HƯỚNG DẪN THỰC HIỆN BÀI TẬP LỚN CHO SINH VIÊN
Học phần Học máy - Học kỳ 1 năm học 2024-2025

I. Nội dung của bài tập lớn

Xây dựng mô hình học máy dựa trên thuật toán đã học để giải quyết vấn đề thực tế dựa trên tập dữ liệu có sẵn. Nội dung của báo cáo bao gồm:

- Mô tả thuật toán.

- Mô tả tập dữ liệu, trình bày quy trình tiền xử lý dữ liệu.

- Phân tích mã nguồn.

- Thử nghiệm mô hình học máy với các bộ tham số khác nhau, lưu lại kết quả sau mỗi lần thử nghiệm.

- Phân tích kết quả thử nghiệm và lựa chọn bộ tham số tốt nhất.

- Hiểu được các đặc điểm của bộ dữ liệu, các hạn chế của thuật toán học máy thể hiện trên bộ dữ liệu.

II. Chủ đề bài tập lớn

1. Xây dựng mô hình K-NN dự đoán việc huỷ đặt phòng khách sạn

   https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset

2. Xây dựng mô hình K-NN dự đoán phê duyệt khoản vay

   https://drive.google.com/file/d/1LIvIdqdHDFEGnfzIgEh4L6GFirzsE3US/view

3. Xây dựng mô hình K-NN dự đoán lương của nhân sự trong lĩnh vực công nghệ dữ liệu

   https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023

4. Xây dựng mô hình K-NN hỗ trợ chẩn đoán bệnh tiểu đường

   https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

5. Xây dựng mô hình K-NN dự đoán khách hàng có ý định chấm dứt sử dụng dịch vụ của ngân hàng

   https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset/data

6. Xây dựng mô hình Naive Bayes phát hiện email lừa đảo

   https://www.kaggle.com/datasets/subhajournal/phishingemails

7. Xây dựng mô hình Naive Bayes đánh giá điểm tín dụng

   https://www.kaggle.com/code/shivangamsoni/credit-score-classofication-naive-bayes

8. Xây dựng mô hình Naive Bayes phân loại lĩnh vực của bản tin

   https://www.kaggle.com/datasets/rmisra/news-category-dataset/data

9. Xây dựng mô hình Naive Bayes để dự đoán mối liên kết giữa các kết nối internet với hoạt động tấn công mạng

   https://www.kaggle.com/datasets/agungpambudi/network-malware-detection-connection-analysis/data

10. Xây dựng mô hình Naive Bayes hỗ trợ chẩn đoán bệnh tim

    https://www.kaggle.com/datasets/data855/heart-disease/data

11. Xây dựng mô hình cây quyết định phát hiện gian lận thẻ tín dụng

    https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023

12. Xây dựng mô hình cây quyết định hỗ trợ chẩn đoán bệnh ung thư vú

    https://www.kaggle.com/datasets/reihanenamdari/breast-cancer

13. Xây dựng mô hình cây quyết định phát hiện gian lận trong thanh toán trực tuyến

    https://www.kaggle.com/datasets/rupakroy/online-payments-fraud-detection-dataset/data

14. Xây dựng mô hình cây quyết định dự đoán cơ hội xin được việc làm của sinh viên sau khi tốt nghiệp

    https://www.kaggle.com/datasets/benroshan/factors-affecting-campus-placement/data

15. Xây dựng mô hình cây quyết định để dự đoán một đơn hàng mua online có được giao đúng hạn không

    https://www.kaggle.com/datasets/prachi13/customer-analytics/data

16. Xây dựng mô hình K-Means để phân nhóm khách hàng nhằm triển khai chiến lược marketing phù hợp (bộ dữ liệu sẽ được tải lên LMS)

17. Xây dựng mô hình K-Means để phân khúc thị trường dựa trên dữ liệu giao dịch bán

buôn

https://www.kaggle.com/datasets/tunguz/online-retail/data

18. Xây dựng mô hình K-Means để phân nhóm tình hình thời tiết

    https://www.kaggle.com/datasets/julianjose/minute-weather/data

19. Xây dựng mô hình K-Means để phân cụm các công ty chứng khoán trên thị trường chứng khoán Việt Nam

    Hướng dẫn: cài đặt thư viện vnstock để tải dữ liệu chứng khoán Việt Nam

    https://pypi.org/project/vnstock/

20. Xây dựng mô hình K-Means để xác định nhóm khách hàng mục tiêu của công ty dựa trên dữ liệu bán lẻ trực tuyến xuyên quốc gia

    https://www.kaggle.com/datasets/hellbuoy/online-retail-customer-clustering

III. Hình thức thực hiện

- Sinh viên tự do chọn nhóm với điều kiện mỗi nhóm từ 2 đến 4 thành viên.
- Đề tài được chọn ngẫu nhiên trong buổi học tuần 6.
- Sinh viên nộp quyển báo cáo bài tập lớn (**ít nhất 30 trang A4**) đảm bảo đầy đủ nội dung theo đúng mẫu quy định của nhà trường.