BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CẦN THƠ
KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

BÁO CÁO
KHAI KHOÁNG DỮ LIỆU

Đề tài

DỰ ĐOÁN TÌNH TRẠNG LY HÔN TẬP DỮ LIỆU DIVORCE

Sinh viên thực hiện :
Trần Sĩ Đạt – B1611134
Lê Hoàng Khương – B1611136

Cần Thơ, 11/2020

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CẦN THƠ
KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

BÁO CÁO
KHAI KHOÁNG DỮ LIỆU

Đề tài

DỰ ĐOÁN TÌNH TRẠNG LY HÔN TẬP DỮ LIỆU DIVORCE

Giáo viên hướng dẫn:
TS.Lưu Tiến Đạo

Sinh viên thực hiện:
Trần Sĩ Đạt – B1611134
Lê Hoàng Khương – B1611136

Cần Thơ, 11/2020

NHẬN XÉT CỦA GIẢNG VIÊN

LỜI CẢM ƠN

Để hoàn thành được đề tài này, em đã nhận được nhiều sự giúp đỡ, hỗ trợ của nhiều Thầy Cô và các bạn. Em xin chân thành cảm ơn tới các Thầy Cô và các bạn đã tạo điều kiện, hỗ trợ em trong quá trình học tập và nghiên cứu đề tài.

Lời đầu tiên, em gửi lời cảm ơn sâu sắc nhất tới thầy TS. Lưu Tiến Đạo đã giảng dạy những kiến thức quan trọng, tạo điều kiện tốt nhất và hỗ trợ, giúp đỡ em trong quá trình thực hiện đề tài

Em cũng xin cảm ơn ban lãnh đạo và các thầy cô trong trường Đại Học Cần Thơ đã tạo môi trường học tập tốt nhất, sự quan tâm, dạy dỗ, chỉ bảo tận tình chu đáo của thầy cô. Đặc biệt em xin gửi lời cảm ơn tới các thầy cô khoa Công Nghệ Thông Tin và Truyền Thông lời chào trân trọng, lời chúc sức khỏe và lời cảm ơn sâu sắc

Em cũng gửi lời cảm ơn tới gia đình, người thân và các bạn đã động viên, hỗ trợ giúp đỡ em trong quá trình thực hiện đề tài này.

Với điều kiện thời gian cũng như kinh nghiệm còn hạn chế, em không thể tránh được những thiếu sót. Em rất mong nhận được sự chỉ bảo, đóng góp ý kiến của các thầy cô để em có điều kiện bổ sung, nâng cao ý thức của mình, phục vụ tốt hơn công tác thực tế sau này.

Cần Thơ, ngày tháng 11 năm 2020

Người viết

MỤC LỤC

DANH MỤC HÌNH

TÓM TẮT

Hiện nay, công nghệ thông tin đã được ứng dụng rộng rãi trong các lĩnh vực như y học, quân sự, giáo dục, kinh doanh, ... nhờ có tốc độ xử lý và khả năng lưu trữ vượt trội của máy vi tính.

Xã hội ngày càng phát triển sẽ đi kèm với lượng lớn dữ liệu cần lưu trữ. Nhưng với các dữ liệu thô sau khi thu thập sẽ không thể sử dụng được vì có thể bị thiếu dữ kiện, sai số, nhiễu... Cần phải lọc, xử lý các dữ liệu này để nhận được thông tin hữu ít.

Khai phá dữ liệu bao gồm việc tìm tòi và phân tích các khối dữ liệu lớn để chắt lọc ra được những mẫu hình và xu hướng có ý nghĩa. Nó được sử dụng trong nhiều mục đích khác nhau như tiếp thị theo cơ sở dữ liệu, quản trị rủi ro tín dụng, phòng chống gian lận, lọc mail rác, hoặc đơn giản là để tìm hiểu tâm lí và ý kiến của người dùng .

Trong đề tài này, chúng em thực hiện dự đoán tình trạng ly hôn trên tập dữ liệu divorce, sử dụng giải thuật DecisionTree để xây dựng mô hình. Xây dựng giao diện web dự đoán tình trạng ly hôn từ các thuộc tính và chọn vào tập tin csv để dự đoán trên tập tin.

PHẦN GIỚI THIỆU

1. Đặt vấn đề

Trong đời sống hôn nhân ngày nay, tình trạng đỗ vỡ hạnh phúc gia đình dẫn đến ly hôn ngày càng nhiều. Với các lý do rất đơn giản mà nhiều gia đình đã không để tâm đến cũng dẫn đến mâu thuẫn gia đình và cuối cùng là ly hôn. Sau khi ly hôn dẫn đến nhiều hệ luỵ ảnh hưởng xấu đến thế hệ trẻ sau này, ảnh hưởng đến xã hội và đất nước.

Vì các lý do trên, chúng em đã chọn đề tài "Dự đoán tình trạng ly hôn trên tập dữ liệu divorce" để giúp các gia đình hiểu các lý do và mức độ ảnh hưởng đến tình trạng hôn nhân và giảm thiểu được tình trạng ly hôn.

2. Lịch sử giải quyết vấn đề

Hiện nay, đã có nhiều tổ chức và cá nhân xây dựng mô hình trên tập dữ liệu divorce bằng ngôn ngữ Python và các giải thuật khác nhau như RandomForest, GaussianDB, Logistic Regression ....

3. Mục tiêu đề tài

- Mục tiêu của đề tài là xây dựng mô hình dự đoán tình trạng ly hôn trên tập dữ liệu divorce
- Thiết kế giao diện trang web dự đoán tình trạng ly hôn dựa trên các giá trị thuộc tính, trong tập dữ liệu này là mức độ các câu hỏi sẽ dẫn đến ly hôn (Tôi thích đi du lịch với vợ tôi ? Tôi biết món ăn yêu thích của vợ / chồng tôi ? ....)
- Dự đoán trên nhiều trường hợp bằng cách chọn tập tin csv
- Triển khai trang web lên Heroku Cloud.

4. Đối tượng và phạm vi nghiên cứu

- Đối tượng nghiên cứu : tập dữ liệu divorce, các giải thuật để xây dựng mô hình, ngôn ngữ Python, framework Flask, Heroku Cloud.

- Phạm vi nghiên cứu : cách thức hoạt động của Flask, cách thức triển khai trang web lên Heroku, cách xây dựng và đánh giá mô hình, cú pháp lập trình của Python, cách xây dựng và thiết kế trang web.

5. Phương pháp nghiên cứu

- Lý thuyết : tìm kiếm, tham khảo giáo trình Khai khoáng dữ liệu và đọc các tài liệu về khai khoáng dữ liệu trên internet. Tìm hiểu tập dữ liệu divorce (thuộc tính, dòng dữ liệu, nhãn, …), tìm hiểu các phương pháp xây dựng và đánh giá mô hình. Tìm hiểu các phương pháp xây dựng trang web bằng Flask và triển khai lên Heroku.

- Thực hành : từ những kiến thức đã học và tìm hiểu được vận dụng vào tạo các file .py để xây dựng và đánh giá mô hình, thiết kế giao diện của trang web, xây dựng trang web bằng framework Flask, triển khai trang web lên Heroku.

6. Kết quả đạt được

- ❖ Xây dựng được mô hình dự đoán ly hôn trên tập dữ liệu divorce
- ❖ Xây dựng được trang web với các chức năng :
  - ✓ Dự đoán tình trạng ly hôn thông qua các câu hỏi và các mức độ
  - ✓ Chọn tập tin csv và trả về kết quả dự đoán bên trong tập tin
  - ✓ Triển khai trang web lên Heroku

7. Bố cục đề tài

Phần giới thiệu

Giới thiệu tổng quát về đề tài.

Phần nội dung

Chương 1 : Mô tả bài toán.

Chương 2 : Thiết kế, cài đặt giải thuật.

Chương 3 : Kết quả thực nghiệm.

Phần kết luận

Trình bày kết quả đạt được và hướng phát triển hệ thống.

PHẦN NỘI DUNG

CHƯƠNG 1

MÔ TẢ BÀI TOÁN

1. Mô tả chi tiết bài toán

- Giới thiệu tập dữ liệu:
    - Tập dữ liệu mô tả về các câu hỏi và các mức độ sẽ dẫn đến ly hôn hay không.
    - Trong đó tập dữ liệu có 170 dòng dữ liệu với 54 thuộc tính và các giá trị từ 1 – 4 tương ứng với các mức độ. Cột nhãn với giá trị 1 (ly hôn) và 0 (không ly hôn)
- Tiền xử lý dữ liệu :
    - Kiểm tra dữ liệu null trong tập dữ liệu
- Xây dựng mô hình :
    - Cài đặt các giải thuật : RandomForest, DecisionTree, GaussianNB
    - Chia tập dữ liệu bằng nghi thức K-Fold với k=170 (leave-one-out)
    - Độ chính xác tổng thể của các giải thuật là
        - RandomForest : 97.64%
        - DecisionTree : 98.82%
        - GaussianNB : 97.05%
    - Chọn giải thuật DecisionTree để xây dựng mô hình

- Xây dựng trang web :
    - Sử dụng framework Flask để xây dựng phần xử lý và thiết kế giao diện bằng HTML, CSS, JS và sử dụng thêm Bootstrap để giao diện đẹp hơn.
    - Gọi mô hình vừa xây dựng được để dự đoán tình trạng ly hôn khi người dùng chọn và gửi tới server giá trị các thuộc tính
    - Viết và cài đặt các hàm xử lý để người dùng có thể chọn và dự đoán trên tập tin csv
- Triển khai hệ thống lên Heroku:
    - Cài đặt heroku và các gói cài đặt cần thiết lên máy
    - Viết các file cần thiết để có thể triển khai trang web

2. Vấn đề và giải pháp liên quan đến bài toán

Hệ thống được viết bằng ngôn ngữ Python, xây dựng và thiết kế trang web bằng Flask, Html, css, js, jquery, bootstrap. Xây dựng mô hình bằng giải thuật RandomForest và đánh giá mô hình bằng phương pháp K-Fold. Triển khai trang web lên Heroku

2.1. Ngôn ngữ :

2.1.1 HTML :

HTML là ngôn ngữ đánh dấu chuẩn để tạo trang web. HTML là viết tắt của Hyper Text Markup Language (ngôn ngữ đánh dấu siêu văn bản). HTML mô tả cấu trúc của trang web bằng các markup. Các phần tử trong HTML là các khối của trang web HTML. Các phần tử trong HTML được đại diện bằng những thẻ đánh dấu (tag). Thẻ đánh dấu HTML chứa các nội dung như ‘paragraph’, ‘heading’, ‘table’... Trình duyệt không hiển thị thẻ HTML nhưng dùng chúng để hiển thị nội dung của trang.

2.1.2 CSS :

CSS là viết tắt của Cascading Style Sheets. CSS mô tả cách các phần tử HTML hiển thị trên màn hình và các phương tiện khác. CSS rất hữu ích và tiện lợi. Nó có thể kiểm soát tất cả các trang trên một website.Các stylesheet ngoài được lưu trữ dưới dạng các tập tin .CSS.

2.1.3 JAVASCRIPT :

JavaScript là một ngôn ngữ lập trình của HTML và WEB. Nó là nhẹ và được sử dụng phổ biến nhất như là một phần của các trang web, mà sự thi hành của chúng cho phép Client-Side script tương tác với người sử dụng và tạo các trang web động. Nó là một ngôn ngữ chương trình thông dịch với các khả năng hướng đối tượng

2.1.4 PYTHON :

Python là một ngôn ngữ lập trình bậc cao cho các mục đích lập trình đa năng, do Guido van Rossum tạo ra và lần đầu ra mắt vào năm 1991. Python được thiết kế với ưu điểm mạnh là dễ đọc, dễ học và dễ nhớ. Python là ngôn ngữ có hình thức rất sáng sủa, cấu trúc rõ ràng, thuận tiện cho người mới học lập trình. Cấu trúc của Python còn cho phép người sử dụng viết mã lệnh với số lần gõ phím tối thiểu

2.1.5 FLASK :

Flask là một web frameworks, nó thuộc loại micro-framework được xây dựng bằng ngôn ngữ lập trình Python. Flask cho phép bạn xây dựng các ứng dụng web từ đơn giản tới phức tạp. Nó có thể xây dựng các api nhỏ, ứng dụng web chẳng hạn như các trang web, blog, trang wiki hoặc một website dựa theo thời gian hay thậm chí là một trang web thương mại. Flask cung cấp cho bạn công cụ, các thư viện và các công nghệ hỗ trợ bạn làm những công việc trên.

Flask là một micro-framework. Điều này có nghĩa Flask là một môi trường độc lập, ít sử dụng các thư viện khác bên ngoài. Do vậy, Flask có ưu điểm là nhẹ, có rất ít lỗi do ít bị phụ thuộc cũng như dễ dàng phát hiện và xử lý các lỗi bảo mật.

2.1.6 BOOTSTRAP :

Bootstrap là một framework bao gồm các HTML, CSS và JavaScript template dùng để phát triển website chuẩn responsive. Bootstrap cho phép quá trình thiết kế website diễn ra nhanh chóng và dễ dàng hơn dựa trên những thành tố cơ bản sẵn có như typography, forms, buttons, tables, grids, navigation, image carousels…

Bootstrap là một bộ sưu tập miễn phí của các mã nguồn mở và công cụ dùng để tạo ra một mẫu webiste hoàn chỉnh. Với các thuộc tính về giao diện được quy định sẵn như kích thước, màu sắc, độ cao, độ rộng…, các designer có thể sáng tạo nhiều sản phẩm mới mẻ nhưng vẫn tiết kiệm thời gian khi làm việc với framework này trong quá trình thiết kế giao diện website.

2.1.7 JQUERY :

jQuery là một thư viện JavaScript đa tính năng, nhỏ gọn, nhanh, được tạo bởi John Resig vào năm 2006 với một phương châm hết sức ý nghĩa: Write less, do more - Viết ít hơn, làm nhiều hơn.

jQuery đơn giản hóa việc duyệt tài liệu HTML, xử lý sự kiện, hoạt ảnh và tương tác Ajax để phát triển web nhanh chóng. Các phân tích web đã chỉ ra rằng, jQuery là thư viện JavaScript được triển khai rộng rãi nhất.

jQuery là một bộ công cụ JavaScript được thiết kế để đơn giản hóa các tác vụ khác nhau bằng cách viết ít code hơn.

2.1.8 HEROKU :

Heroku là nền tảng đám mây cho phép các lập trình viên xây dựng, triển khai, quản lý và mở rộng ứng dụng (PaaS – Platform as a service).

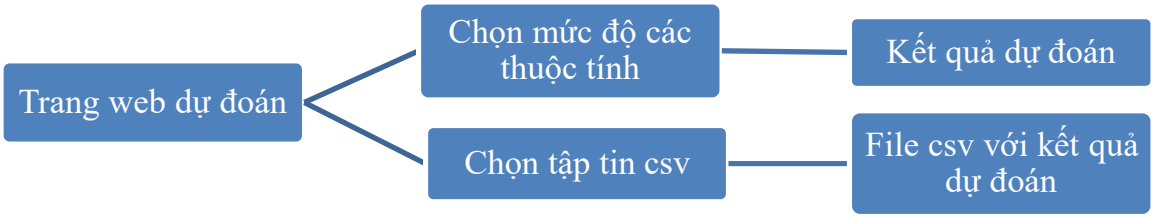
Nó rất linh hoạt và dễ sử dụng, cung cấp cho một con đường đơn giản nhất để đưa sản phẩm tiếp cận người dùng. Nó giúp các nhà phát triển tập trung vào phát triển sản phẩm mà không cần quan tâm đến việc vận hành máy chủ hay phần cứng…

2.2 : Yêu cầu hệ thống :

- **Máy chủ (server):**
  - Hệ điều hành: Windows 7 trở lên hoặc bất kì Ubuntu, LinuxMint, CentOS, và các distro linux khác có thể chạy được Apache, MySQL.
  - Phần mềm: cài đặt heroku, git, python và các gói cài cần thiết.
  - Phần cứng: CPU Pentium trở lên, RAM từ 1 GB, HDD hoặc SSD từ 50GB trở lên.
- **Máy khách (Client) :**
  - Laptop, máy tính bàn nếu chạy localhost.
  - Tất cả các thiết bị có thể truy cập Internet bằng trình duyệt truy cập đến link app trên Heroku

CHƯƠNG 2
THIẾT KẾ VÀ CÀI ĐẶT

1. Thiết kế hệ thống



Hình 1 - Mô hình hệ thống

- Khi người dùng truy cập vào trang web chọn các mức độ của thuộc tính và dự đoán sẽ nhận về kết quả có ly hôn hoặc không ly hôn

- Khi người dùng chọn vào file tải lên tập tin csv sẽ nhận được kết quả dự đoán bên trong tập tin

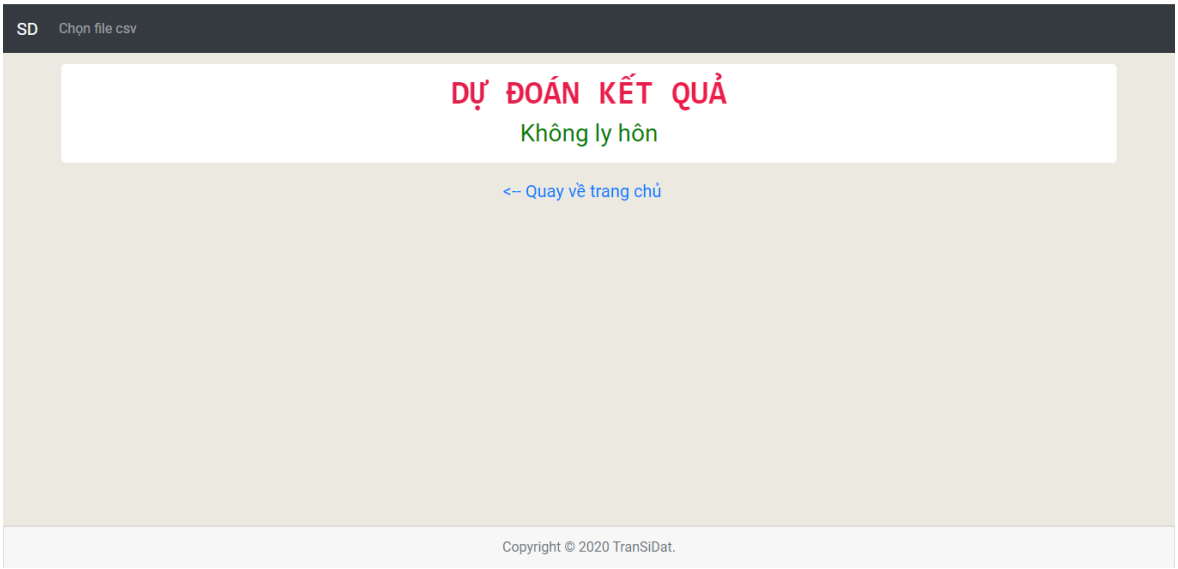
2. Cài đặt giải thuật

- Cài đặt giải thuật DecisionTree để xây dựng mô hình, sử dụng nghi thức K-Fold để kiểm tra và sử dụng chỉ số accuracy_score để đánh giá mô hình
- Sử dụng Flask để xây dựng phần xử lý, và sử dụng HTML, CSS, JS để xây dựng giao diện
- Cài đặt và triển khai hệ thống lên Heroku

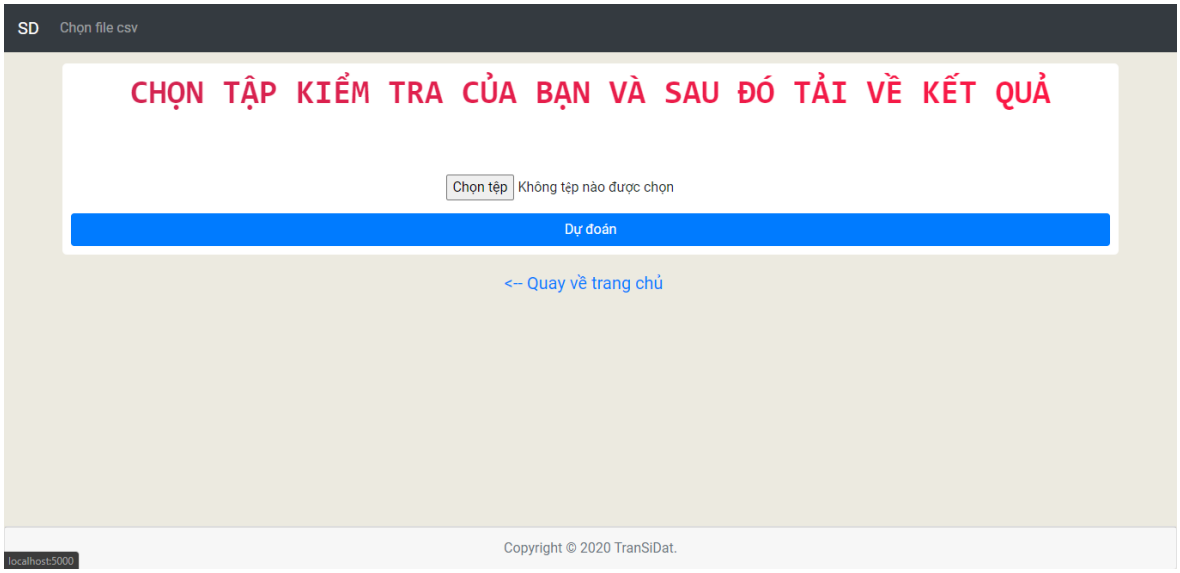
CHƯƠNG 3
KẾT QUẢ THỰC NGHIỆM

1. Kết quả kiểm tra

Hình 2 - Giao diện trang chủ



Hình 3 - Giao diện kết quả dự đoán



Hình 4 - Giao diện chọn file csv

Qua kết quả thực nghiệm, hệ thống cho kết quả chạy tốt. Có thể dự đoán trên một trường hợp bằng cách chọn các giá trị thuộc tính và dự đoán nhiều trường hợp bằng cách chọn tập tin csv

PHẦN KẾT LUẬN

1. Kết quả đạt được

- ✓ Xây dựng được website dự đoán tình trạng ly hôn trên các thuộc tính
- ✓ Chọn tập tin csv và trả về kết quả trong tập tin
- ✓ Triển khai website lên Heroku

2. Hướng phát triển

- ✓ Cải thiện tốc độ trang web, tối ưu code
- ✓ Thiết kế giao diện đẹp mắt và dễ sử dụng

TÀI LIỆU THAM KHẢO

1. https://vietnambiz.vn/khai-pha-du-lieu-data-mining-la-gi-nhung-dac-diem-can-luu-y-20191130175442498.htm (vietnambiz.vn : khai phá dữ liệu)

2. https://vi.wikipedia.org/wiki/Python_(ng%C3%B4n_ng%E1%BB%AF_l%E1%BA%ADp_tr%C3%ACnh) (wikipedia : ngôn ngữ python)

3. Flask Python là gì? Thư viện Flask trong lập trình Python (nguyenvanhieu.vn) (nguyenvanhieu.vn : Flask)

4. https://topdev.vn/blog/heroku-la-gi/ (topdev.vn : heroku)