Khai phá dữ liệu

A. Project nhóm giữa kì

1. Tham khảo các cơ sở dữ liệu công khai trên các trang web sau:

- Trang web của Trung tâm Máy tính và Hệ thống Thông tin (UCI) của Đại học California.
  Link: https://archive.ics.uci.edu/ml/datasets.php
- Trang web của Kaggle - một cộng đồng chuyên về Khoa học Dữ liệu và Máy học.
  Link: https://www.kaggle.com/datasets
- Trang web của Google Dataset Search.
  Link: https://datasetsearch.research.google.com

Khi chọn bộ dữ liệu cần xác định các tiêu chí:

- **Xác định loại bài toán cụ thể**: Bài toán phân loại (classification), bài toán phân cụm (clustering), bài toán phân tích kết hợp (association analysis).
- **Kích thước dữ liệu**: tối thiểu 200 dòng.
- **Chiều của dữ liệu**: tối thiểu 5 cột.
- **Dữ liệu đầu vào**: Kiểm tra tính đầy đủ và độ chính xác của dữ liệu đầu vào.
- **Kết quả đầu ra**: Xác định dữ liệu đầu ra để có thể đưa ra kết luận từ phân tích dữ liệu.

2. Phân tích dữ liệu khám phá

2.1. Tiền xử lý dữ liệu

- Xác định và loại bỏ các dữ liệu không cần thiết
- Xử lý dữ liệu thiếu
- Xử lý dữ liệu bị trùng
- Xử lý dữ liệu ngoại lai
- Xử lý dữ liệu nhiễu

2.2. Khám phá dữ liệu

2.2.1. Phân tích thống kê mô tả

- Sử dụng các thống kê mô tả như giá trị trung bình, trung vị, độ lệch chuẩn, phân vị để hiểu về phân bố và đặc điểm của các biến trong dữ liệu.

2.2.2. Phân tích đơn biến

- Xem xét phân bố của từng biến trong dữ liệu bằng cách sử dụng histogram, box plot, kernel density plot.
- Phân tích các giá trị ngoại lai (outliers) và các giá trị bất thường (anomalies) trong dữ liệu.

2.2.3. Phân tích đa biến

- Tìm kiếm các mối quan hệ giữa các biến trong dữ liệu bằng cách sử dụng ma trận tương quan, heatmap.
- Thực hiện phân tích chuỗi thời gian (time series analysis) nếu dữ liệu là dữ liệu chuỗi thời gian.

3. Thuyết trình

- Mỗi nhóm thuyết trình về nội dung mình đã làm trong 15 phút.

B. Project cá nhân cuối kì

1. Mỗi cá nhận đặt ra ít nhất 2 câu hỏi về dữ liệu.

Bài toán phân loại

- Dựa vào yếu tố nào để xác định một email là thư rác hay thư thường.
- Yếu tố nào giúp chúng ta xác định một khách hàng tiềm năng hay không tiềm năng.

Bài toán phân cụm

- Phân cụm khách hàng theo đặc tính mua sắm, ví dụ như phong cách thời trang yêu thích, loại sản phẩm thường mua, tần suất mua hàng.
- Phân cụm khách hàng của một ngân hàng theo loại tài khoản và hình thức thanh toán sử dụng.
- Phân cụm sản phẩm theo thuộc tính như kích thước, màu sắc, giá cả.
- Phân cụm bài viết trên mạng xã hội theo chủ đề và mức độ phổ biến.

Bài toán phân tích kết hợp

- Tìm ra các quy luật và mối liên hệ giữa các sản phẩm được mua cùng nhau trên một trang web mua sắm.
- Phân tích mối tương quan giữa giá cổ phiếu và chỉ số tài chính để dự đoán giá cổ phiếu trong tương lai.

- Phân tích dữ liệu về đơn đặt hàng và dữ liệu khách hàng để tìm ra mối liên hệ giữa khách hàng tiềm năng và các sản phẩm đang được bán trên trang web.
- Phân tích dữ liệu về lượng truy cập và thông tin về khách hàng để tìm ra các xu hướng trong quá trình mua sắm trực tuyến.

2. **Dựa trên kết quả phân tích giữa kì, mỗi cá nhân áp dụng từ 2 đến 3 thuật toán máy học để khai phá dữ liệu.**

   Thuật toán phân loại (Classification):

   - Logistic Regression
   - Naive Bayes Classifier
   - Decision Tree
   - Random Forest
   - K-Nearest Neighbors (KNN)
   - Support Vector Machine (SVM)
   - XGboost

   Thuật toán phân cụm (Clustering):

   - K-Means Clustering
   - Hierarchical Clustering
   - Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
   - Mean Shift Clustering
   - Fuzzy C-Means Clustering (FCM)

   Thuật toán phân tích kết hợp (Association):

   - Apriori Algorithm
   - Eclat Algorithm
   - FP-Growth Algorithm
   - K-Means Association Rule Mining
   - Classification Based Association Rule Mining (CBA)
   - Sequential Pattern Mining Algorithm (SPM)

3. Đánh giá mô hình và trả lời 2 câu hỏi đã đề ra.

4. Viết báo cáo.

I - Giới thiệu: Phần này giới thiệu về đề tài, mục đích và phạm vi của dự án khai phá dữ liệu.

II - Mô tả bộ dữ liệu: Phần này mô tả về bộ dữ liệu được sử dụng trong dự án, bao gồm nguồn gốc của dữ liệu, số lượng mẫu, số lượng thuộc tính và các giá trị bị khuyết trong dữ liệu.

II - Tiền xử lý dữ liệu: Phần này mô tả về quá trình tiền xử lý dữ liệu, bao gồm các bước như lọc dữ liệu, xử lí giá trị bị khuyết, rút trích đặc trưng và chuẩn hóa dữ liệu.

III – Phân tích dữ liệu khám phá: Phần này mô tả về quá trình khai phá dữ liệu, bao gồm các bước như phân tích đa biến, phân tích đơn biến, phân tích tương quan.

IV - Khai phá dữ liệu: Phần này đặt ra các câu hỏi về dữ liệu và đề xuất các thuật toán phù hợp để khai phá dữ liệu và trả lời câu hỏi.

V - Đánh giá và chọn thuật toán: Phần này đánh giá kết quả của các thuật toán được áp dụng trong quá trình khai phá dữ liệu và chọn thuật toán tốt nhất để giải quyết các câu hỏi trong đề tài.

VI - Kết quả và thảo luận: Phần này trình bày các kết quả của quá trình khai phá dữ liệu và thảo luận về những khía cạnh quan trọng của kết quả, bao gồm những điểm mạnh và điểm yếu của kết quả.

VII - Kết luận: Phần này tổng kết lại các kết quả của dự án khai phá dữ liệu và đưa ra kết luận về hiệu quả và khả năng áp dụng của quá trình khai phá dữ liệu.

Tài liệu tham khảo: Phần này liệt kê các tài liệu đã sử dụng để tham khảo trong quá trình thực hiện dự án.