TRƯỜNG ĐẠI HỌC VĂN LANG

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO ĐỒ ÁN THỰC TẬP 1,2**

NGÀNH: CÔNG NGHỆ THÔNG TIN

***Đề tài:***

**ỨNG DỤNG MÔ HÌNH HỌC MÁY ĐỂ DỰ ĐOÁN GIÁ TRỊ DINH DƯỠNG TRONG MÓN ĂN**

**SVTH: Nguyễn Thành Đạt**

**MSSV: 2174802010644**

***GVHD*: TH.S Phan Hồ Viết Trường**

TP. Hồ Chí Minh – năm 2024

Lời cảm ơn

Lời đầu tiên em xin được gửi lời cảm ơn sâu sắc tới thầy cô giáo trong khoa Công Nghệ Thông tin đã dìu dắt, giúp đỡ, hướng dẫn em trong suốt khoảng thời gian em học tập và làm báo cáo thực tập ở trường.

Em xin được gửi lời cảm ơn chân thành tới Ban giám hiệu trường Đại Học Văn Lang, Ban chủ nhiệm Khoa Công Nghệ Thông Tin đã cho em một môi trường học tập và rèn luyện bản thân thật tốt. Em cảm ơn đã cho em cảm nhận được không khí của đại gia đình Văn Lang từ những ngày đầu bước chân vào ngưỡng cửa đại học và bắt đầu quá trình trưởng thành tại trường.

Em cũng xin bày tỏ lòng biết ơn sâu sắc tới anh Hoàng Minh Phụng và chị Nguyễn Thị Yến Nhi công ty TNHH CUỘC SỐNG VUI KHỎE đã cho em cơ hội được trải nghiệm và phát triển bản thân khi thực tập tại công ty. Chị rất nhiệt tình chào đón và hỗ trợ em thực hiện đề tài một cách tận tình. Và quan trọng hơn chị đóng vai trò cung cấp tài liệu và chia sẻ một khối lượng không nhỏ những kiến thức thực tế không chỉ về mặt công nghệ mà còn về mặt xã hội, đời sống, khoa học và con người.

Em xin gửi lời cảm ơn chân thành và sâu sắc tới thầy Phan Hồ Viết Trường đã hướng dẫn em hoàn thành khóa thực tập này.

Em xin trân trọng cảm ơn!

*TP. Hồ Chí Minh, ngày tháng năm* 2024

**Sinh viên thực hiện**

*(Kí và ghi rõ họ tên)*

**CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**

**Độc lập – Tự do – Hạnh phúc**

**PHIẾU NHẬN XÉT SINH VIÊN**

**CỦA ĐƠN VỊ THỰC TẬP**

**Họ và tên sinh viên**: Nguyễn Thành Đạt

**Lớp**: CNTT22 **Khóa**: K27 **Khoa**: Công nghệ thông tin

**Trực thuộc Trường**: Đại học Văn Lang

Trong thời gian từ ngày 27 tháng 5 năm 2024 đến ngày 18 tháng 8 năm 2024

**Tại Công ty**: Công ty TNHH Cuộc sống Vui khỏe

**Địa chỉ**: 82 Phan Đăng Lưu, Phường 5, Phú Nhuận, TP.Hồ Chí Minh

Sau quá trình thực tập tại đơn vị của sinh viên, chúng tôi có một số nhận xét đánh giá như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Mục nhận xét** | **Nội dung nhận xét** | **Điểm**  **(thang điểm 10)** |
| 1 | ***Về ý thức tổ chức kỷ luật, thái độ*** | - Tuân thủ các quy định của công ty  - Có trách nhiệm với công việc được giao |  |
| 2 | ***Về kỹ năng (giao tiếp, quan hệ, học tập, làm việc)*** | - Giao tiếp tốt  - Có tinh thần tự học hỏi và hoàn thành công việc đúng hạn |  |

***Đánh giá chung sau khi thực tập:***

Ngày tháng năm 2024

**Người hướng dẫn trực tiếp Xác nhận của đơn vị thực tập**

*(Ký và ghi rõ họ tên) (Ký và ghi rõ họ tên, và đóng mộc Cty)*

**Hoàng Minh Phụng**

**CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**

**Độc lập – Tự do – Hạnh phúc**

**PHIẾU NHẬN XÉT SINH VIÊN**

**CỦA GIẢNG VIÊN HƯỚNG DẪN**

**Họ và tên sinh viên:** Nguyễn Thành Đạt

**MSSV**: 2174802010644.. **Lớp**: K27CNTT22 **Khoa:** Công nghệ thông tin

**Đơn vị thực tập**: Công ty TNHH Cuộc sống Vui khỏe

**Giảng viên hướng dẫn**: ThS. Phan Hồ Viết Trường

**Nhận xét**:

*TP. Hồ Chí minh, ngày tháng năm 2024*

**Giảng viên hướng dẫn**

*(Ký và ghi rõ họ và tên)*

**ThS. Phan Hồ Viết Trường**

**Mục Lục**

[CHƯƠNG 1: MỞ ĐẦU 8](#_Toc3940)

[1.1 Lý do chọn đề tài nghiên cứu: 8](#_Toc19033)

[CHƯƠNG 2: TÌM HIỂU CÔNG TY TNHH CUỘC SỐNG VUI KHỎE 8](#_Toc12752)

[2.1 Tổng quan về công ty 8](#_Toc12053)

[2.2 VuiKhoe.ai – Nền tảng chăm sóc sức khỏe trực tuyến 9](#_Toc26170)

[2.3 Liên hệ 10](#_Toc5667)

[CHƯƠNG 3: TỔNG QUAN VỀ DỰ ĐOÁN VỚI MÔ HÌNH RANDOM FOREST 11](#_Toc8145)

[3.1 Giới thiệu Flask 11](#_Toc10811)

[3.1.1 Lịch sử phát triển của Flask 11](#_Toc31118)

[3.1.2 Định nghĩa Flask 12](#_Toc20750)

[3.1.3 So sánh Flask với các framework khác 14](#_Toc1686)

[3.2 Nguyên lý hoạt động của Flask 14](#_Toc19609)

[3.2.1 Nhận yêu cầu (Request) 14](#_Toc13998)

[3.2.2 Xử lý yêu cầu 14](#_Toc13379)

[3.2.3 Trả về kết quả (Response) 15](#_Toc30581)

[3.3 Ví dụ minh họa Flask cơ bản: 15](#_Toc27582)

[3.4 Giới thiệu mô hình Random Forest 17](#_Toc16440)

[3.5 Kiến trúc mô hình Random Forest 19](#_Toc11768)

[3.6 Ứng dụng của Random Forest trong thực tiễn 21](#_Toc26014)

[1. Dự báo 21](#_Toc13482)

[2. Phân loại 21](#_Toc15231)

[3. Phát hiện gian lận 21](#_Toc2370)

[4. Y học 21](#_Toc32417)

[5. Khoa học tự nhiên 22](#_Toc12862)

[6. Kinh doanh 22](#_Toc3376)

[7. Xử lý ngôn ngữ tự nhiên 22](#_Toc18475)

[CHƯƠNG 4: ÁP DỤNG MÔ HÌNH RANDOM FOREST ĐỂ DỰ BÁO MỨC ĐỘ DINH DƯỠNG CỦA MỘT MÓN ĂN 22](#_Toc27696)

[4.1 Giới thiệu dinh dưỡng trong bữa ăn 22](#_Toc6128)

[4.2 Lợi ích dinh dưỡng đối với con người 23](#_Toc18742)

[4.3 Phát biểu vấn đề 24](#_Toc25605)

[4.4 Thu thập dữ liệu 25](#_Toc25382)

[4.4.1 Nguồn dữ liệu 25](#_Toc9411)

[4.5 Thiết kế Model 27](#_Toc23002)

[4.6 Thiết kế giao diện 27](#_Toc26622)

[4.7 Kết quả chương trình 27](#_Toc9161)

[CHƯƠNG 5: KẾT LUẬN VÀ ĐỀ XUẤT 28](#_Toc12919)

[5.1 Kết quả thực tập: 28](#_Toc8751)

[5.2 Các kiến nghị rút ra từ kết quả của đợt thực tập: 28](#_Toc6644)

[KẾ HOẠCH THỰC TẬP 29](#_Toc20300)

[TÀI LIỆU THAM KHẢO 31](#_Toc17660)

Mục lục hình ảnh

*[Hình 1.1: Logo Vuikhoe.ai](#_Toc8109)* [10](#_Toc8109)

*[Hình 2.2: Website Vuikhoe.ai](#_Toc9610)* [11](#_Toc9610)

*[Hình 2.3: Logo Flask](#_Toc13977)* [13](#_Toc13977)

*[Hình 3.4: Ứng dụng Flask đơn giản hiển thị "Hello, World!"](#_Toc11821)* [18](#_Toc11821)

*[Hình 3.5: Chạy ứng dụng](#_Toc20252)* [18](#_Toc20252)

*[Hình 3.6: Kết quả khi chạy ứng dụng Flask](#_Toc32123)* [19](#_Toc32123)

*[Hình 3.7: Kiến trúc mô hình Random Forest](#_Toc1613)* [23](#_Toc1613)

*[Hình 4.8: Tháp dinh dưỡng](#_Toc6587)* [27](#_Toc6587)

*[Hình 4.9: Trang web Fatsecret](#_Toc22708)* [29](#_Toc22708)

*[Hình 4.10: Trang web Nutritioninx](#_Toc26414)* [30](#_Toc26414)

*[Hình 4.11: Trang web Cookpad](#_Toc29699)* [30](#_Toc29699)

*[Hình 4.12: Trang web Wellplated](#_Toc5479)* [31](#_Toc5479)

**CHƯƠNG 1: MỞ ĐẦU**

**1.1 Lý do chọn đề tài nghiên cứu:**

Trong cuộc sống hiện đại, việc quan tâm đến chế độ ăn uống lành mạnh ngày càng trở nên phổ biến. Tuy nhiên, việc xác định chính xác giá trị dinh dưỡng của một món ăn lại không hề đơn giản. Với sự đa dạng của thực phẩm và cách chế biến, việc tính toán thủ công lượng calo, protein, chất béo và các chất dinh dưỡng khác là một công việc tốn thời gian và đòi hỏi kiến thức chuyên môn. Điều này khiến nhiều người gặp khó khăn trong việc xây dựng một chế độ ăn uống cân đối và khoa học.

Để giải quyết vấn đề trên, em đề xuất ứng dụng mô hình học máy, cụ thể là thuật toán Random Forest, để dự đoán giá trị dinh dưỡng trong món ăn. Mô hình này có khả năng học hỏi từ một lượng lớn dữ liệu về thành phần dinh dưỡng của các loại thực phẩm và cách chúng kết hợp với nhau trong các món ăn khác nhau. Nhờ đó, mô hình có thể đưa ra dự đoán chính xác về giá trị dinh dưỡng của một món ăn bất kỳ chỉ dựa trên danh sách các nguyên liệu và cách chế biến.

Ứng dụng này sẽ mang lại nhiều lợi ích cho người dùng Việt Nam. Đầu tiên, nó giúp người dùng dễ dàng quản lý chế độ ăn uống, đặc biệt là những người đang có nhu cầu giảm cân, tăng cân hoặc mắc các bệnh mãn tính. Thứ hai, ứng dụng cung cấp thông tin dinh dưỡng chi tiết, giúp người dùng hiểu rõ hơn về giá trị dinh dưỡng của các món ăn và đưa ra những lựa chọn tốt hơn. Cuối cùng, ứng dụng này có tiềm năng trở thành một công cụ hỗ trợ đắc lực cho các chuyên gia dinh dưỡng trong việc tư vấn cho khách hàng.

**CHƯƠNG 2: TÌM HIỂU CÔNG TY TNHH CUỘC SỐNG VUI KHỎE**

**2.1 Tổng quan về công ty**

Công ty TNHH Cuộc Sống Vui Khỏe là một doanh nghiệp hoạt động trong lĩnh vực công nghệ thông tin và chăm sóc sức khỏe. Với trụ sở chính đặt tại 82 Phan Đăng Lưu, phường 5, quận Phú Nhuận, Thành phố Hồ Chí Minh, công ty sở hữu nền tảng trực tuyến VuiKhoe.ai, cung cấp thông tin sức khỏe, dinh dưỡng và các giải pháp chăm sóc bản thân cho cộng đồng.



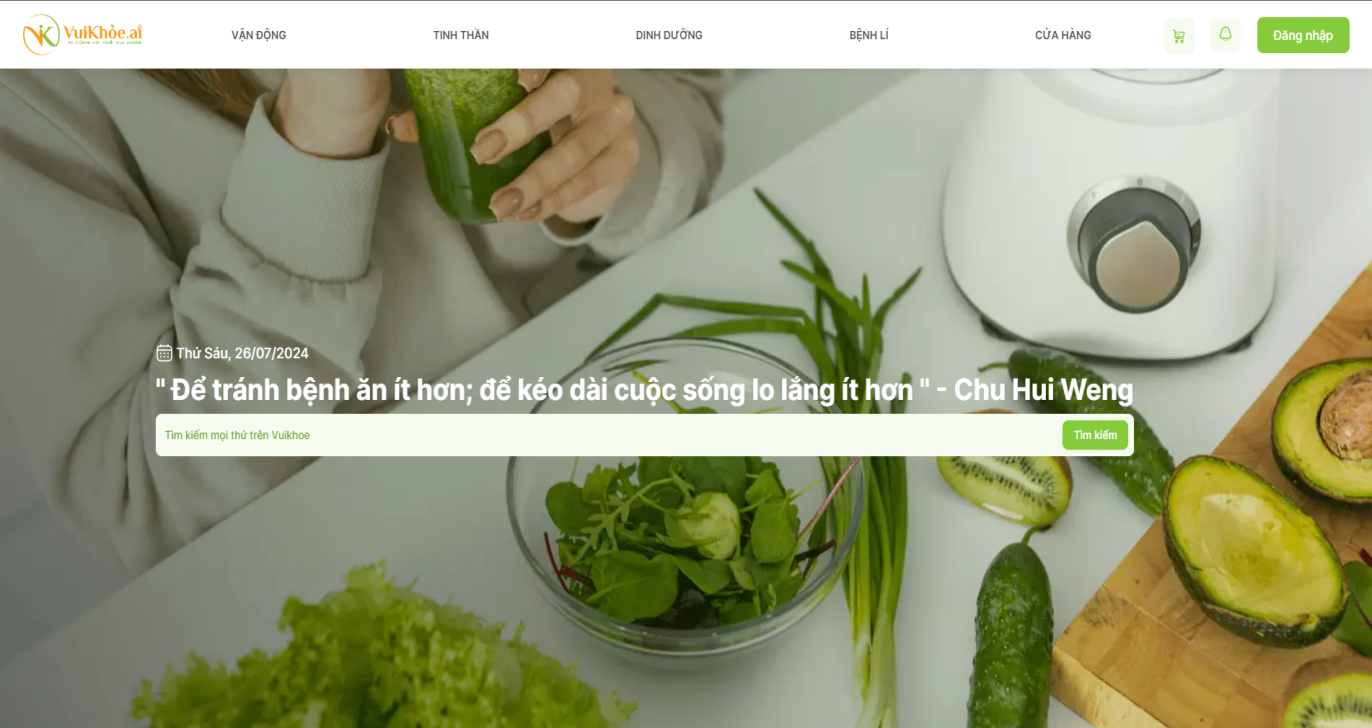
*Hình 1.1: Logo Vuikhoe.ai*

**2.2 VuiKhoe.ai – Nền tảng chăm sóc sức khỏe trực tuyến**

VuiKhoe.ai là một nền tảng trực tuyến toàn diện, cung cấp thông tin sức khỏe, dinh dưỡng và các giải pháp chăm sóc bản thân một cách khoa học và dễ hiểu. Với giao diện thân thiện và nội dung được cập nhật thường xuyên, VuiKhoe.ai đã trở thành người bạn đồng hành tin cậy của hàng triệu người Việt Nam trên con đường chăm sóc sức khỏe. Tính năng nổi bật:

* **Kho tàng kiến thức**: Cung cấp thông tin chi tiết về các bệnh lý, dinh dưỡng, sức khỏe tâm thần,...
* **Cộng đồng chia sẻ**: Nơi mọi người có thể trao đổi, chia sẻ kinh nghiệm về sức khỏe.
* **Cửa hàng trực tuyến**: Đa dạng sản phẩm chăm sóc sức khỏe, từ thực phẩm chức năng đến thiết bị y tế.
* **Tư vấn trực tuyến**: Kết nối với các chuyên gia y tế để được tư vấn và giải đáp thắc mắc.

VuiKhoe.ai hướng đến tất cả mọi người quan tâm đến sức khỏe của bản thân và gia đình, đặc biệt là những người bận rộn, không có nhiều thời gian tìm hiểu thông tin.

*Hình 2.2: Website Vuikhoe.ai*

**2.3 Liên hệ**

Địa chỉ: 82 Phan Đăng Lưu, phường 5, quận Phú Nhuận, Thành phố Hồ Chí Minh

Điện thoại: 0827000248

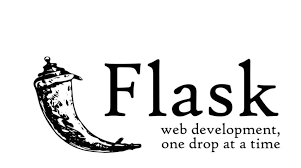
Email: email@vuikhoe.ai

**CHƯƠNG 3: TỔNG QUAN VỀ DỰ ĐOÁN VỚI MÔ HÌNH RANDOM FOREST**

**3.1 Giới thiệu Flask**

**3.1.1 Lịch sử phát triển của Flask**

Flask được phát triển bởi Armin Ronacher và cộng đồng Pocco. Nó được xây dựng dựa trên bộ công cụ Werkzeug WSGI và template engine Jinja2. Flask nhanh chóng trở thành một trong những framework phổ biến nhất trong cộng đồng Python nhờ sự đơn giản, linh hoạt và mạnh mẽ của nó.



*Hình 2.3: Logo Flask*

**Các cột mốc quan trọng:**

* **2010:** Flask được phát hành lần đầu tiên, nhanh chóng thu hút sự chú ý của cộng đồng Python nhờ sự gọn nhẹ và dễ tiếp cận.
* **Sử dụng Werkzeug và Jinja2:** Flask được xây dựng dựa trên hai dự án khác của Pocco là Werkzeug (một bộ công cụ WSGI) và Jinja2 (một template engine). Sự kết hợp này đã tạo nên nền tảng vững chắc cho Flask.
* **Cộng đồng phát triển mạnh mẽ:** Flask nhanh chóng có một cộng đồng người dùng lớn và sôi động. Điều này dẫn đến việc tạo ra nhiều extension và plugin, mở rộng khả năng của Flask.
* **Các phiên bản phát hành:** Flask liên tục được cập nhật với các phiên bản mới, bổ sung thêm tính năng và cải thiện hiệu năng.
* **Ứng dụng rộng rãi:** Flask được sử dụng để xây dựng nhiều loại ứng dụng web khác nhau, từ các ứng dụng đơn giản đến các API phức tạp.

**3.1.2 Định nghĩa Flask**

Flask là một micro-framework được xây dựng trên ngôn ngữ Python, được sử dụng để phát triển các ứng dụng web. Điều này có nghĩa là nó cung cấp các tính năng cốt lõi cần thiết để xây dựng một ứng dụng web, nhưng không bao gồm các thành phần quá phức tạp hoặc chuyên biệt. Điều này giúp cho Flask trở nên nhẹ nhàng, dễ học và dễ tùy chỉnh.

Các thành phần chính của Flask:

* **Werkzeug**: Là một bộ công cụ WSGI (Web Server Gateway Interface) cung cấp các lớp và hàm để xây dựng các ứng dụng web.
* **Jinja2**: Là một template engine mạnh mẽ, cho phép tạo ra HTML động một cách dễ dàng.
* **Router**: Là một thành phần giúp Flask xác định URL nào sẽ được xử lý bởi hàm nào.
* **Request và Response**: Là các đối tượng đại diện cho yêu cầu và phản hồi HTTP.

Ưu điểm của Flask:

* **Linh hoạt**: Cho phép tùy chỉnh mọi thứ theo ý muốn.
* **Dễ học**: Cú pháp đơn giản, dễ tiếp cận.
* **Cộng đồng lớn**: Có một cộng đồng người dùng đông đảo và sôi động, giúp dễ dàng tìm kiếm tài liệu và hỗ trợ.
* **Hiệu suất cao**: Nhờ cấu trúc đơn giản và tối ưu hóa, Flask có hiệu suất rất tốt.
* **Mở rộng**: Có rất nhiều extension và plugin giúp mở rộng chức năng của Flask.

Nhược điểm của Flask:

* **Ít tính năng mặc định**: So với các framework khác như Django, Flask cung cấp ít tính năng mặc định hơn. Điều này có thể yêu cầu tự cài đặt và cấu hình thêm các thư viện khác.
* **Độ dốc học tập có thể cao hơn**: Mặc dù Flask dễ học, nhưng để xây dựng các ứng dụng phức tạp, cần hiểu rõ hơn về các khái niệm về web development.

**3.1.3 So sánh Flask với các framework khác**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tính năng** | **Flask** | **Django** | **Bottle** |
| **Mức độ linh hoạt** | Rất cao | Cao | Rất cao |
| **Kích thước** | Nhỏ | Trung bình | Rất nhỏ |
| **Cộng đồng** | Lớn | Rất lớn | Nhỏ |
| **Tính năng mặc định** | Ít | Nhiều | Ít |
| **Dự án phù hợp** | Dự án nhỏ, vừa, API | Dự án lớn, phức tạp | Dự ản nhỏ, API |

**3.2 Nguyên lý hoạt động của Flask**

Flask hoạt động dựa trên kiến trúc request-response cycle. Khi một người dùng truy cập vào một URL trên ứng dụng Flask, một loạt các sự kiện diễn ra như sau:

**3.2.1 Nhận yêu cầu (Request)**

Khi người dùng nhập một URL vào trình duyệt, một yêu cầu HTTP được gửi đến server. Flask sẽ nhận và phân tích yêu cầu này, bao gồm:

* **Method**: Phương thức HTTP (GET, POST, PUT, DELETE,...)
* **URL**: Đường dẫn đến tài nguyên
* **Headers**: Các thông tin bổ sung về yêu cầu
* **Body**: Dữ liệu được gửi kèm theo yêu cầu (ví dụ: dữ liệu từ form)

**3.2.2 Xử lý yêu cầu**

**Routing**: Flask sẽ so sánh URL trong yêu cầu với các route đã được định nghĩa trong ứng dụng để tìm ra hàm xử lý (view function) phù hợp.

**Gọi view function**: Hàm view này sẽ được gọi để thực hiện các xử lý logic:

* Truy xuất dữ liệu từ cơ sở dữ liệu
* Thực hiện các phép tính
* Gọi các hàm hỗ trợ khác
* Render template (nếu cần)

**3.2.3 Trả về kết quả (Response)**

Sau khi xử lý xong, hàm view sẽ trả về một đối tượng Response. Đối tượng này chứa:

* **Status code**: Mã trạng thái HTTP (200 OK, 404 Not Found,...)
* **Headers**: Các thông tin bổ sung về phản hồi
* **Body:** Nội dung của phản hồi (thường là HTML)

Flask sẽ gửi response này về trình duyệt của người dùng.

**3.3 Ví dụ minh họa Flask cơ bản:**



*Hình 3.4: Ứng dụng Flask đơn giản hiển thị "Hello, World!"*

* **from flask import Flask**: Dòng này nhập thư viện Flask vào chương trình.
* **app = Flask(\_\_name\_\_)**: Tạo một đối tượng ứng dụng Flask. **\_\_name\_\_** ở đây thường bằng **\_\_main\_\_** khi chạy trực tiếp file này.
* **@app.route('/')**: Decorator này liên kết hàm **hello** với URL gốc **(/)**. Điều này có nghĩa là khi người dùng truy cập vào địa chỉ **http://127.0.0.1:5000/** (mặc định), hàm **hello** sẽ được gọi.
* **def hello():**: Đây là hàm view, nó sẽ được gọi khi có yêu cầu đến route **/**.
* **return 'Hello, World!'**: Hàm này trả về chuỗi "Hello, World!". Chuỗi này sẽ được hiển thị trên trình duyệt của người dùng.
* **if \_\_name\_\_ == '\_\_main\_\_':**: Đoạn code này đảm bảo rằng ứng dụng chỉ chạy khi file được thực thi trực tiếp, không phải khi được import làm module.
* **app.run()**: Khởi chạy ứng dụng Flask.

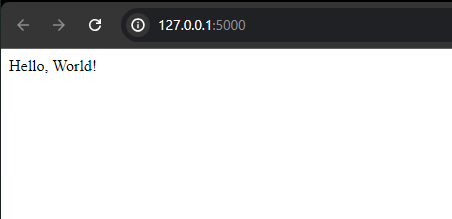
Cách chạy ứng dụng:

* **Lưu file**: Lưu đoạn code trên thành một file Python (ví dụ: **test.py**).
* **Mở terminal**: Mở terminal hoặc command prompt tại thư mục chứa file **test.py**.
* **Chạy ứng dụng**: Gõ lệnh **python test.py** và nhấn Enter.
* **Mở trình duyệt**: Mở trình duyệt và nhập địa chỉ **http://127.0.0.1:5000/**. Bạn sẽ thấy dòng chữ "Hello, World!" hiển thị trên màn hình.



*Hình 3.5: Chạy ứng dụng*

Kết quả sau khi chạy ứng dụng:



*Hình 3.6: Kết quả khi chạy ứng dụng Flask*

**3.4 Giới thiệu mô hình Random Forest**

Random Forest là một thuật toán học máy có giám sát, được sử dụng cho cả bài toán phân loại và hồi quy. Nó hoạt động bằng cách xây dựng một rừng gồm nhiều cây quyết định (Decision Tree), mỗi cây được xây dựng trên một tập dữ liệu huấn luyện ngẫu nhiên và một tập thuộc tính ngẫu nhiên.

Nguyên lý hoạt động:

1. **Tạo nhiều cây quyết định:**

* **Sampling**: Từ tập dữ liệu huấn luyện ban đầu, ngẫu nhiên chọn ra một tập dữ liệu con có kích thước bằng với tập dữ liệu gốc (với thay thế).
* **Feature selection**: Ngẫu nhiên chọn một tập con các thuộc tính từ tập thuộc tính ban đầu.
* **Xây dựng cây quyết định**: Trên tập dữ liệu con và tập thuộc tính con này, xây dựng một cây quyết định theo thuật toán Decision Tree.

1. **Kết hợp kết quả:**

* **Phân loại**: Để phân loại một mẫu dữ liệu mới, mỗi cây quyết định sẽ đưa ra một dự đoán. Dự đoán cuối cùng sẽ được quyết định bằng cách bỏ phiếu đa số.
* **Hồi quy**: Giá trị dự đoán cuối cùng sẽ là trung bình của các giá trị dự đoán từ các cây quyết định.

Ưu điểm của Random Forest:

* **Hiệu suất cao**: Thời gian huấn luyện và dự đoán tương đối nhanh.
* **Khả năng tổng quát hóa tốt**: Ít bị ảnh hưởng bởi nhiễu dữ liệu.
* **Không yêu cầu chuẩn hóa dữ liệu**: Có thể xử lý được cả dữ liệu số và dữ liệu danh mục.
* **Có thể xử lý dữ liệu thiếu**: Có nhiều cách để xử lý dữ liệu thiếu trong Random Forest.

Nhược điểm của Random Forest:

* **Mô hình phức tạp**: Rừng ngẫu nhiên có thể khó giải thích và hiểu rõ cơ chế hoạt động bên trong.
* **Thời gian huấn luyện có thể lâu**: Khi số lượng cây quyết định và số lượng dữ liệu lớn, thời gian huấn luyện có thể khá lâu.

Các công thức toán học liên quan:

* **Entropy**: Đo lường mức độ không chắc chắn của một tập dữ liệu. Entropy càng cao, dữ liệu càng hỗn loạn và khó phân loại..

Trong đó:

S: Tập dữ liệu

P(i): Tỷ lệ phần tử thuộc lớp i trong tập S

* **Information Gain**: Đo lường mức độ giảm đi của entropy sau khi phân chia dữ liệu theo một thuộc tính. Thuộc tính có information gain càng cao thì càng hữu ích trong việc phân chia dữ liệu.

Trong đó:

Entropy(P): Entropy của tập dữ liệu ban đầu (trước khi chia)

Entropy(i): Entropy của tập dữ liệu con thứ i sau khi chia theo thuộc tính

: Số lượng mẫu dữ liệu trong tập dữ liệu con thứ i

n: Tổng số mẫu dữ liệu trong tập dữ liệu ban đầu

k: Số lượng tập dữ liệu con sau khi chia

* **Gini Impurity**: Một thước đo khác để đánh giá độ tinh khiết của một tập dữ liệu. Gini impurity càng thấp, tập dữ liệu càng đồng nhất.

Trong đó:

S: Tập dữ liệu

P(i): Tỷ lệ phần tử thuộc lớp i trong tập S

**3.5 Kiến trúc mô hình Random Forest**

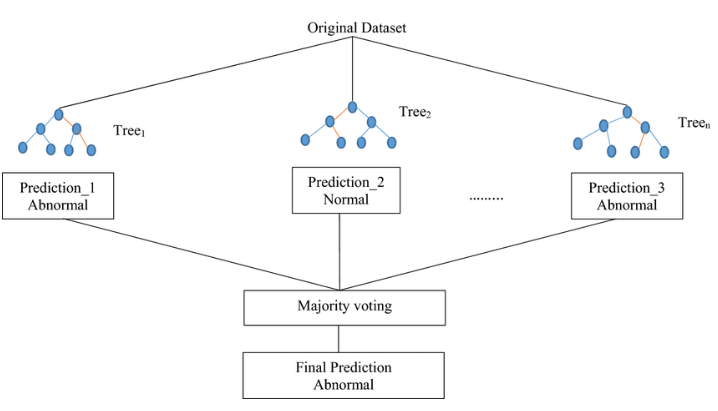
Để hiểu rõ hơn về cách thức hoạt động của Random Forest, chúng ta cần tìm hiểu kỹ hơn về kiến trúc của nó.

Kiến trúc Cấp cao

* **Rừng cây**: Mô hình Random Forest được hình dung như một "rừng" gồm nhiều cây quyết định. Mỗi cây quyết định trong rừng là một mô hình độc lập, được xây dựng trên một tập dữ liệu huấn luyện và tập thuộc tính ngẫu nhiên khác nhau.
* **Quyết định bằng đa số**: Khi có một mẫu dữ liệu mới cần dự đoán, mỗi cây trong rừng sẽ đưa ra một dự đoán. Dự đoán cuối cùng của rừng sẽ được quyết định bằng cách bỏ phiếu đa số.

Kiến trúc Cấp thấp (Chi tiết về mỗi cây quyết định)

* **Cây quyết định**: Mỗi cây quyết định là một mô hình đơn giản, gồm các nút và các cạnh.
* **Nút**: Mỗi nút trong cây đại diện cho một thuộc tính (feature) của dữ liệu.
* **Cạnh**: Các cạnh nối giữa các nút đại diện cho các giá trị có thể có của thuộc tính đó.
* **Lá**: Các nút lá đại diện cho kết quả dự đoán cuối cùng (lớp hoặc giá trị).



*Hình 3.7: Kiến trúc mô hình Random Forest*

**3.6 Ứng dụng của Random Forest trong thực tiễn**

Random Forest là một thuật toán học máy vô cùng linh hoạt và hiệu quả, được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau. Dưới đây là một số ứng dụng điển hình của Random Forest:

### 1. ****Dự báo****

* **Dự báo thời tiết:** Dự đoán nhiệt độ, lượng mưa, áp suất khí quyển,... dựa trên các dữ liệu lịch sử và các yếu tố ảnh hưởng khác như độ ẩm, gió.
* **Dự báo tài chính:** Dự đoán giá cổ phiếu, tỷ giá hối đoái, doanh thu của công ty,... dựa trên các chỉ số kinh tế, tin tức thị trường.
* **Dự báo nhu cầu:** Dự đoán nhu cầu về sản phẩm, dịch vụ để tối ưu hóa quy trình sản xuất và kinh doanh.

### 2. ****Phân loại****

* **Phân loại hình ảnh:** Phân loại các đối tượng trong hình ảnh (ví dụ: chó, mèo, xe hơi), nhận dạng khuôn mặt, phân tích cảm xúc từ biểu cảm khuôn mặt.
* **Phân loại văn bản:** Phân loại email (spam/không spam), phân loại tin tức theo chủ đề, phân tích sentiment (tình cảm tích cực/tiêu cực) trong các bình luận.
* **Phân loại khách hàng:** Phân loại khách hàng tiềm năng, phân loại khách hàng theo hành vi mua sắm.

### 3. ****Phát hiện gian lận****

* **Phát hiện gian lận thẻ tín dụng:** Phân biệt các giao dịch bình thường và các giao dịch gian lận dựa trên các đặc trưng như địa điểm, thời gian, số tiền giao dịch.
* **Phát hiện gian lận bảo hiểm:** Phát hiện các yêu cầu bồi thường gian lận dựa trên các thông tin về người được bảo hiểm, loại hình bảo hiểm, lịch sử yêu cầu bồi thường.

### 4. ****Y học****

* **Chẩn đoán bệnh:** Dự đoán bệnh dựa trên các triệu chứng, kết quả xét nghiệm.
* **Phân loại tế bào:** Phân loại tế bào ung thư và tế bào bình thường.
* **Dự đoán khả năng sống sót:** Dự đoán khả năng sống sót của bệnh nhân dựa trên các thông tin về bệnh án, điều trị.

### 5. ****Khoa học tự nhiên****

* **Phân loại loài:** Phân loại các loài thực vật, động vật dựa trên các đặc trưng hình thái, di truyền.
* **Dự báo thảm họa tự nhiên:** Dự báo động đất, sóng thần, bão lũ dựa trên các dữ liệu địa chất, khí tượng.

### 6. ****Kinh doanh****

* **Xây dựng hệ thống khuyến nghị:** Đề xuất sản phẩm cho khách hàng dựa trên lịch sử mua hàng và sở thích.
* **Phân tích rủi ro:** Đánh giá rủi ro tín dụng, rủi ro đầu tư.

### 7. ****Xử lý ngôn ngữ tự nhiên****

* **Phân tích sentiment:** Phân tích cảm xúc trong văn bản.
* **Nhận dạng thực thể:** Nhận dạng các thực thể như người, địa điểm, tổ chức trong văn bản.

**CHƯƠNG 4:   
ÁP DỤNG MÔ HÌNH RANDOM FOREST ĐỂ DỰ BÁO MỨC ĐỘ DINH DƯỠNG CỦA MỘT MÓN ĂN**

**4.1 Giới thiệu dinh dưỡng trong bữa ăn**

Dinh dưỡng đóng vai trò quan trọng trong việc duy trì sức khỏe và năng lượng cho cơ thể. Một bữa ăn đầy đủ dinh dưỡng cung cấp cho cơ thể các chất cần thiết để hoạt động tốt. Hãy cùng tìm hiểu sâu hơn về các yếu tố dinh dưỡng quan trọng trong mỗi bữa ăn và cách xây dựng một chế độ ăn uống cân bằng nhé.

Các yếu tố dinh dưỡng cần thiết trong bữa ăn

* **Carbohydrate**: Là nguồn năng lượng chính cho cơ thể. Có trong các loại ngũ cốc nguyên hạt, khoai lang, trái cây, và các loại đậu.
* **Protein**: Cần thiết cho việc xây dựng và sửa chữa các tế bào, mô. Có trong thịt, cá, trứng, đậu nành, các loại hạt.
* **Chất béo**: Cung cấp năng lượng, giúp hấp thụ các vitamin tan trong dầu và bảo vệ các cơ quan nội tạng. Có trong dầu ô liu, quả bơ, các loại hạt.
* **Vitamin**: Giúp cơ thể thực hiện nhiều chức năng quan trọng. Ví dụ: vitamin C tăng cường hệ miễn dịch, vitamin D giúp hấp thu canxi.
* **Khoáng chất**: Cần thiết cho sự phát triển xương, răng, duy trì cân bằng điện giải. Ví dụ: canxi, sắt, kẽm.
* **Nước**: Chiếm phần lớn trọng lượng cơ thể, giúp điều hòa nhiệt độ, vận chuyển chất dinh dưỡng.



*Hình 4.8: Tháp dinh dưỡng*

**4.2 Lợi ích dinh dưỡng đối với con người**

Dinh dưỡng đóng vai trò vô cùng quan trọng trong việc duy trì sức khỏe và đảm bảo sự phát triển toàn diện của cơ thể. Một chế độ ăn uống cân bằng và cung cấp đầy đủ các chất dinh dưỡng cần thiết sẽ mang lại nhiều lợi ích cho sức khỏe con người.

Các lợi ích nổi bật:

* **Tăng cường hệ miễn dịch**: Dinh dưỡng đầy đủ giúp cơ thể sản sinh ra các tế bào miễn dịch, giúp chống lại các tác nhân gây bệnh, bảo vệ cơ thể khỏi nhiễm trùng.
* **Cung cấp năng lượng**: Các chất dinh dưỡng như carbohydrate, protein và chất béo cung cấp năng lượng cần thiết cho các hoạt động hàng ngày của cơ thể.
* **Hỗ trợ quá trình tăng trưởng và phát triển**: Đặc biệt quan trọng đối với trẻ em và thanh thiếu niên, dinh dưỡng đầy đủ giúp xương chắc khỏe, cơ bắp phát triển và não bộ hoạt động tốt.
* **Ngăn ngừa bệnh tật**: Một chế độ ăn uống lành mạnh giúp giảm nguy cơ mắc các bệnh mãn tính như tim mạch, tiểu đường, béo phì và một số loại ung thư.
* **Cải thiện tâm trạng**: Các chất dinh dưỡng như vitamin B và omega-3 có tác dụng tích cực đến tâm trạng, giúp giảm stress và cải thiện giấc ngủ.
* **Làm đẹp da và tóc**: Dinh dưỡng đầy đủ cung cấp các chất cần thiết để duy trì làn da khỏe mạnh, mái tóc bóng mượt và móng tay chắc khỏe.

**4.3 Phát biểu vấn đề**

Với sự phổ biến của các ứng dụng theo dõi sức khỏe và xu hướng ăn uống lành mạnh, việc xác định chính xác giá trị dinh dưỡng của một món ăn trở nên ngày càng quan trọng. Tuy nhiên, truy xuất thông tin dinh dưỡng cho các món ăn thường khó khăn và tốn thời gian, đặc biệt đối với các món ăn tự chế biến hoặc có nguồn gốc phức tạp. Ví dụ như:

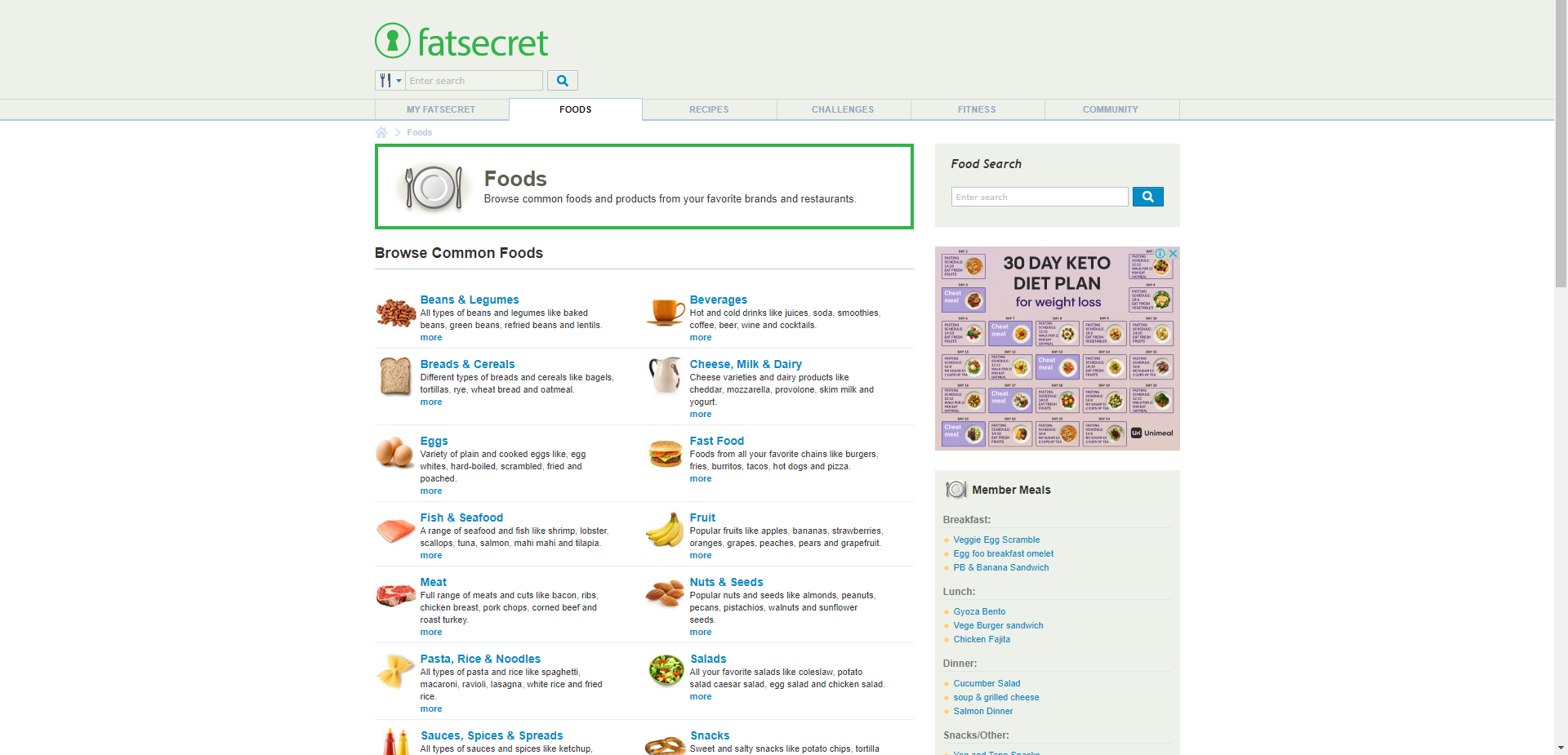
* Bạn có thể không biết lượng calo, chất béo, protein và carbs chính xác trong một bát phở.
* Bạn có thể khó ước tính lượng chất dinh dưỡng trong món salad tự chế biến của bạn, tùy thuộc vào các thành phần và nước sốt.

Ứng dụng mô hình học máy, cụ thể là thuật toán Random Forest, để dự đoán giá trị dinh dưỡng trong món ăn. Thuật toán này có khả năng học hỏi từ một lượng lớn dữ liệu về thành phần dinh dưỡng của các loại thực phẩm và cách chúng kết hợp với nhau trong các món ăn khác nhau. Dựa vào đó, mô hình có thể dự đoán giá trị dinh dưỡng của một món ăn bất kỳ chỉ dựa trên danh sách các nguyên liệu và cách chế biến.

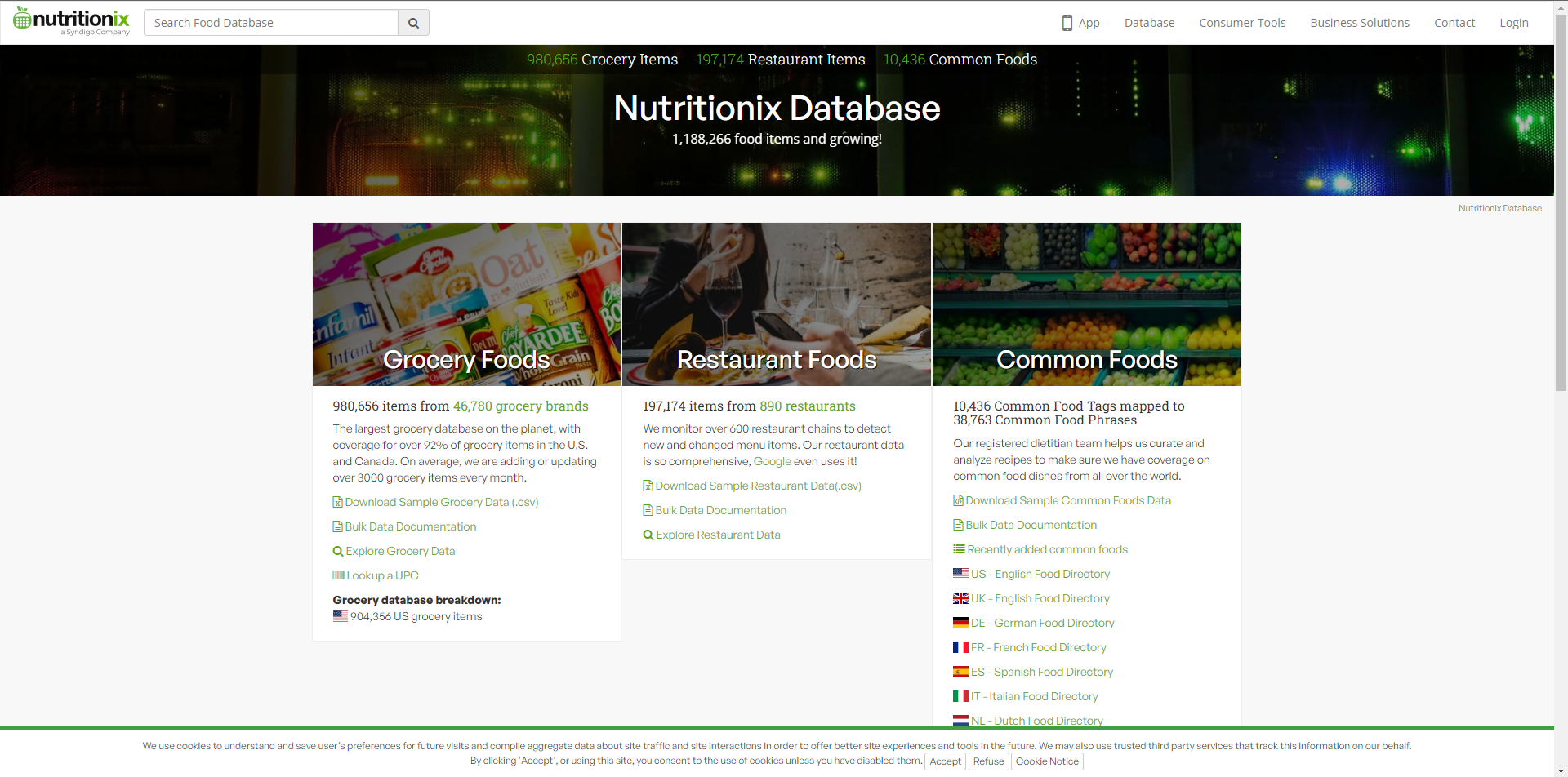
**4.4 Thu thập dữ liệu**

**4.4.1 Nguồn dữ liệu**

Em đã chọn hai nguồn chính để thu thập dữ liệu dinh dưỡng: Fatsecret và Nutritionix. Đây là những trang web nổi tiếng và đáng tin cậy, cung cấp thông tin chi tiết về giá trị dinh dưỡng của hàng ngàn thực phẩm. Fatsecret và Nutritionix đều cung cấp dữ liệu phong phú về hàm lượng calo, protein, carbohydrate, chất béo, và các vi chất dinh dưỡng khác. Điều này cho phép chúng tôi có được dữ liệu chính xác và đa dạng để phục vụ cho các phân tích về dinh dưỡng

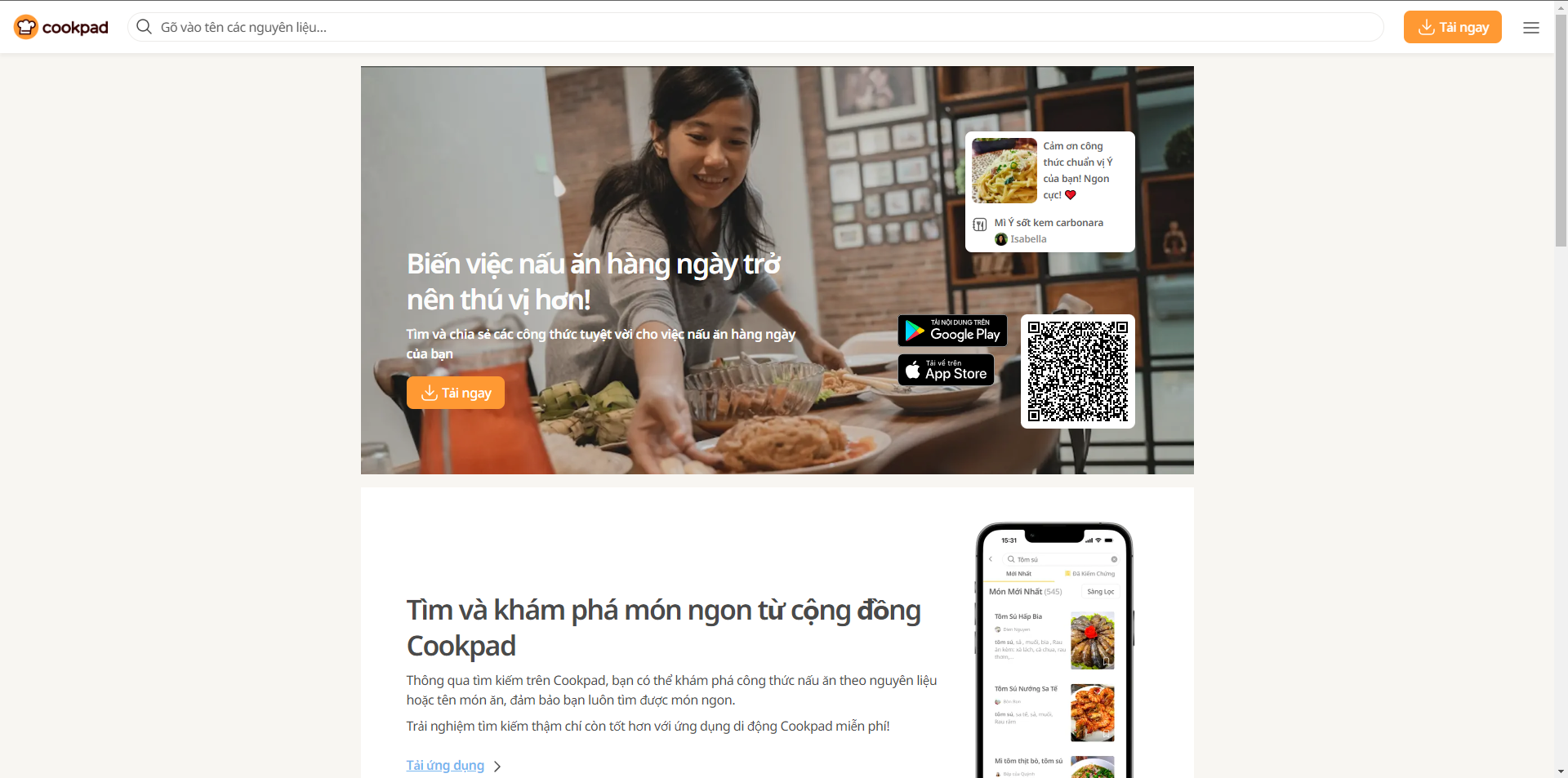


*Hình 4.9: Trang web Fatsecret*

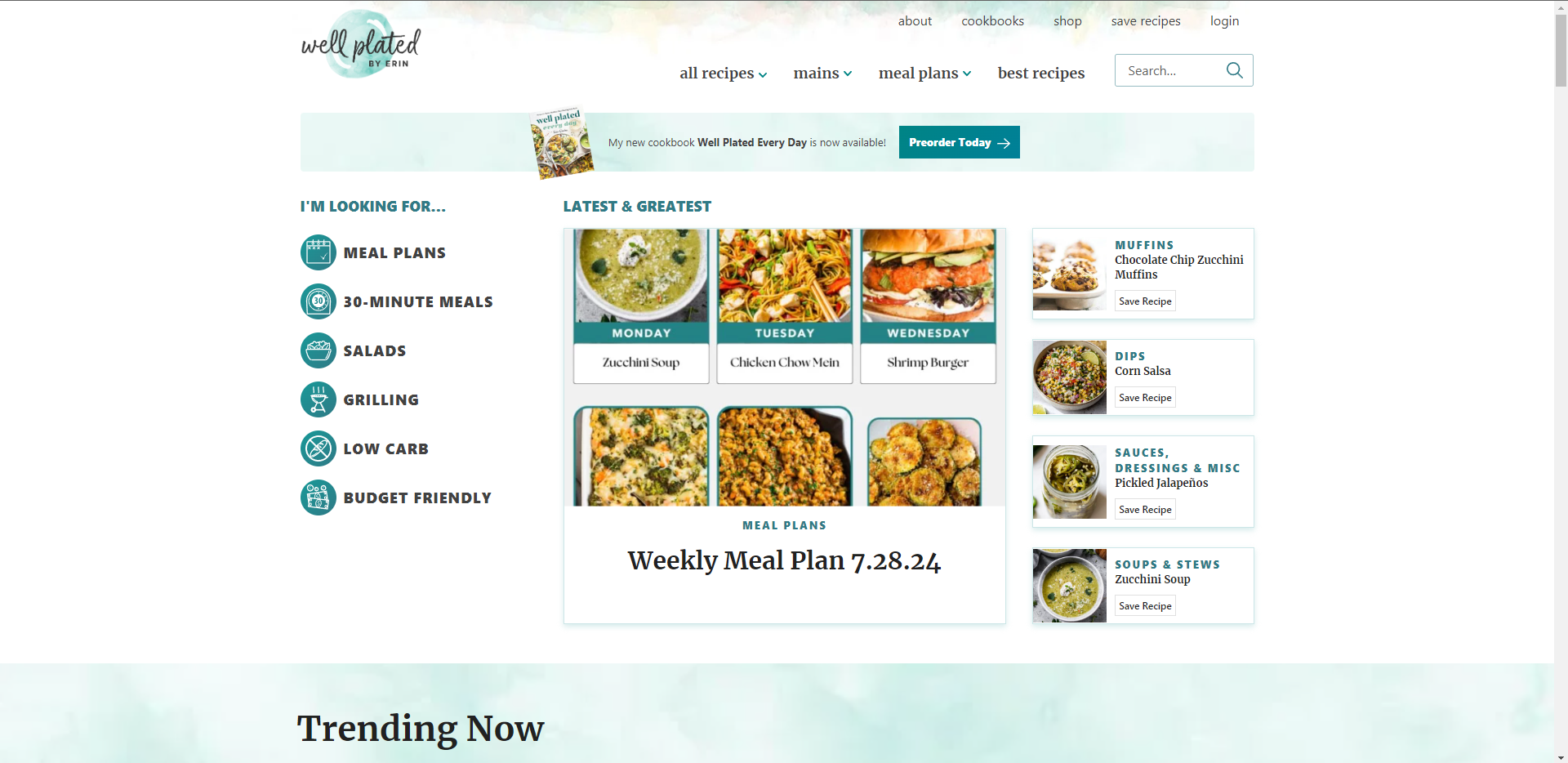


*Hình 4.10: Trang web Nutritioninx*

Về phần các công thức nấu ăn, chúng tôi thu thập từ các trang web hướng dẫn nấu ăn thông dụng. Các trang này cung cấp một kho tàng công thức đa dạng từ các món ăn hàng ngày đến các món ăn đặc sản. Các trang web này thường bao gồm các công thức từ nhiều nền ẩm thực khác nhau, mang lại sự phong phú và đa dạng trong lựa chọn công thức.

**

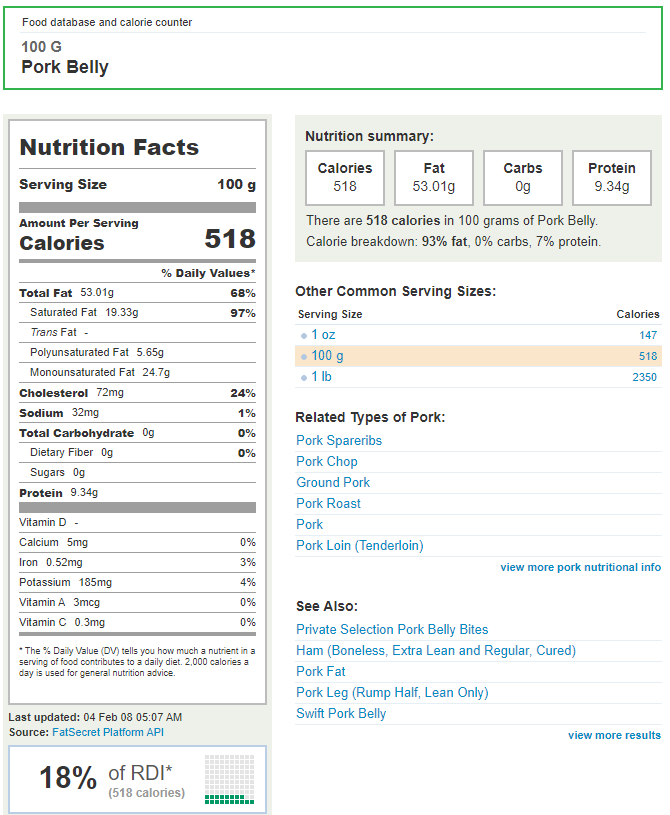
*Hình 4.11: Trang web Cookpad*



*Hình 4.12: Trang web Wellplated*

**4.4.2 Phương pháp thu thập dữ liệu**

Quá trình thu thập dữ liệu dinh dưỡng được thực hiện thủ công. Chúng tôi tìm kiếm bảng dinh dưỡng của từng loại nguyên liệu trên Fatsecret và Nutritionix, sau đó ghi lại các giá trị dinh dưỡng tương ứng với các khối lượng khác nhau. Việc thu thập thủ công này đảm bảo rằng chúng tôi có thể kiểm tra và xác nhận thông tin cho từng nguyên liệu cụ thể, đảm bảo tính chính xác và chi tiết của dữ liệu.



*Hình 4.13: Bảng dinh dưỡng*

Đối với các công thức nấu ăn, chúng tôi chọn lọc các công thức từ các trang web hướng dẫn nấu ăn uy tín. Mỗi công thức được kiểm tra kỹ lưỡng về các thành phần và quy trình nấu ăn để đảm bảo tính toàn vẹn và chính xác của thông tin.

**4.4.3 Lọc và Xử lý dữ liệu**

Sau khi thu thập dữ liệu, chúng tôi tiến hành quá trình lọc và xử lý để đảm bảo chất lượng. Điều này bao gồm loại bỏ các dữ liệu không hợp lệ hoặc không đủ thông tin, chuẩn hóa các đơn vị đo lường để đảm bảo tính nhất quán, và xác nhận lại các giá trị dinh dưỡng để đảm bảo độ chính xác.

Thu thập dữ liệu thủ công là một quá trình tốn thời gian và công sức. Một trong những thách thức lớn nhất là việc đảm bảo tính nhất quán và chính xác của dữ liệu từ các nguồn khác nhau. Ngoài ra, một số nguyên liệu và công thức đặc biệt có thể không có sẵn thông tin chi tiết trên các trang web, đòi hỏi phải tìm kiếm và xác minh thêm từ các nguồn khác.

**4.5 Thiết kế Model**

**4.6 Thiết kế giao diện**

**4.7 Kết quả chương trình**

**CHƯƠNG 5: KẾT LUẬN VÀ ĐỀ XUẤT**

**5.1 Kết quả thực tập:**

Sau khoảng thời gian thực tập, tôi nghiên cứu đã được kết quả như sau:

- Hiểu thêm được phân tích dự báo và một trong những bước để phân tích dự báo nhu cầu của khách hàng.

- Phân tích, xử lý dữ liệu và trực quan hóa bằng các charts trong mô hình dự báo.

- Vận dụng cái kỹ năng kiến thức đã học ở trường để giải quyết công việc, tuy nhiên em đã học hỏi thêm từ các anh chị từ đó mà tiến bộ hơn

- Trao dòi thêm kỹ năng tự học qua việc đọc tài liệu, sách và các bài viết về Phân tích dự báo (predictive analytics) từ các nguồn khách nhau để mở rộng thêm kiến thức và tư duy.

Tuy nhiên, thời gian có hạn nên còn nhiều kiến thức về Phân tích dự báo và một số công việc vẫn chưa hoàn thành kịp.

**5.2 Các kiến nghị rút ra từ kết quả của đợt thực tập:**

Đợt thực tập này đã cho em một trải nghiệm một phần thiết thực về phân tích dự báo và những kiến thức hoàn toàn mới. Tuy bước đầu thực hiện còn nhiều khó khăn, song nhờ sự giúp đỡ nhiệt tình của các cô và anh chị trong công ty, em đã có thể hoàn thành đề tài này.

Thời gian thực tập còn ngắn nên em vẫn chưa học hỏi được nhiều những nghiệp vụ cũng như tiếp cận nhiều hơn với các công cụ các mô hình dự báo. Hy vọng đợt thực tập sắp tới sẽ có thời gian nhiều hơn để nhóm có thể có nhiều cơ hội trải nghiệp hơn ở doanh nghiệp.

Trong tương lai, em sẽ cố gắng học hỏi và tìm hiểu các kiến thức liên quan đến phân tích dự báo, để có thể ứng dụng vào công việc và ngày càng hoàn thiện hơn trong lĩnh vực này.

**KẾ HOẠCH THỰC TẬP**

|  |  |  |  |
| --- | --- | --- | --- |
| **Thời gian** | **Nội dung công việc** | | |
| **Sinh viên** | **Giảng viên hướng dẫn tại Khoa** | **Người hướng dẫn tại doanh nghiệp** |
| **Tuần đăng ký** | * **Đăng ký làm Đồ án thực tập 2** | * **Gửi thông tin đăng ký thực tập** * **Xét tiêu chuẩn làm Đồ án thực tập 2 của sinh viên đăng ký.** * **Phân công GVHD.** * **Công bố danh sách làm Đồ án thực tập 2, công ty thực tập và GVHD** | * **Cung cấp các yêu cầu cho chương trình thực tập tại doanh nghiệp** |
| **Tuần chuẩn bị** | * **Liên lạc với GVHD** | * **Hướng dẫn quy trình làm thực tập (công việc, deadline)** * **Lên kế hoạch báo cáo với GVHD.** * **Cung cấp các biểu mẫu thực tập.** |  |
| **Tuần thứ nhất** | * **Liên lạc với doanh nghiệp.** * **Hoàn tất các thủ tục theo yêu cầu của đơn vị nhận thực tập.** * **Sinh viên đến làm việc thực tế tại đơn vị thực tập và tìm hiểu doanh nghiệp.** * **Lập kế hoạch thực tập.** | * **Hướng dẫn SV chọn đề tài** * **Hướng dẫn SV lập kế hoạch thực tập.** |  |
| **Tuần thứ 2** | * **Sinh viên làm việc thực tế tại đơn vị thực tập và tìm hiểu các công việc thực hiện.** * **Chọn đề tài/lĩnh vực mình muốn nghiên cứu/thực tập (có cân nhắc đến thời gian hoàn thành và phạm vi nghiên cứu phù hợp với yêu cầu của học phần).** * **Viết và nộp đề cương chi tiết cho GVHD.** | * **Kiểm tra tiến độ thực hiện của sinh viên.** * **Hướng dẫn viết đề cương chi tiết.** * **Chỉnh sửa và thông qua đề cương chi tiết.** |  |
| **Tuần thứ 3-7** | * **Sinh viên làm việc thực tế tại đơn vị thực tập và phải thực hiện các công việc mà đơn vị thực tập giao** * **Sinh viên phải quan sát, tìm hiểu các hoạt động thực tế có liên quan chủ đề thực tập; phỏng vấn trực tiếp những người có liên quan; thu thập thông tin và dữ liệu chuẩn bị cho việc viết báo cáo Đồ án thực tập 2.** | * **Kiểm tra tiến độ thực hiện của sinh viên** * **Trao đổi và hướng dẫn sinh viên các vấn đề liên quan đến thực tập và viết báo cáo Đồ án thực tập 2.** * **Nhận Đề cương chi tiết của sinh viên và ký tên xác nhận đã duyệt.** |  |
| **Tuần thứ 8** | * **Sinh viên làm việc thực tế tại đơn vị thực tập.** * **Hoàn thiện và chỉnh sửa báo cáo Đồ án thực tập 2.** | * **Kiểm tra tiến độ thực hiện của sinh viên.** * **Kiểm tra nội dung báo cáo Đồ án thực tập 2** |  |
| **Tuần thứ 9** | * **Sinh viên làm việc thực tế tại đơn vị thực tập.** * **Gửi báo cáo Đồ án thực tập 2 đến cơ quan thực tập xin xác nhận.** * **Nộp báo cáo Đồ án thực tập 2 cho khoa.** | * **Kiểm tra tiến độ thực hiện của sinh viên.** | * **Nhận xét và đánh giá báo cáo Đồ án thực tập 2 của sinh viên.** |
| **Tuần thứ 10** | * **Trình bày và báo cáo Đồ án thực tập 2** | * **Chấm điểm báo cáo Đồ án thực tập 2 qua buổi báo cáo.** |  |
| **Tuần thứ 11** | * **Xem điểm tổng kết và phản hồi nếu có.** | * **Tổng hợp và công bố điểm cho sinh viên** * **Gửi email cảm ơn doanh nghiệp.** |  |

*TP. Hồ Chí Minh, ngày …. tháng …. năm 2024*

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn**  *(Ký và ghi rõ họ tên)* | **Sinh viên thực hiện**  *(Ký và ghi rõ họ tên)* |

**TÀI LIỆU THAM KHẢO**

[1] TechFlow, “Dataiku trong AI”, 2022  
link:<https://techflow.vn/tech-data-hop-tac-voi-dataiku-de-thuc-day-ung-dung-tri-tue-nhan-tao-ai-trong-doanh-nghiep/>

[2] Cision, “Dataiku phân tích dữ liệu”, 2021  
link: <https://vn.prnasia.com/story/60613-11.shtml>

[3] VietNambiz, “Hệ hỗ trợ quyết định trong Dataiku”, 2019

Link: <https://vietnambiz.vn/he-ho-tro-quyet-dinh-decision-support-system-la-gi-20190810192517123.htm>

[4] MeeyLand, “Dataiku là gì?”, 2022   
link: <https://meeyland.com/dau-tu/dss-la-gi-nhung-loi-ich-ma-dss-dem-den-cho-doanh-nghiep/>

1. BigData Solution, “Tổng quan về Preditive analytics” truy cập ngày 6/11/2022, link:<https://bigdatauni.com/tin-tuc/tong-quan-ve-predictive-analytics-phan-tich-du-bao-phan-1-khai-niem-va-loi-ich.html>
2. devoteam, “What is Dataiku ?”, truy cập ngày 6/11/2022, link: <https://www.devoteam.com/expert-view/what-is-dataiku/>
3. VIBLO, “Confusion Matrix/Ma trân nhầm lẫn/Ma trân lỗi <P1>, truy cập ngày 7/11/2022 link: <https://viblo.asia/p/confusion-matrix-ma-tran-nham-lan-ma-tran-loi-p1-V3m5WQB7ZO7>
4. Dataiku Knowledge, “Model Evaluation”, truy cập ngày 8/11/2022, link: <https://knowledge.dataiku.com/latest/courses/intro-to-ml/predictive-modeling/model-eval-summary.html?highlight=roc%20curve>
5. GIM, “Understanding the AUC-ROC Curve in Machine Learning Classification”, truy cập ngày 13/11/2022, link: <https://analyticsindiamag.com/understanding-the-auc-roc-curve-in-machine-learning-classification/>
6. Microsolf, “Lift Chart (Analysis Services - Data Mining)", tuy cập ngày 17/11/2022,

link: <https://learn.microsoft.com/en-us/analysis-services/data-mining/lift-chart-analysis-services-data-mining?view=asallproducts-allversions>

1. Course, “Calibration Curves”, truy câph ngày 19/11/2022,

link: <https://www.geeksforgeeks.org/calibration-curves/>

1. CESTI, “Trực quan hóa dữ liệu – Phần 5: Các dạng biểu đồ thể hiện sự phân phối của dữ liệu” , truy cập ngày 21/11/2022,

Link: http://thongke.cesti.gov.vn/dich-vu-thong-ke/tai-lieu-phan-tich-thong-ke/963-truc-quan-hoa-du-lieu-phan-5-cac-dang-bieu-do-the-hien-su-phan-phoi-cua-du-lieu