DỰ ÁN (PROJECT ASSIGNMENT)

DEADLINE: FRIDAY, 14 MARCH 2025

A. YÊU CẦU:

- Bạn sẽ phát triển một ứng dụng có sử dụng các kỹ thuật Máy học (Machine Learning) thú vị. Ứng dụng mà bạn phát triển có thể là ứng dụng Web, Windows Form, hoặc Mobile nhưng cần có sử dụng và triển khai các kỹ thuậ Máy học đã học.

- Demo, thuyết trình và báo cáo — Bạn cần đưa ra bản demo của ứng dụng mà bạn phát triển, một video thuyết trình tối đa 15 phút để giải thích những gì bạn đã thực hiện dự án của mình.

- Bạn cũng cần gửi cho tôi mã nguồn Python, các SQL scripts (nếu có), báo cáo, video thuyết trình về dự án theo thời gian deadline đã sắp xếp.

B. YÊU CẦU NỘP BÁO CÁO:

- Nhóm cần nộp 1 bản mềm (soft copy) và 1 bản cứng (hard copy)
  - Bản mềm (soft copy) gồm Python source code, SQL scripts, file báo cáo, và 1 video thuyết trình về chức năng của *app* mà bạn phát triển, đặc biệt nhấn mạnh vào các kỹ thuật máy học, việc thu thập dữ liệu, tiền xử lý, phân tích dữ liệu, các mô hình, thuật toán đã sử dụng, trong tối đa 15 phút.
  - Bản cứng (hard copy) là bản in của file báo cáo (in 2 mặt và in trắng đen).

- Bản báo cáo cần mô tả các chức năng của ứng dụng, use case diagram, kiến trúc ứng dụng, thiết kế cơ sở dữ liệu, công nghệ được sử dụng. Bản báo cáo cần được định dạng theo như các mẫu ví dụ sau:
  - QuanLyGiaPha
  - Website_NhaTroSV

- Sinh viên có thể tham khảo lại cách định dạng bản report cho bài tập dự án dùng Microsoft Word:
  - Microsoft Word - Tổng quan
  - Microsoft Word - Quản lý gia phả - 01 (Styles)
  - Microsoft Word - Quản lý gia phả - 02 (Page Section, Number)
  - Microsoft Word - Quản lý gia phả - 03 (Table of Contents, Figures, Tabs, Shapes, SmartArt)

- Microsoft Word - Quản lý gia phả - 04 (Cover page, Footnote, Endnote, Convert text to Table)

- Sinh viên có thể tham khảo các video thuyết trình dự án
  - Website tra cứu thông tin nhà ở và văn phòng cho thuê
  - Website bán đồ chơi
  - Future Kids App Demo
  - Music NFTs Streaming App Demo

- Cấu trúc thư mục (folder) và tập tin của bài nộp cần được thực hiện như sau:

```
<ProjectName>
        |__Source
        |      |__SQL Scripts (nếu có)
        |      |__Python Code
        |
        |__ MSSV_HoVaTen_Report.docx
        |__ MSSV_HoVaTen_Video.mp4
```

- Bản nộp cuối cùng, toàn bộ thư mục **<ProjectName>**, cần được nén rồi upload lên OneDrive và share cho email longhn@hiu.vn. Link bài nộp cũng cần được nộp trên Elearning.

C. DANH MỤC CÔNG VIỆC TRONG DỰ ÁN MÁY HỌC

Danh sách các công việc này có thể là kim chỉ nam trong các dự án Máy học của bạn. Sinh viên tham khảo để phát triển dự án Máy học của mình. Nó bao gồm tám bước chính sau:

1. Định hình bài toán và xem xét bức tranh tổng thể.
2. Thu thập dữ liệu.
3. Khám phá dữ liệu để có được thông tin chi tiết.
4. Chuẩn bị dữ liệu để tạo ra những các dữ liệu đầu vào phù hợp cho các thuật toán Máy học.
5. Khám phá các mô hình Máy học khác nhau và chọn ra những mô hình tốt nhất.
6. Tinh chỉnh các mô hình và kết hợp chúng lại thành một giải pháp hoàn hảo hơn.
7. Trình bày giải pháp của bạn.
8. Khởi chạy, giám sát, và bảo trì hệ thống của bạn.

Tất nhiên là bạn có thể tự do điều chỉnh danh sách trên cho phù hợp với nhu cầu cụ thể.

I. Định hình Bài toán và Xem xét Bức tranh Tổng thể

1) Xác định mục tiêu theo phương diện kinh doanh.
2) Giải pháp của bạn sẽ được sử dụng như thế nào?
3) Các giải pháp hiện tại (nếu có) là gì?
4) Bạn sẽ định hình bài toán này như thế nào (có giám sát/không giám sát, trực tuyến/ngoại tuyến, v.v.)?
5) Chất lượng của giải pháp sẽ được đo lường như thế nào?
6) Thước đo chất lượng này có phù hợp với mục tiêu kinh doanh không?
7) Chất lượng mô hình tối thiểu cần thiết để đạt được mục tiêu kinh doanh là bao nhiêu?
8) Các bài toán tương tự là gì? Bạn có thể sử dụng lại kinh nghiệm hoặc công cụ có sẵn được không?
9) Yếu tố kiến thức chuyên môn có cần thiết không?
10) Bạn sẽ giải quyết bài toán theo cách thủ công như thế nào?
11) Liệt kê những giả định mà bạn (hoặc những người khác) đã đưa ra cho đến thời điểm hiện tại.
12) Kiểm chứng các giả định nếu có thể.

II. Thu thập Dữ liệu

Lưu ý: hãy tự động hóa càng nhiều càng tốt để có thể thu thập được dữ liệu mới một cách dễ dàng.

1) Liệt kê loại dữ liệu và số lượng mà bạn cần.
2) Tìm và ghi lại nơi bạn có thể thu thập được dữ liệu đó.
3) Kiểm tra xem dữ liệu sẽ chiếm bao nhiêu dung lượng.
4) Kiểm tra các nghĩa vụ pháp lý và xin quyền sử dụng nếu cần.
5) Xin quyền truy cập.
6) Tạo một môi trường làm việc (có đủ dung lượng lưu trữ).
7) Thu thập dữ liệu.
8) Chuyển đổi dữ liệu thành định dạng mà bạn có thể dễ dàng thao tác (mà không làm thay đổi thông tin dữ liệu).
9) Đảm bảo các thông tin nhạy cảm phải được bảo vệ hoặc xóa bỏ (ví dụ, dạng ẩn danh).
10) Kiểm tra kích thước và dạng dữ liệu (dạng chuỗi thời gian, mẫu, không gian địa lý, v.v.).
11) Lấy mẫu để tạo tập dữ liệu kiểm tra, để nó sang một bên và không được đụng vào (không được xem lén dữ liệu!).

III. Khám phá Dữ liệu

Lưu ý: hãy cố gắng nhờ một chuyên gia trong ngành để hiểu rõ hơn về dữ liệu khi thực hiện các bước này.

1) Tạo một bản sao dữ liệu dành cho việc khám phá (lấy mẫu với kích thước nhỏ hơn để có thể xử lý dễ dàng nếu cần).
2) Tạo một Jupyter notebook để ghi lại quá trình khám phá dữ liệu.
3) Nghiên cứu từng thuộc tính và đặc điểm của nó:
   - Tên
   - Dạng dữ liệu (dạng hạng mục, dạng số nguyên/số thực, dạng bị chặn/không bị chặn, dạng văn bản, dạng có cấu trúc, v.v.)
   - Phần trăm của những giá trị bị thiếu
   - Độ nhiễu và dạng nhiễu (ngẫu nhiên, ngoại lai, lỗi làm tròn số, v.v.)
   - Tính hữu ích cho tác vụ
   - Dạng phân phối của dữ liệu (Gauss, đều, logarit, v.v.)
4) Đối với các tác vụ học có giám sát, hãy xác định (các) thuộc tính mục tiêu.
5) Trực quan hóa dữ liệu.
6) Nghiên cứu mối tương quan giữa các thuộc tính.
7) Nghiên cứu cách bạn sẽ giải quyết bài toán theo cách thủ công.
8) Xác định các phép biến đổi tiềm năng mà bạn có thể áp dụng.
9) Xác định dữ liệu bổ sung hữu ích (quay lại bước Thu thập Dữ liệu).
10) Ghi lại những gì bạn đã học được.

IV.Chuẩn bị Dữ liệu

Lưu ý:

- Làm việc trên các bản sao của dữ liệu thay vì tập dữ liệu gốc (giữ nguyên tập dữ liệu gốc).
- Viết hàm cho tất cả các phép biến đổi dữ liệu, vì năm lý do sau:
  - Dễ dàng hơn trong việc chuẩn bị dữ liệu trong lần tiếp theo khi bạn có một tập dữ liệu mới
  - Có thể sử dụng các phép biến đổi này trong những dự án tương lai
  - Làm sạch và chuẩn bị tập kiểm tra
  - Làm sạch và chuẩn bị các mẫu dữ liệu mới khi giải pháp được triển khai
  - Dễ dàng để coi các lựa chọn trong việc chuẩn bị dữ liệu như các siêu tham số

1) Làm sạch dữ liệu
   - Sửa hoặc loại bỏ các điểm ngoại lai (không bắt buộc).
   - Điền vào các giá trị còn thiếu (ví dụ với 0, trung bình, trung vị, v.v.) hoặc loại bỏ hàng (hoặc cột) chứa các giá trị đó.

2) Lựa chọn đặc trưng (không bắt buộc):
   - Loại bỏ các thuộc tính không hữu ích cho tác vụ.
3) Thiết kế đặc trưng, nếu thích hợp:
   - Biến các đặc trưng liên tục thành rời rạc.
   - Phân tách các đặc trưng (ví dụ như hạng mục, thời gian, v.v..)
   - Thêm các phép biến đổi đặc trưng tiềm năng (ví dụ như log(x), $\sqrt{x}$, x^2 , v.v..)
   - Tổng hợp các đặc trưng thành đặc trưng mới tiềm năng.
4) Co giãn đặc trưng:
   - Chuẩn hóa hoặc chuẩn tắc hóa các đặc trưng.

V. Rút gọn Danh sách các Mô hình Tiềm năng

Lưu ý:

- Nếu tập dữ liệu lớn, bạn có thể lấy mẫu các tập huấn luyện nhỏ hơn để huấn luyện nhiều mô hình khác nhau trong một khoảng thời gian hợp lý (lưu ý rằng phương pháp này không hiệu quả với những mô hình phức tạp như các mạng nơ-ron lớn hoặc Rừng Ngẫu nhiên).
- Một lần nữa, hãy cố gắng tự động hóa các bước này nhiều nhất có thể.

1) Huấn luyện nhiều loại mô hình đơn giản khác nhau (ví dụ như tuyến tính, naive Bayes, SVM, Rừng Ngẫu nhiên, Mạng nơ-ron, v.v..) sử dụng các tham số tiêu chuẩn.
2) Đo lường và so sánh chất lượng của chúng.
   - Với mỗi mô hình, sử dụng kiểm định N-fold, tính trung bình và độ lệch chuẩn của chất lượng mô hình trên N fold.
3) Phân tích những biến quan trọng nhất của từng thuật toán.
4) Phân tích những loại lỗi mà các mô hình gặp phải.
   - Con người sẽ sử dụng dữ liệu nào để tránh những lỗi này?
5) Thực hiện nhanh một lượt lựa chọn và thiết kế đặc trưng.
6) Thực hiện nhanh một hoặc hai lượt cả năm bước ở trên.
7) Tạo danh sách rút gọn từ ba tới năm mô hình có tiềm năng nhất, ưu tiên các mô hình có các loại lỗi khác nhau.

VI. Tinh chỉnh Hệ thống

Lưu ý:

- Bạn nên sử dụng càng nhiều dữ liệu càng tốt tại bước này, đặc biệt là trong giai đoạn cuối của quá trình tinh chỉnh.
- Như thường lệ, cố gắng tự động hóa càng nhiều càng tốt.

1) Tinh chỉnh các siêu tham số sử dụng kiểm định chéo:
   - Coi các lựa chọn biến đổi dữ liệu như các siêu tham số, đặc biệt là khi bạn không chắc chắn về cách lựa chọn chúng (ví dụ nếu bạn phân vân trong việc thay thế giá trị thiếu bằng 0 hay bằng giá trị trung vị, hay chỉ đơn giản là bỏ luôn các hàng đó).
   - Trừ khi có rất ít các giá trị siêu tham số để thử nghiệm, hãy ưu tiên tìm kiếm ngẫu nhiên thay vì tìm kiếm dạng lưới. Nếu quá trình huấn luyện tốn nhiều thời gian, bạn có thể sử dụng phương pháp tối ưu hóa Bayes.
2) Thử các phương pháp Ensemble. Kết hợp các mô hình tốt nhất thường sẽ cho kết quả tốt hơn so với từng mô hình riêng biệt.
3) Một khi bạn đã tự tin với mô hình cuối cùng, hãy đo lường chất lượng trên tập kiểm tra để ước lượng lỗi tổng quát hóa.

VII. Trình bày Giải pháp

1) Ghi chép lại những gì bạn đã làm.
2) Thực hiện một bài thuyết trình hay.
   - Hãy đảm bảo rằng bạn sẽ nhấn mạnh bức tranh tổng thể trước.
3) Giải thích tại sao giải pháp của bạn đạt được mục tiêu kinh doanh.
4) Đừng quên trình bày những điểm thú vị mà bạn tìm được trong quá trình phát triển hệ thống.
   - Mô tả những thứ hoạt động và không hoạt động.
   - Liệt kê các giả định cũng như các hạn chế của hệ thống.
5) Đảm bảo những phát hiện chính của bạn được truyền đạt bằng trực quan đẹp mắt hoặc mệnh đề dễ nhớ (ví dụ như "thu nhập trung bình là nhân tố quan trọng nhất để dự đoán giá nhà ở").

VIII. Triển khai!

1) Chuẩn bị để triển khai giải pháp (đưa dữ liệu đầu vào vào hệ thống, viết unit test, v.v..).
2) Viết mã giám sát để kiểm tra định kỳ chất lượng của hệ thống trong quá trình hoạt động và kích hoạt cảnh bảo khi chất lượng đi xuống.
   - Cẩn thận với sự xuống cấp chậm của hệ thống: các mô hình có xu hướng "mục nát" khi dữ liệu thay đổi.
   - Việc đo lường chất lượng có thể yêu cầu nhân công con người (ví dụ như thông qua các dịch vụ crowdsourcing).
   - Đồng thời theo dõi chất lượng của đầu vào (ví dụ cảm biến bị trục trặc gửi đi các giá trị ngẫu nhiên, hoặc đầu ra của một nhóm khác không được cập

nhật theo thời gian thực). Điều này đặc biệt quan trọng với các hệ thống học trực tuyến.

3) Huấn luyện lại các mô hình theo định kỳ với dữ liệu mới (tự động hóa nhiều nhất có thể).

D. Ý TƯỞNG CHO DỰ ÁN:

Sau đây là một vài ý tưởng cho dự án máy học mà bạn có thể phát triển. Các bạn có thể thảo luận với tôi trong trường hợp chọn đề tài khác.

Project 1. **Dự đoán xếp hạng khách sạn (prediction of hotels' rating)** - Nghiên cứu xây dựng hệ thống có thể đưa ra dự đoán chính xác về xếp hạng cho một khách sạn khi mới ra mắt, sử dụng một số mô tả về khách sạn đó. Xếp hạng thuộc về {1*, 2*, 3*, 4*, 5*}.

- Input: mô tả, thông tin về khách sạn.
- Output: đánh giá rating cho khách sạn đó.
- Phương pháp có thể sử dụng: Random Forest,….
- Dataset: dữ liệu về khác sạn, mô tả chi tiết về các khác sạn. Dữ liệu có thể được thu thập từ https://www.agoda.com/.

Project 2. **Dự đoán xếp hạng ứng dụng (Prediction of apps' rating)** - Nghiên cứu xây dựng hệ thống có thể đưa ra dự đoán chính xác về xếp hạng trung bình cho một ứng dụng, sử dụng một số mô tả về ứng dụng.

- Input: mô tả, thông tin về ứng dụng (app).
- Output: xếp hạng trung bình từ người dùng cho một ứng dụng.
- Phương pháp có thể sử dụng: Ridge regression hoặc neural network,…
- Dataset: Danh sác các ứng dụng và mô tả của chúng dưới dạng văn bản, mỗi ứng dụng có xếp hạng được thu thập từ App Store.

Project 3. **Sở thích của người dùng về âm nhạc (Users' preference in music)** – Nghiên cứu xây dựng ứng dụng phân tích sở thích/quan tâm của người dùng trực tuyến về âm nhạc, theo nhân khẩu học/thời gian/giới tính,…

- Input: tập hợp các bài hát/MV và tập hợp người dùng cũng như sự tương tác của họ với bài hát/MV
- Output: sở thích âm nhạc, phát hiện mới, trực quan hoá về sở thích âm nhạc,…
- Phương pháp có thể sử dụng: phân cụm theo K-means, phân loại bằng Random forest,…
- Dataset: dữ liệu về các bài hát/MV cũng như tập hợp người dùng và sự tương tác của họ với bài hát/MV. Dữ liệu có thể được thu thập từ https://www.youtube.com/ .

Project 4. **Phân loại chất lượng rượu vang (Wine Quality Classification)** - Phát triển mô hình phân loại chất lượng rượu vang dựa trên các đặc trưng hóa học như độ pH, lượng đường, độ cồn.

- Input: Đặc trưng hóa học của rượu.
- Output: Phân loại chất lượng rượu (từ 1 đến 10).
- Phương pháp có thể sử dụng: Support Vector Machine (SVM), Neural Networks.
- Dataset: Bộ dữ liệu "Wine Quality" từ UCI Machine Learning Repository.

Project 5. **Hệ thống dự đoán điểm tín dụng (Credit Scoring System)** - Xây dựng hệ thống dự đoán mức độ tin cậy của khách hàng vay tín dụng dựa trên thông tin cá nhân và lịch sử tài chính.

- Input: Thông tin khách hàng (tuổi, thu nhập, lịch sử tín dụng, số dư tài khoản, v.v.).
- Output: Điểm tín dụng hoặc phân loại (tốt, trung bình, xấu).
- Phương pháp có thể sử dụng: Logistic Regression, Gradient Boosting.
- Dataset: Dữ liệu tín dụng từ Kaggle hoặc các nguồn công khai.

Project 6. **Dự đoán khả năng đậu kỳ thi (Prediction of Exam Success)** - Phân tích và dự đoán khả năng đậu kỳ thi của học sinh dựa trên điểm số, thời gian học tập, và các yếu tố khác.

- Input: Thông tin về quá trình học tập và kết quả trước đó.
- Output: Khả năng đậu kỳ thi (0 hoặc 1).
- Phương pháp có thể sử dụng: Decision Tree, Naive Bayes.
- Dataset: Bộ dữ liệu giáo dục từ các nguồn công khai.

Project 7. **Dự đoán lưu lượng giao thông (Traffic Flow Prediction)** - Phát triển hệ thống dự đoán lưu lượng giao thông tại một địa điểm cụ thể dựa trên dữ liệu thời gian thực như ngày, giờ, và các yếu tố môi trường.

- Input: Thông tin thời gian và đặc trưng môi trường (nhiệt độ, độ ẩm, ngày lễ).
- Output: Lưu lượng giao thông (cao, trung bình, thấp).
- Phương pháp có thể sử dụng: Time Series Analysis, LSTM.
- Dataset: Bộ dữ liệu giao thông từ các thành phố lớn (NYC, San Francisco, HCM,….).

Project 8. **Xây dựng hệ thống nhận dạng khuôn mặt và dự đoán tuổi của con người** - Nghiên cứu và xây dựng hệ thống dự đoán độ tuổi trên khuôn mặt người. Hệ thống sẽ nhận dạng toàn bộ những khuôn mặt có trong ảnh, real-time video và dự đoán xem khuôn mặt ấy đang trong 5 mức độ tuổi: 1-14, 14-25, 25-40, 40-60, trên 60 tuổi. Yêu cầu bài toán:

- Phát hiện đúng khuôn mặt có trong ảnh, video.

- Mô hình đạt được tỉ lệ chính xác cao, tối thiểu sự sai số về độ tuổi giúp người dùng tin tưởng để sử dụng.
- Bảo đảm sự mượt mà khi chạy real-time với webcam.

Có thể chia nhỏ bài toán thành 2 bài toán con cần quan tâm và giải quyết:

- Bài toán 1: Phát hiện toạ độ khuôn mặt trong ảnh, video.
- Bài toán 2: Sau khi đã xác định được khuôn mặt, dự đoán độ tuổi trên khuôn mặt đó.

Dataset: Sinh viên có thể sử dụng tập dữ liệu ở link sau:

https://www.kaggle.com/datasets/mariafrenti/age-prediction

Project 9. **Lọc thư rác** - Nghiên cứu xây dựng ứng dụng lọc thư rác, xác định (phân loại) những thư điện tử là thư rác (spam e-mails)

- Input: Biểu diễn nội dung của một e-mail (vd: một vectơ các từ khóa – có/không có trọng số)
- Output: Thư rác ("spam") hoặc thư hợp lệ ("normal")
- Phương pháp có thể sử dụng: Phân loại Naïve Bayes
- Tập dữ liệu: Một tập các ví dụ; mỗi ví dụ bao gồm biểu diễn nội dung của một e-mail và nhãn lớp ("spam" hoặc "normal")

Project 10. **Phân loại các trang Web** - Nghiên cứu xây dựng ứng dụng phân loại các trang web. Với một tập các trang Web, hệ thống cần phải gán (phân loại) mỗi trang Web vào một trong số các thể loại (vd: "Kinh doanh", "Thể thao", "Công nghệ", ...)

- Input: Biểu diễn nội dung của một trang Web (vd: một vectơ các tần xuất xuất hiện của các từ khóa)
- Output: Thể loại phù hợp của trang Web đó
- Phương pháp có thể sử dụng: Phân loại Naïve Bayes, hoặc Mạng nơ-ron nhân tạo
- Tập dữ liệu: Một tập các ví dụ; mỗi ví dụ bao gồm biểu diễn của một trang Web và nhãn lớp (thể loại)

Project 11. **Trợ lý ảo hỗ trợ học tập kiến thức môn học CTDL&GT** - Phát triển trợ lý ảo hỗ trợ học tập kiến thức môn học Cấu trúc dữ liệu & Giải thuật. Mục tiêu của dự án:

- Nghiên cứu các kỹ thuật xây dựng trợ lý ảo (chatbot), bao gồm: các kỹ thuật xử lý ngôn ngữ tự nhiên, xử lý voice, phân loại ý định, trích xuất thông tin, quản lý hội thoại...
- Nghiên cứu, sử dụng các thư viện hỗ trợ xây dựng chatbot như LangChain, Ollama và các kỹ thuật như RAG, Prompt, Text embedding, để xây dựng và huấn luyện mô hình cho môn học CTDL&GT.

- Xây dựng ứng dụng web có tích hợp chatbot (text, voice) hỗ trợ học tập môn CTDL&GT

Project 12. **Trợ lý ảo phân tích cảm xúc trong văn bản** – Mục tiêu của dự án:

- Nghiên cứu các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) để phân tích cảm xúc trong văn bản, bao gồm các phương pháp học máy truyền thống và học sâu.
- Tích hợp các thư viện như NLTK, spaCy, Hugging Face để thực hiện xử lý dữ liệu, tạo mô hình và đánh giá hiệu quả.
- Xây dựng ứng dụng web hỗ trợ người dùng nhập văn bản (hoặc upload file) để phân tích cảm xúc tự động (tích cực, tiêu cực, trung tính).

Project 13. **Trợ lý ảo hỗ trợ tuyển sinh và các vấn đề học vụ cho sinh viên** – Nghiên cứu, phát triển trợ lý ảo và ứng dụng quản lý trợ lý ảo hỗ trợ tư vấn tuyển sinh các ngành của Khoa Công nghệ - Kỹ thuật và học vụ cho sinh viên Trường Đại học Quốc tế Hồng Bàng sử dụng https://dify.ai/

Project 14. **Hệ thống gợi ý xem phim** - Nghiên cứu phát triển hệ thống gợi ý xem phim online. Mục tiêu của dự án:

- Nghiên cứu về máy học và về các kỹ thuật xây dựng hệ thống khuyến nghị như collaborating filtering, content-based filtering, hybrid-recommender systems, các giải pháp khuyến nghị phim và các thư viện Python, scikit-learn, numpy, pandas, keras..
- Xây dựng dụng web khuyến nghị xem phim.
- Nghiên cứu, sử dụng Angular và các thư viện biểu diễn chart để trực quan hoá các kết quả thống kê của ứng dụng.

Project 15. **Hệ thống khuyến nghị sách học thuật cá nhân hóa** - Mục tiêu của dự án:

- Nghiên cứu thuật toán hệ thống khuyến nghị như Collaborative Filtering, Content-Based Filtering và Hybrid Recommendation.
- Áp dụng thư viện Scikit-Learn, Surprise, hoặc TensorFlow để xây dựng hệ thống khuyến nghị.
- Tích hợp hệ thống lên web hoặc app, cung cấp gợi ý sách học thuật phù hợp với sở thích và lịch sử học tập của người dùng.

Project 16. **Ứng dụng tóm tắt truyện, bài đọc** - Xây dựng ứng dụng tóm tắt truyện, bài đọc dành cho học sinh tiểu học. Mục tiêu của dự án:

- Nghiên cứu về xử lý ngôn ngữ tự nhiên (natural language processing), mô hình ngôn ngữ lớn (large language model - LLM), các kỹ thuật tóm tắt văn bản text (Text Summarization)
- Nghiên cứu, sử dụng các thư viện hỗ trợ tóm tắt văn bản như Transformers, T5, LangChain,...

- Xây dựng ứng dụng web để tóm tắt truyện đọc, bài đọc kèm hình ảnh minh hoạ dành cho học sinh tiểu học.
- Nghiên cứu, sử dụng Angular và các thư viện biểu diễn chart để trực quan hoá các kết quả thống kê của ứng dụng.

Project 17. **Chuyển đổi câu truy vấn tự nhiên thành SQL cơ bản (Basic Text-to-SQL Conversion)** - Xây dựng hệ thống chuyển đổi các câu truy vấn tự nhiên (tiếng Anh hoặc tiếng Việt) thành câu lệnh SQL cơ bản.

- Input: Câu truy vấn tự nhiên (VD: "Tìm tất cả sinh viên có điểm trên 8").
- Output: Câu lệnh SQL tương ứng (VD: SELECT * FROM students WHERE grade > 8;).
- Phương pháp có thể sử dụng: Seq2Seq, Transformer, hoặc các mô hình pre-trained như T5, GPT.
- Dataset: Bộ dữ liệu Text-to-SQL công khai như Spider, WikiSQL.

Project 18. **Hệ thống trợ lý truy vấn cơ sở dữ liệu (Database Query Assistant)** - Xây dựng một trợ lý ảo giúp người dùng truy vấn cơ sở dữ liệu bằng ngôn ngữ tự nhiên và trả về kết quả.

- Input: Câu truy vấn tự nhiên.
- Output: Kết quả từ cơ sở dữ liệu (kèm theo câu SQL được tạo).
- Phương pháp có thể sử dụng: Fine-tune các mô hình NLP như BERT hoặc GPT để sinh SQL từ text.
- Dataset: Spider, Kaggle datasets hoặc tự xây dựng tập dữ liệu từ các bảng cơ sở dữ liệu mẫu.

Project 19. **Phân tích cú pháp câu truy vấn để tối ưu hóa SQL (Query Parsing and SQL Optimization)** - Phân tích các câu truy vấn tự nhiên và tối ưu hóa các câu lệnh SQL sinh ra để giảm thời gian thực thi trên cơ sở dữ liệu lớn.

- Input: Câu truy vấn tự nhiên.
- Output: Câu SQL tối ưu hóa.
- Phương pháp có thể sử dụng: BiLSTM, Tree-LSTM để phân tích cú pháp kết hợp với thuật toán tối ưu hóa.
- Dataset: Bộ dữ liệu chứa các truy vấn SQL thực tế và thời gian thực thi.

Project 20. **Hệ thống Text-to-SQL đa ngôn ngữ (Multilingual Text-to-SQL System)** - Xây dựng hệ thống Text-to-SQL hỗ trợ nhiều ngôn ngữ (tiếng Anh, tiếng Việt, tiếng Nhật, v.v.) để phục vụ nhu cầu truy vấn toàn cầu.

- Input: Câu truy vấn tự nhiên bằng nhiều ngôn ngữ.
- Output: Câu SQL tương ứng.
- Phương pháp có thể sử dụng: Fine-tune mô hình mBERT hoặc XLM-R cho bài toán sinh SQL.

- Dataset: Dữ liệu Text-to-SQL quốc tế như Spider, kết hợp với dữ liệu tự gán nhãn.

Project 21. **Hệ thống kiểm tra lỗi và đề xuất sửa câu truy vấn (SQL Error Detection and Correction)** - Phát triển hệ thống kiểm tra câu SQL được sinh ra từ truy vấn tự nhiên và đề xuất sửa lỗi nếu phát hiện lỗi cú pháp hoặc logic.

- Input: Câu SQL được sinh ra từ truy vấn.
- Output: Câu SQL đúng hoặc gợi ý sửa lỗi.
- Phương pháp có thể sử dụng: Transformer-based models kết hợp với Grammar Rules.
- Dataset: Bộ dữ liệu các câu SQL lỗi và câu sửa lỗi (tự xây dựng hoặc thu thập từ log của các hệ quản trị cơ sở dữ liệu).

Project 22. **Optical Music Recognition** - Nghiên cứu về Học Sâu (Deep Learning) và các kiến trúc liên quan để xây dựng mô hình OMR (Optical Music Recognition) nhằm tạo ứng dụng tự động lật trang cho nhạc sĩ.

Project 23. **Phân loại hình ảnh hỗ trợ chẩn đoán y khoa** - Mục tiêu:

- Nghiên cứu các mạng CNN (Convolutional Neural Network) và ứng dụng thư viện TensorFlow hoặc PyTorch để phân loại hình ảnh y khoa.
- Huấn luyện mô hình trên các bộ dữ liệu hình ảnh công khai (như X-quang, MRI).
- Xây dựng ứng dụng cho phép tải hình ảnh lên và nhận kết quả phân loại kèm theo tỷ lệ chính xác.

Project 24. **Ứng dụng phát hiện gian lận giao dịch tài chính** - Mục tiêu:

- Nghiên cứu các kỹ thuật Machine Learning như Decision Tree, Random Forest, và XGBoost để phát hiện gian lận.
- Sử dụng các bộ dữ liệu giao dịch tài chính (giả lập hoặc công khai) để huấn luyện và đánh giá mô hình.
- Xây dựng giao diện hiển thị phân tích và cảnh báo khi phát hiện các giao dịch đáng nghi ngờ.

Project 25. **Hệ thống nhận diện biển số xe tự động (Automatic License Plate Recognition System)** - Xây dựng hệ thống nhận diện biển số xe từ hình ảnh hoặc video trong thời gian thực.

- Input: Ảnh hoặc video chứa xe và biển số.
- Output: Văn bản chứa số và chữ trên biển số xe.
- Phương pháp có thể sử dụng: YOLO (You Only Look Once), Tesseract OCR, OpenCV.
- Dataset: Bộ dữ liệu biển số xe Việt Nam hoặc quốc tế (có thể tự thu thập hoặc tải từ Kaggle)

Project 26. **Phân loại biển số xe theo quốc gia (License Plate Country Classification)** - Phân loại biển số xe theo quốc gia dựa trên đặc điểm như phông chữ, màu sắc, kích thước.

- Input: Ảnh biển số xe.
- Output: Quốc gia sở hữu biển số (VD: Việt Nam, Mỹ, Nhật Bản).
- Phương pháp có thể sử dụng: Convolutional Neural Networks (CNNs).
- Dataset: Bộ dữ liệu biển số xe quốc tế từ các nguồn công khai.

Project 27. **Phát hiện biển số xe không hợp lệ (Invalid License Plate Detection)** - Xây dựng hệ thống kiểm tra và phát hiện biển số xe giả hoặc không hợp lệ dựa trên các quy chuẩn định dạng.

- Input: Ảnh biển số xe.
- Output: Kết quả phân loại (Hợp lệ/Không hợp lệ).
- Phương pháp có thể sử dụng: SVM, Random Forest hoặc mạng RNN để phân tích chuỗi ký tự.
- Dataset: Bộ dữ liệu về biển số xe hợp lệ và không hợp lệ.

Project 28. **Hệ thống kiểm soát bãi đỗ xe thông minh (Smart Parking Lot Management)** - Tích hợp nhận diện biển số xe để ghi nhận và quản lý xe ra/vào bãi đỗ tự động.

- Input: Video hoặc hình ảnh xe ra/vào.
- Output: Văn bản chứa biển số xe và thời gian ra/vào.
- Phương pháp có thể sử dụng: Deep Learning kết hợp OpenCV và xử lý thời gian thực.
- Dataset: Bộ dữ liệu biển số xe và video quay bãi đỗ (có thể tự thu thập).

Project 29. **Trợ lý ảo hỗ trợ tuyển sinh và các vấn đề học vụ cho sinh viên** – Nghiên cứu, phát triển trợ lý ảo và ứng dụng quản lý trợ lý ảo hỗ trợ tư vấn tuyển sinh các ngành của Khoa Công nghệ - Kỹ thuật và học vụ cho sinh viên Trường Đại học Quốc tế Hồng Bàng sử dụng ChatGPT's API hoặc DeepSeek's API.

Project 30. **Xây dựng ứng dụng với thư viện, dự án nguồn mở về Machine Learning** – Sinh viên nghiên cứu một trong các dự án ở danh sách sau https://github.com/ml-tooling/best-of-ml-python và xây dựng ứng dụng.

---------- THE END ----------