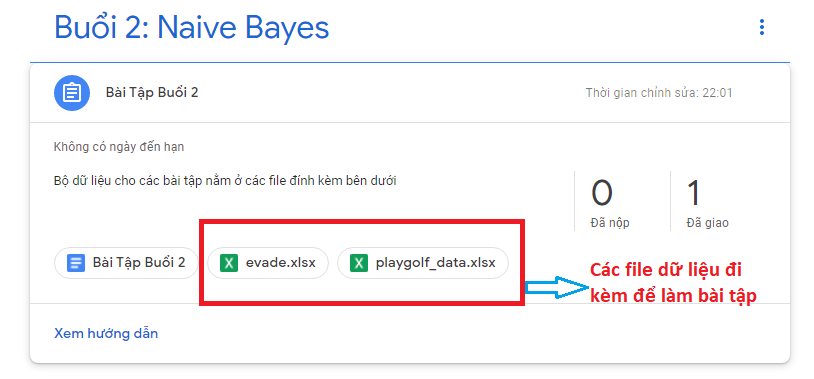
**Bài Tập Buổi 2**

Lưu ý: Bộ dữ liệu đi kèm để làm các bài tập được đính kèm chung với file này trong classroom (Xem hình dưới)



**Exercise 1: Binary Classification**

Cho bộ dữ liệu Evade nằm trong file Evade.xlsx.

Yêu cầu:

1. Đọc dữ liệu và huấn luyện mô hình Naive Bayes trên bộ dữ liệu này
2. Giả sử có 1 mẫu mới : (Refund = “No”, Marital\_Status = 1, Taxable\_Income = 120000), hãy tiến hành dự đoán nhãn cho mẫu này.
3. Giả sử có 1 mẫu mới : (Refund = “Yes”, Marital\_Status = 0, Taxable\_Income = 25000), hãy tiến hành dự đoán nhãn cho mẫu này
4. Viết code để save và load model ở dạng pickle (Gợi ý: [Save and Load Machine Learning Models in Python with scikit-learn - MachineLearningMastery.com](https://machinelearningmastery.com/save-load-machine-learning-models-python-scikit-learn/))

**Học viên chỉ submit duy nhất 1 file jupyter notebook cho bài tập này**

**Exercise 2: News Classification**

Cho bộ dữ liệu 20 news group của Sklearn ([sklearn.datasets.fetch\_20newsgroups — scikit-learn 1.2.2 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_20newsgroups.html)). Học viên thực hiện các yêu cầu sau:

Yêu cầu:

1. Đọc dữ liệu từ 3 nhóm:

comp.graphics', 'rec.sport.baseball', 'sci.electronics' với 2 subset là train và test, và convert data đọc được về 2 dataframe là train\_data và test\_data

1. Xây dựng mô hình Naive Bayes thích hợp và đánh giá hiệu quả của mô hình trên tập test thông qua hàm [classification\_report](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html)
2. Nhận xét mô hình thu được (ví dụ: mô hình có bị overfit/underfit không? Mô hình có hiệu quả không?)
3. Câu hỏi nâng cao: Theo bạn, model Naive Bayes có quan tâm tới trật tự các từ trong 1 mẫu câu hay không? Nếu có/ nếu không có, giải thích xem điều này có tác hại như thế nào, hay nói cách khác mô hình Naive Bayes có điểm yếu như thế nào **từ nhận định này**?

Lưu ý: Học viên save code ở dạng jupyter notebook, đối với câu 3 & 4, học viên có thể trả lời bằng cách tạo thêm 1 Markdown cell trong Jupyter notebook. **Học viên chỉ submit duy nhất 1 jupyter notebook cho câu hỏi này.**

**Exercise 3:**

Cho bộ dữ liệu PlayGolf nằm trong file PlayGolf.xlsx

**Yêu cầu:**

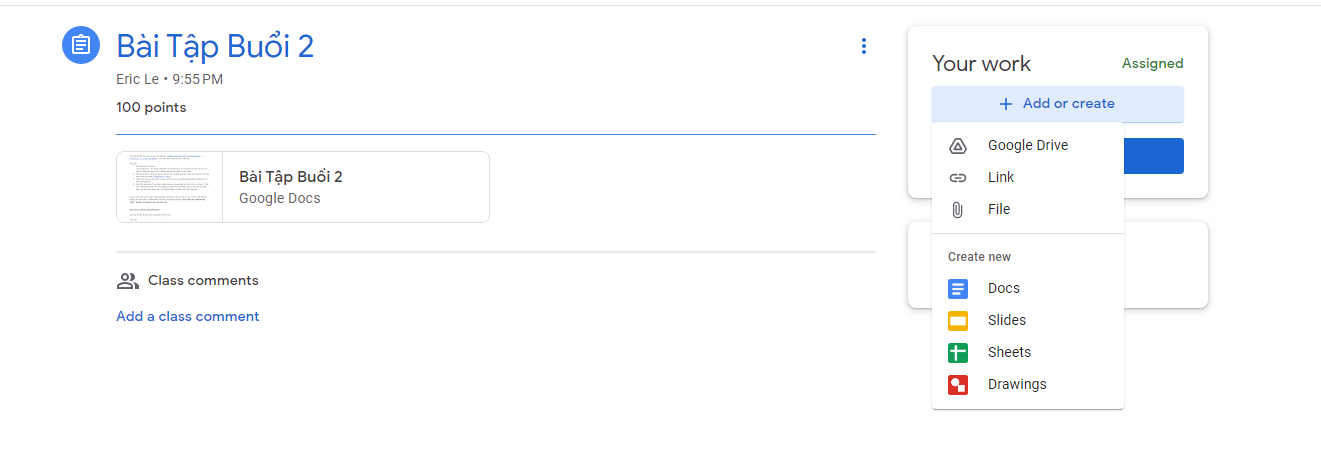
1. Đọc dữ liệu và huấn luyện mô hình Bernoulli Naive Bayes trên bộ dữ liệu này
2. Tiến hành dự đoán nhãn cho các mẫu sau: 

**Học viên chỉ submit duy nhất 1 file jupyter notebook cho bài tập này**

**Hướng dẫn submit:**

Học viên nén các file jupyter notebook trong 1 file zip và submit trực tiệp trên Google Classroom.

Học viên chọn mục “Bài Tập Buổi 2” trên classroom, ở khung bên phải, ấn “Add or create” >> File để upload file zip như hình bên dưới.

****