

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**



# **BÁO CÁO ĐỒ ÁN THỰC HÀNH**

**|Đề tài|**

**THỰC HIỆN QUY TRÌNH KHOA HỌC DỮ LIỆU  
VÀ MÔ HÌNH HÓA DỮ LIỆU**

**|Trợ giảng phụ trách đồ án|**

**Trần Đại Chí  
Nguyễn Bảo Long  
Lê Nhật Nam  
Nguyễn Thái Vũ**

**Môn học: Nhập môn Khoa học dữ liệu**

**Thành phố Hồ Chí Minh - 2022**

---

## THÔNG TIN THÀNH VIÊN

---

MSSV	Họ tên
20120357	Nguyễn Đức Minh Quân
20120402	Nguyễn Hoàng Việt
20120120	Nguyễn Việt Khoa
20120565	Nguyễn Tấn Sơn

---

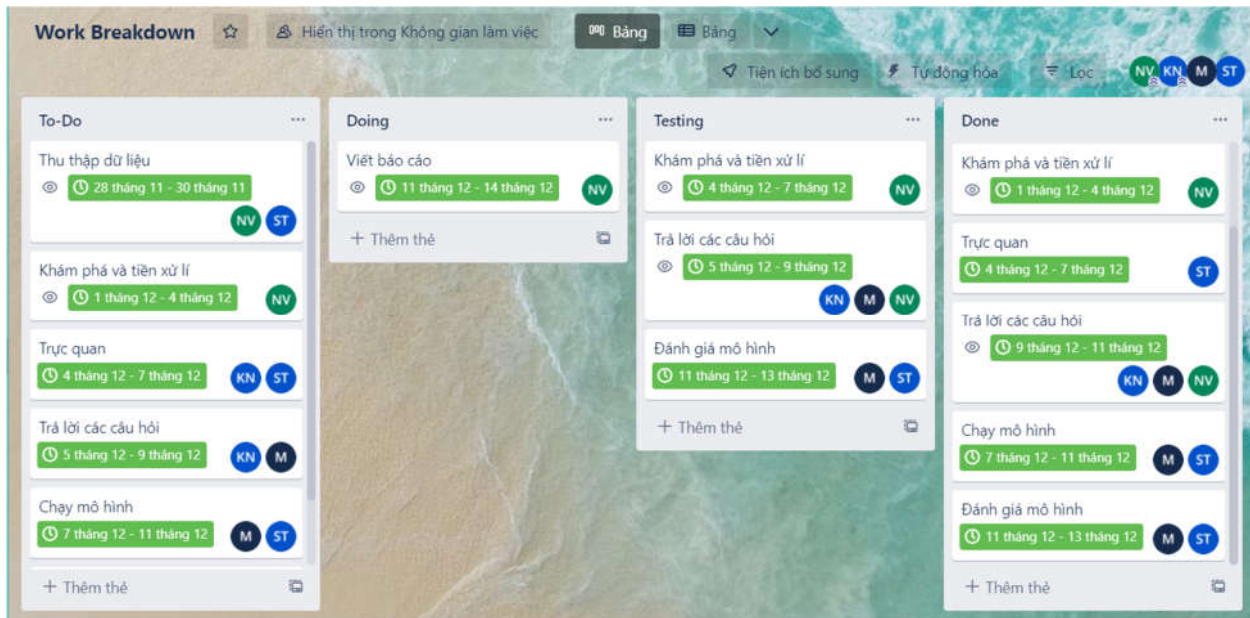
## MỤC LỤC

---

### Nội dung

THÔNG TIN THÀNH VIÊN.....	2
MỤC LỤC .....	2
PHÂN CHIA CÔNG VIỆC NHÓM.....	3
BÁO CÁO TỪNG THÀNH VIÊN .....	5

## PHÂN CHIA CÔNG VIỆC NHÓM



Việc xác định dữ liệu để thu thập và khám phá do cả nhóm bỏ phiếu và chọn ra. Công việc cụ thể:

- Nguyễn Hoàng Việt: thu thập dữ liệu, khám phá và tiền xử lý; kiểm tra tính đúng đắn cho các câu trả lời của phần đặt câu hỏi; phân chia công việc cho nhóm trên Trello và viết báo cáo.
- Nguyễn Tấn Sơn: thu thập dữ liệu, trực quan hóa dữ liệu và chạy mô hình.
- Nguyễn Việt Khoa: tham gia trực quan hóa dữ liệu, đặt ra các câu hỏi và trả lời các câu hỏi đó.
- Nguyễn Đức Minh Quân: đặt ra 1 câu hỏi, khắc phục những vấn đề của mô hình và đánh giá mô hình.

## LỊCH SỬ LÀM VIỆC TRÊN GITHUB

Link repository Github: [https://github.com/kun-zero162/NMKHDL-20\\_21-Final-Project](https://github.com/kun-zero162/NMKHDL-20_21-Final-Project)

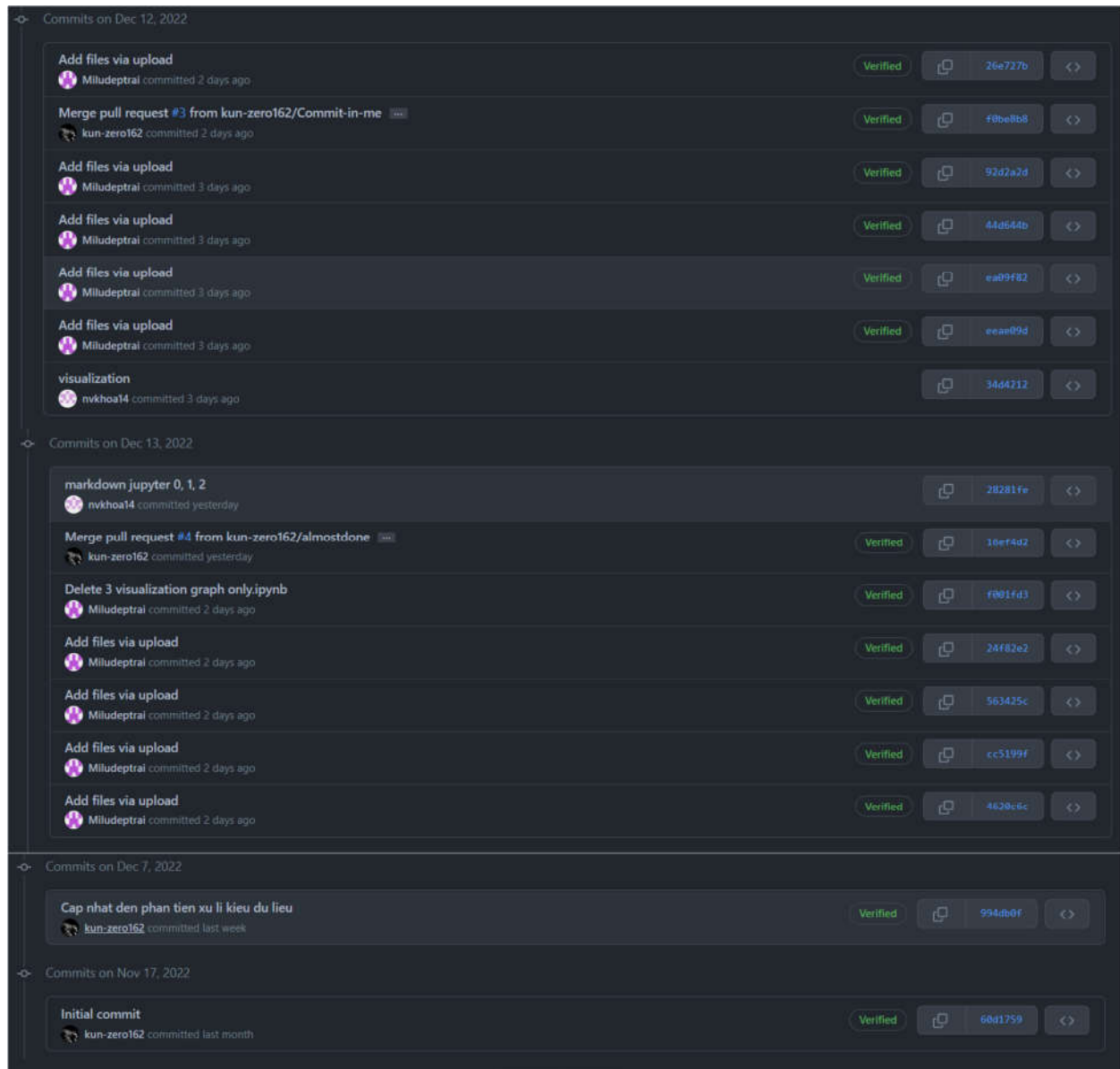
## Lịch sử commit:

The screenshot displays the commit history for the 'main' branch on GitHub. The interface is dark-themed. At the top, there's a dropdown menu for the branch, currently set to 'main'. Below this, a section titled 'Commits on Dec 14, 2022' lists the following commits:

- final of final** by kun-zero162, committed 3 minutes ago. Hash: d78ca82.
- Merge remote-tracking branch 'origin/visual\_model'** by kun-zero162, committed 4 minutes ago. Hash: 5180f8b.
- rename model** by kun-zero162, committed 5 minutes ago. Hash: 8471645.
- Merge branch 'visual\_model'** by kun-zero162, committed 10 minutes ago. Hash: f9ead0c.
- fix model nb** by kun-zero162, committed 11 minutes ago. Hash: 8fc9348.
- xu ly conflict 'origin/visual\_model'** by kun-zero162, committed 18 minutes ago. Hash: 8406ee7.
- Merge branch 'main' of https://github.com/kun-zero162/NMKHDL-20\_21-Fi...** by kun-zero162, committed 28 minutes ago. Hash: fb62d68.
- kiem tra tat ca + upload bao cao** by kun-zero162, committed 1 hour ago. Hash: 75b6c01.
- maybe\_final** by minhquanlk2002, committed 5 hours ago. Hash: 8aff6e6.
- Maybe\_final\_visualization** by minhquanlk2002, committed 5 hours ago. Hash: 70de7eb.

Below this list, another section titled 'Commits on Dec 14, 2022' shows more commits:

- Merge branch 'main' of https://github.com/kun-zero162/NMKHDL-20\_21-Fi...** by kun-zero162, committed 1 minute ago. Hash: fb62d68.
- kiem tra tat ca + upload bao cao** by kun-zero162, committed 3 minutes ago. Hash: 75b6c01.
- Merge pull request #8 from kun-zero162/visual\_model** by kun-zero162, committed 5 hours ago. Hash: 44a0513. Status: Verified.
- final\_model** by minhquanlk2002, committed 5 hours ago. Hash: fc11730.
- Merge pull request #7 from kun-zero162/visual\_model** by kun-zero162, committed 5 hours ago. Hash: 3a8a7cc. Status: Verified.
- final\_model** by minhquanlk2002, committed 5 hours ago. Hash: 6584161.
- Merge branch 'markdown' vao main** by kun-zero162, committed 7 hours ago. Hash: 6d1c136.
- Merge pull request #5 from kun-zero162/visual\_model** by minhquanlk, committed yesterday. Hash: 8e94e9c. Status: Verified.
- modify** by minhquanlk2002, committed yesterday. Hash: 099ff3d.



## BÁO CÁO TỪNG THÀNH VIÊN

Họ và tên: Nguyễn Hoàng Việt

MSSV: 20120402

### Những khó khăn gặp phải khi làm đồ án:

- Gặp khó khăn trong việc tìm kiếm ý tưởng cho các câu hỏi.
- Có ý tưởng về câu hỏi nhưng không thể tìm được câu trả lời "hợp lý" từ dữ liệu.

### Những điều học được thông qua đồ án lần này:

- Có thêm kinh nghiệm trong việc khám phá và tiền xử lý dữ liệu.
- Biết thêm nhiều thư viện cũng như công cụ bổ ích trong quá trình trực quan và mô hình hóa dữ liệu.
- Phân tích và đặt ra các câu hỏi có ý nghĩa thực tiễn từ dữ liệu đã khám phá
- Nâng cao kỹ năng làm việc nhóm, quản lý mã nguồn qua github
- Tìm ra lỗi sai của bản thân khi phân tích dữ liệu và khắc phục lỗi sai đó.

**Những việc nhóm sẽ làm khi có thêm thời gian:**

- Tìm và thu thập dữ liệu đa dạng hơn, có thể thử nghiệm thu thập dữ liệu bằng cách parse HTML hoặc get API.
- Mô hình hóa dữ liệu với nhiều thuật toán khác.
- Đặt thêm nhiều câu hỏi về dữ liệu hơn để trả lời.

-----//-----

Họ và tên: Nguyễn Việt Khoa

MSSV: 20120120

**Những khó khăn gặp phải khi làm đồ án:**

- Có ý tưởng về câu hỏi nhưng chưa tiếp cận thêm về dữ liệu.
- Khó khăn trong xây dựng và giải thích biểu đồ.

**Những điều học được thông qua đồ án lần này:**

- Tìm hiểu thêm được nhiều biểu đồ hay và hiệu quả trong trực quan hóa.
- Đào sâu phân tích thêm dữ liệu để trả lời câu hỏi.
- Nâng cao kỹ năng làm việc nhóm qua trello, quản lý mã nguồn thông qua github.

**Những việc sẽ làm khi có thêm thời gian:**

- Mô hình hóa dữ liệu với nhiều thuật toán khác.
- Đặt thêm nhiều câu hỏi về dữ liệu hơn để trả lời.

-----//-----

Họ và tên: Nguyễn Tấn Sơn

MSSV: 20120565

**Những khó khăn gặp phải khi làm đồ án:**

- Hiệu chỉnh selector và xpath cho đến khi lấy được data, fix các trường hợp đặc biệt.
- Khó khăn trong việc tìm cách thay giá trị thiếu/nhiều.

- Thư viện plotly tương đối là mới mẻ đối với nhóm và tài liệu về thư viện này cũng khó có thể tiếp cận.
- Thiếu cột giá trị binary dùng để phân lớp, phải dùng html request trên 1 trang khác.
- Các siêu tham số như Binary cross entropy loss khiến việc lập công thức khó khăn.
- Đánh giá RMSE cho cả bảng dữ liệu khá khó khăn khi các giá trị ở mỗi cột có đơn vị đo khác nhau.

**Những điều học được thông qua đồ án lần này:**

- Biết được cách copy selector và xpath trong browser.
- Cách dùng thư viện plotly.
- Học được nhiều kiến thức về việc lập model, thư viện sklearn, đánh giá model.

**Những việc sẽ làm khi có thêm thời gian:**

- Mô hình hóa dữ liệu với nhiều thuật toán khác.
- Đặt thêm nhiều câu hỏi về dữ liệu hơn để trả lời.

-----//-----

Họ và tên: Nguyễn Đức Minh Quân

MSSV: 20120357