Đề thi:

BIG DATA IN MACHINE LEARNING

Hạn chót nộp bài: 23h30 – Thứ Hai ngày 27/11/2023

**** HV tạo 1 thư mục **LDS9_K289_ONLINE_HoVaTen_Cuoi_ky** trong thư mục **LDS9_K289_ONLINE_HoVaTen** trên Google Drive đã share, lưu tất cả bài làm vào để chấm điểm ****

**** HV sẽ bị trừ điểm nếu bài làm giống nhau ****

**** HV gửi mail đính kèm link của thư mục **LDS9_K289_ONLINE_HoVaTen_Cuoi_ky** đúng hạn nộp bài, sau hạn nộp bài nếu HV không gửi thì sẽ không được chấm điểm ****

Chú ý, với mỗi câu:

- HV cần kiểm tra xem dữ liệu đã sạch, chuẩn và dùng được hay chưa, nếu chưa thì cần tiền xử lý trước khi làm bài.
- Cần hiển thị thông tin chung của dữ liệu để có cái nhìn ban đầu về dữ liệu.
- Trong dữ liệu có thể có rất nhiều thông tin (feature/column), cần xác định xem thông tin nào cần thiết dùng trong thuật toán thì đưa vào, không cần thiết thì không đưa.
- Mỗi câu là một file viết trên jupyter notebook/colab, các yêu cầu nhận xét kết quả trong từng câu được viết trong cell dưới định dạng Markdown.
- Mỗi câu đều phải đưa ra nhận xét, giải pháp cho các lựa chọn.
- Câu nào có trực quan hóa kết quả thì vừa phải trực quan vừa phải giải thích.
- Từ *Question 1 đến Question 3*, cần phải đề xuất từ 2 thuật toán trở lên cho mỗi câu sau đó dựa trên kết quả để chọn thuật toán phù hợp hơn. Nêu rõ lý do tại sao chọn.
- *Cần phải in các dòng hiển thị kết quả ngay sau từng bước để GV đọc và chấm điểm*, GV chỉ « run » lại bài làm của HV khi thấy bài làm có vấn đề vì thời gian để thực thi cho một bài khá dài.

Question 1: Regression - Gemstone Price Prediction (1.0 mark)

Use **cubic_zirconia.csv** dataset (in folder Gemstone) to build a model to predict "price of gemstone" (Inputs: select suitable features, Output: **price**)
Then, make new prediction:

- If the information about a gemstone is as follows:

carat	cut	color	clarity	depth	table	x	y	z
1.45	Fair	G	VS2	65.2	54	7.2	7.1	4.6

- What is its price?

Read more information here:
https://www.kaggle.com/datasets/colearninglounge/gemstone-price-prediction

Question 2: Classification – Home Loan Approval (1.0 mark)

Use **loan_sanction_train.csv** dataset (in folder HomeLoan) to build a model to predict "Whether a customer will be approved for a loan or not" (Inputs: select suitable features, Output: **Loan_Status**)
Then, make new prediction for the new customers in **loan_sanction_test.csv** (in folder HomeLoan):

- Will they be approved for a loan or not?

Read more information here:

https://www.kaggle.com/datasets/rishikeshkonapure/home-loan-approval/data

Question 3: Classification - Fake and real news (1.0 mark)

Use **fake-and-real-news-dataset** to build a model to determine "if an article is fake news or not".

Read more information here:

https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

Question 4: Clustering – Movies Clustering (1.0 mark)

Build a clustering model to cluster the movies in **tmdb_5000_movies.csv** dataset. Explain the main characteristics of each cluster. Use Word Cloud to visualize each cluster.

(Hint: Use some features such as 'title', 'tagline', 'overview', 'popularity'…)

Câu 5: Recommendation - Amazon – Toys and Games (1.0 mark)

If you have more than two laptops/computers:

Use the information "reviewerID" (first column), "asin" (ProductID, second column), and "overall" (users' ratings for each product, third column) in dataset **ratings_Toys_and_Games.csv** to build a model to predict overalls for products that have not been selected by users.

Then make recommendations to some users: AGJ5VZ9QDM7JK, A30Q4EO6S4NG3D, AJFC5966PYS6N

If you have one laptop/computer:

Use the information "reviewerID", "asin" (ProductID), and "overall" (users' ratings for each product) in dataset **reviews_Toys_and_Games_5.json.gz** to build a model to predict overalls for products that have not been selected by users. Then make recommendations to some users: A3GJPLCZCDXXG6, A34U85WY8ZWBPV, A2VIY2TL6QPYLG

Read more information here:

https://cseweb.ucsd.edu/~jmcauley/datasets/amazon/links.html

Câu 6: Association Rules – BAKERY (1.0 mark)

Use dataset **75000** (select suitable files in this folder) to build the model to identify sets of items that are frequently bought together with two cases: use Id and use Flavor and Food name (in **goods.csv**).

Read more information here:

http://users.csc.calpoly.edu/~dekhtyar/466-Spring2018/labs/lab2.466.pdf

--- Good luck 👌 ---