principal parâmetro para o k-Means é a distância do ponto até o centroide, o que pode causar confusão em regiões de fronteira. **Conclusão:** Os modelos apresentaram um desempenho satisfatório, tanto para a tarefa de classificação quanto para a tarefa de agrupamento. Os resultados foram condizentes com o que se esperava com base na literatura.

PALAVRAS-CHAVE: machine learning, análise de dados, python.

AGRADECIMENTOS: Esse projeto de iniciação científica só foi possível com os auxílios do CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico – e do orientador Prof. Dr. Cosme Eustaquio Rubio Mercedes. Meus sinceros agradecimentos a todos os agentes que possibilitam o ensino, pesquisa e extensão no Brasil.