## Mar 8, 2024 | [ML-BTL](https://www.google.com/calendar/event?eid=NmhnZmUzYzRuNGh0Y292b3V2ZmZpZWFobG4gbmd1eWVudGhpbWFpY2hpMTcxMTIwMDNAbQ)

Attendees:

Notes:

* Đề tài: Lọc email spam
* Chọn phương thức xử lý bài toán

Action items

* Tìm data set (web dataset)
* Làm sạch data (hd)
* Tiền xử lý (tuần sau)
* Chọn đề tài:
  + NLP
  + Linear Regression
    - Logistic Regression
  + NB
  + Cây quyết định (Decision tree)
  + K Means, Hierarchical Clustering
* Data set:
  + <https://www.kaggle.com/datasets>
* Video tutorial:
  + Mail - NB: <https://youtu.be/2sXAYoPIz3A?si=gjHSQQ63ohLWHydy>
  + Preprocess: [Preprocessing Data in Python](https://youtu.be/bFCe_e4raaY?si=t6CtOmGWjBb14zCn)
  + Làm sạch data: [Data Cleaning Tutorial | Cleaning Data With Python and Pandas](https://youtu.be/OS2m0f2gVJ0?si=69k-vu0-pVxd780U)
  + Visual data: [Data Visualizations in Python](https://youtu.be/YFkFhKPNA20?si=0v5PDC7nUAwHyPmP)
  + Mail - NLP: <https://youtu.be/13ZNT9RU_aw?si=Ut1ZKAZgVNmueJ7p>
  + Classifier: [Email Spam Classifier | SMS Spam Classifier | End to End Project | Heroku Deployment](https://youtu.be/YncZ0WwxyzU?si=pji_48FpSRtoMt-0)

Mai Chi

Lọc email spam (no phishing) <3/4>

* + Logistic Regression
  + NB
  + Cây quyết định (Decision tree)
  + NLP

Sok Sokong

Xây Dựng Chương Trình Phân Loại Văn Bản Tiếng Việt <2/4>

* + NLP

Giao

Dự đoán số tiền nhà <1/4>

* + Linear Regression

Hình ảnh phân loại <2/4>

Kiên

## Mar 15, 2024 8:00 PM | [ML-BTL](https://www.google.com/calendar/event?eid=NmhnZmUzYzRuNGh0Y292b3V2ZmZpZWFobG4gbmd1eWVudGhpbWFpY2hpMTcxMTIwMDNAbQ)

Attendees: - Kiên

Notes:

* Mô tả bài toán
  + Input: mail
  + Output: label -> spam / ham
    - Ham -> spam không chấp nhận
    - Spam -> ham một lượng nhỏ
* Các phương thức xử lý bài toán và ưu, nhược
  + Logistic Regression (hồi quy logic)
  + NB (Xác suất bayes)
    - Words not phrase (có thể cải tiến)
    - Lỗi chính tả (clean data)
    - Từ mới (không có trong kho dữ liệu)
      * Alpha (a/b < (a+c)/(b+c)) (chọn alpha)
    - Stop with long mail (thống kê bên ngoài: đặt điều kiện time exceed → spam)
  + Cây quyết định (Decision tree)
  + NLP (xử ngôn ngữ tự nhiên)
* Vấn đề còn tồn tại
  + Tiền xử lý dữ liệu
    - Gán nhãn
    - Lọc nhiễu
      * Chính tả
      * Thiếu thuộc tính (label, body)
* Cách xử lý / giảm thiểu vấn đề
  + The whitelist -> ham
    - Contain word/ not contain -> spam / ham
  + The blacklist (block)