Đề thi:

BIG DATA IN MACHINE LEARNING

Hạn chót nộp bài : 23h00 ngày 30/04/2020

**** HV tạo 1 project là LDS9_HoVaTen, lưu tất cả bài làm vào để nộp chấm điểm ****

**** HV được sử dụng tài liệu ****

**** Điểm trừ : Bài làm giống nhau ****

Chú ý, với mỗi câu:

- HV cần kiểm tra xem dữ liệu có bị thiếu (NaN, null, hoặc để trống) hay không, nếu có thì cần chuẩn hóa trước khi làm bài.
- Cần hiển thị thông tin chung của dữ liệu show(), printSchema()... để có cái nhìn ban đầu về dữ liệu.
- Lần lượt thực hiện các bước làm bài như đã được hướng dẫn làm bài tập trong lớp.
- Mỗi câu là 1 file, các yêu cầu nhận xét kết quả trong từng câu được viết trong cell dưới định dạng Markdown.

Câu 1: Women's E-Commerce Clothing Reviews

Can you use **womens-ecommerce-clothing-reviews** dataset (file Womens_Clothing_E_Commerce_Reviews.xlsx, sheetName: **Reviews**) to build a model to predict products' ratings (based on Review Text and other optional features)
Please predict ratings for products in Womens_Clothing_E_Commerce_Reviews.xlsx, sheetName: **new_reviews**.
Read more information here:
https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews

Câu 2: Fake and real news dataset

Can you use **fake-and-real-news-dataset** 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

Câu 3: Combined Cycle Power Plant

Can you use **CCPP** dataset to build the model to predict the net hourly electrical energy output (EP) of the plant based on Features consist of hourly average ambient variables Temperature (T), Ambient Pressure (AP), Relative Humidity (RH) and Exhaust Vacuum (V)
Read more information here:
http://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant

Câu 4: CBC News Coronavirus articles

Can you build a clustering model to cluster the articles in **cbc-news-coronavirus-articles-march-26** dataset? Explain the main characteristics of each cluster.
(Hint: Use feature *description* and/or *text*)
Read more information here:
https://www.kaggle.com/ryanxjhan/cbc-news-coronavirus-articles-march-26

Câu 5: Amazon - Grocery and Gourment Food

Use the information "asin" (ProductID), "reviewerID" and overall (users' ratings for each product) in dataset **ratings_Grocery_and_Gourmet_Food.csv** to build a model to predict overalls for products that haven't been selected by users. Then make recommendations to some users: A3ABZBEG3KZ0L, A2BSUJYATHI7WW, A26LKBXTSIHQV2
Read more information here:
http://jmcauley.ucsd.edu/data/amazon/

Câu 6: BAKERY

Use dataset **75000** (select one file in this folder that is suitable for you) to build the model to identify sets of items that are frequently bought together (please use Flavor and Food name (in **goods.csv**) instead of Id).
Read more information here: http://users.csc.calpoly.edu/~dekhtyar/466-Fall2010/labs/lab2/extendedBakery.pdf
Dataset: http://users.csc.calpoly.edu/~dekhtyar/466-Spring2018/labs/lab01.html

--- Chúc các bạn làm bài tốt!!! ---