KHOA KỸ THUẬT VÀ CÔNG NGHỆ

**BỘ MÔN CÔNG NGHỆ THÔNG TIN**

****

**THỰC TẬP ĐỒ ÁN CƠ SỞ NGÀNH**

**HỌC KỲ I NĂM HỌC 2023**

**TRỰC QUAN HÓA DỮ LIỆU**

**BẰNG ORANGE**

*Giáo viên hướng dẫn:*ThS. NGUYỄN KHẮC QUỐC

*Sinh viên thực hiện:*

Họ tên: PHẠM KHÁNH ĐĂNG

MSSV: 110121009

Lớp: DA21TTB

***Trà Vinh, tháng 01 năm 2024***

KHOA KỸ THUẬT VÀ CÔNG NGHỆ

**BỘ MÔN CÔNG NGHỆ THÔNG TIN**



**THỰC TẬP ĐỒ ÁN CƠ SỞ NGÀNH**

**HỌC KỲ I NĂM HỌC 2023**

**TRỰC QUAN HÓA DỮ LIỆU**

**BẰNG ORANGE**

*Giáo viên hướng dẫn:*ThS. NGUYỄN KHẮC QUỐC

*Sinh viên thực hiện:*

Họ tên: PHẠM KHÁNH ĐĂNG

MSSV: 110121009

Lớp: DA21TTB

***Trà Vinh, tháng 01 năm 2024***

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

*Trà Vinh, ngày ….. tháng …… năm ……*

**Giáo viên hướng dẫn**

*(Ký tên và ghi rõ họ tên)*

*Trà Vinh, ngày ….. tháng …… năm ……*

**Giáo viên hướng dẫn**

*(Ký tên và ghi rõ họ tên)*

**LỜI CẢM ƠN**

**NHẬN XÉT CỦA THÀNH VIÊN HỘI ĐỒNG**

*Trà Vinh, ngày ….. tháng …… năm ……*

**Thành viên hội đồng**

*(Ký tên và ghi rõ họ tên)*

Em tên là Phạm Khánh Đăng sinh viên lớp DA21TTB em xin chân thành cảm ơn đến thầy THs Nguyễn Khắc Quốc đã tận tình hướng dẫn, giúp đỡ em trong suốt thời gian qua. Nhờ có thầy hướng dẫn mà em đã hoàn thành đồ án theo đúng tiến độ.

Em xin chân thành cảm ơn!

**MỤC LỤC**

[**MỤC LỤC** 6](#_Toc155557906)

[**TÓM TẮT ĐỒ ÁN CƠ SỞ NGÀNH** 9](#_Toc155557907)

[**MỞ ĐẦU** 11](#_Toc155557908)

[**CHƯƠNG 1: TỔNG QUAN** 13](#_Toc155557909)

[1. Tại sao lại cần khai phá dữ liệu. 13](#_Toc155557910)

[2. Khai phá dữ liệu. 13](#_Toc155557911)

[3. Các chức năng chính của khai phá dữ liệu 13](#_Toc155557912)

[4. Ứng dụng của khai phá dữ liệu 14](#_Toc155557913)

[**CHƯƠNG 2: NGHIÊN CỨU LÝ THUYẾT** 15](#_Toc155557914)

[I. Trực quan hóa dữ liệu 15](#_Toc155557915)

[1. Khái niệm 15](#_Toc155557916)

[2. Ứng dụng 15](#_Toc155557917)

[3. Những điều cần lưu ý 16](#_Toc155557918)

[II. Orange 16](#_Toc155557919)

[1. Khái niệm 16](#_Toc155557920)

[2. Các bước tạo ra một biểu đồ 18](#_Toc155557921)

[4. Các tính năng của Orange 19](#_Toc155557927)

[5. Các mô hình phân lớp dữ liệu 22](#_Toc155557928)

[6. Quy trình phân lớp dữ liệu 25](#_Toc155557929)

[**CHƯƠNG 3: HIỆN THỰC HÓA NGHIÊN CỨU** 28](#_Toc155557932)

[1. Giao diện chính của phần mềm 28](#_Toc155557933)

[2. Thực hiện trực quan hóa 28](#_Toc155557934)

[**CHƯƠNG 4: KẾT QUẢ NGHIÊN CỨU** 37](#_Toc155557936)

[**CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN** 38](#_Toc155557937)

[1. Kết Luận 38](#_Toc155557938)

[2. Hướng phát triển 38](#_Toc155557939)

[**DANH MỤC TÀI LIỆU THAM KHẢO** 39](#_Toc155557940)

**DANH MỤC HÌNH ẢNH – BẢNG BIỂU**

[Hình 2. 1. Biểu đồ đường 16](#_Toc155558349)

[Hình 2. 2. Biểu đồ cột 16](#_Toc155558350)

[Hình 2. 3. Biểu đồ tròn 16](#_Toc155558351)

[Hình 2. 4. Biểu đồ tán xạ 17](#_Toc155558352)

[Hình 2. 5. Biểu đồ miền 17](#_Toc155558353)

[Hình 2. 6. Chức năng của Data 19](#_Toc155558354)

[Hình 2. 7. Chức năng của Visualize 19](#_Toc155558355)

[Hình 2. 8. Chức năng của Model 20](#_Toc155558356)

[Hình 2. 9. Chức năng của Evaluate 20](#_Toc155558357)

[Hình 2. 10. Chức năng của Unsuperived 21](#_Toc155558358)

[Hình 2. 11. Mô hình Logistic Regression 21](#_Toc155558359)

[Hình 2. 12. Mô hình Decision tree 22](#_Toc155558360)

[Hình 2. 13. Mô hình Support Vector Machine 23](#_Toc155558361)

[Hình 2. 14. Phân cụm dữ liệu 24](#_Toc155558362)

[Hình 2. 15. Phân cụm phân cấp bằng phương pháp Diana 25](#_Toc155558363)

[Hình 2. 16. Kết quả phân cụm bằng Diana 25](#_Toc155558364)

[Hình 2. 17. Phân cụm phân hoạch bằng thuật toán K-means 26](#_Toc155558365)

[Hình 3. 1. Giao diện chính của Orange 27](#_Toc155558366)

[Hình 3. 2. File Excel 28](#_Toc155558367)

[Hình 3. 3. Lệnh file 29](#_Toc155558368)

[Hình 3. 4. Hộp thoại file 29](#_Toc155558369)

[Hình 3. 5. Chọn file 30](#_Toc155558370)

[Hình 3. 6. Hiện cột của file 30](#_Toc155558371)

[Hình 3. 7 Liên kết file với Data Table 31](#_Toc155558372)

[Hình 3. 8. Mở Data Table 31](#_Toc155558373)

[Hình 3. 9. Trực quan hóa các giá trị số 32](#_Toc155558374)

[Hình 3. 10. Thanh công cụ Visualize 32](#_Toc155558375)

[Hình 3. 11. Kếtt nối Bar Plot 33](#_Toc155558376)

[Hình 3. 12. Biểu đồ được trực quan hóa 33](#_Toc155558377)

[Hình 3. 13. Thay đổi thuộc tính 34](#_Toc155558378)

[Hình 3. 14. Kết nối nhiều lại biểu đồ 34](#_Toc155558379)

[Hình 3. 15. Scatter Plot 35](#_Toc155558380)

[Hình 3. 16. Box Plot 35](#_Toc155558381)

**TÓM TẮT ĐỒ ÁN CƠ SỞ NGÀNH**

Trực quan hóa dữ liệu bằng Orange là quá trình sử dụng công cụ phân tích dữ liệu Orange để biểu diễn dữ liệu dưới dạng đồ thị, biểu đồ hoặc hình ảnh để hiểu và phân tích dữ liệu một cách dễ dàng và trực quan. Orange cung cấp nhiều công cụ và chức năng để trực quan hóa dữ liệu, bao gồm: biểu đồ phân phối, biểu đồ tương quan, biểu đồ cây quyết định, biểu đồ mạng, biểu đồ thời gian. Orange là một thư viện quy trình và đối tượng cốt lõi của C++ kết hợp rất nhiều thuật toán Data mining và học máy tiêu chuẩn và không tiêu chuẩn. Nó là một công cụ trực quan hóa dữ liệu Data mining và học máy mã nguồn mở. Orang là một môi trường có thể viết kịch bản để tạo mẫu nhanh cho các thuật toán và mẫu thử nghiệm mới nhất.

Hướng tiếp cận: Bước đầu tiên là xác định mục tiêu của việc trực quan hóa dữ liệu. Điều này sẽ giúp chúng ta xác định loại dữ liệu cần được thu thập và cách dữ liệu sẽ được hiển thị. Bước tiếp theo là thu thập dữ liệu cần thiết. Chúng ta có thể thu thập dữ liệu từ nhiều nguồn khác nhau, chẳng hạn như cơ sở dữ liệu, bảng tính hoặc khảo sát. Bước thứ ba là xử lí dữ liệu, điều này bao gồm việc làm sạch dữ liệu, loại bỏ các lỗi và mâu thuẫn. Bạn cũng có thể cần phải chuyển đổi dữ liệu sang định dạng phù hợp với loại biểu đồ hoặc đồ thị mà bạn muốn tạo. Bước thứ tư là chọn biểu đồ hoặc đồ thị phù hợp để hiển thị dữ liệu lên. Có nhiều loại đồ thị và biểu đồ khác nhau, mỗi loại có ưu điểm và nhược điểm riêng. Bước cuối cùng là tạo biểu đồ hoặc đồ thị. Chúng ta có thể sử dụng một công cụ trực quan hóa dữ liệu hoặc tạo thủ công bằng phần mềm đồ họa. Khi tạo biểu đồ hoặc đồ thị, điều quan trọng là phải làm cho chúng rõ ràng và dễ hiểu. Chúng ta cũng cần sử dụng các màu sắc và biểu tượng phù hợp để truyền tải thông tin một cách hiệu quả.

Hãy dành thời gian để xem xét loại dữ liệu đang hiển thị, mục tiêu của việc trực quan hóa dữ liệu và người xem.

Màu sắc và biểu tượng là một cách để làm cho dữ liệu trở nên trực quan hơn. Hãy sử dụng màu sắc và biểu tượng một cách hợp lý. Hãy sử dụng các màu sắc và biểu tượng phù hợp với nhau và không quá nhiều.

Các nhãn và chú thích là điều cần thiết để giúp người xem hiểu dữ liệu. Nếu không có các nhãn và chú thích rõ ràng, người xem có thể gặp khó khăn trong việc hiểu ý nghĩa của dữ liệu. Hãy luôn sử dụng các nhãn và chú thích rõ ràng và dễ hiểu. Hãy đảm bảo rằng các nhãn và chú thích của chúng ta cung cấp thông tin cần thiết để người xem hiểu được dữ liệu.

Trước khi tạo bất kỳ biểu đồ hoặc đồ thị nào, hãy luôn kiểm tra xem dữ liệu có chính xác hay không. Dữ liệu không chính xác có thể dẫn đến kết quả trực quan hóa dữ liệu không chính xác. Hãy luôn sử dụng các nguồn dữ liệu đáng tin cậy và kiểm tra dữ liệu kỹ lưỡng trước khi tạo bất kỳ biểu đồ hoặc đồ thị nào.

Trực quan hóa dữ liệu bằng Orange giúp chúng ta hiểu rõ hơn về dữ liệu, phát hiện thông tin tìm ẩn, truyền đạt thông tin một cách dễ hiểu, hỗ trợ quyết định và tạo r báo cáo và trình bày.

**MỞ ĐẦU**

Lí do chọn đề tài:

* **Trực quan hóa dữ liệu là một lĩnh vực quan trọng của khoa học dữ liệu, có ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau.**
* **Trực quan hóa dữ liệu giúp chúng ta đưa ra các quyết định, dự báo hoặc cải thiện hiệu suất của các hệ thống một cách chính xác và hiệu quả hơn.**

Cụ thể, trực quan hóa dữ liệu có những lợi ích sau:

* **Giúp chúng ta hiểu dữ liệu một cách trực quan và dễ dàng hơn.**
* **Giúp chúng ta phát hiện các mẫu, xu hướng và mối quan hệ trong dữ liệu mà chúng ta không thể nhìn thấy bằng mắt thường.**
* **Giúp chúng ta truyền đạt thông tin về dữ liệu một cách hiệu quả hơn.**
* **Giúp chúng ta đưa ra các quyết định, dự báo hoặc cải thiện hiệu suất của các hệ thống một cách chính xác và hiệu quả hơn.**

Trong thời đại công nghệ thông tin hiện nay, dữ liệu ngày càng trở nên phong phú và phức tạp hơn. Trực quan hóa dữ liệu là một công cụ quan trọng giúp chúng ta hiểu và phân tích dữ liệu một cách hiệu quả. Do đó, việc nghiên cứu về trực quan hóa dữ liệu là một lĩnh vực có nhiều tiềm năng phát triển.

Mục đích của việc trực quan hóa dữ liệu là để biến dữ liệu trở nên dễ hiểu và dễ tiếp cận hơn. Trực quan hóa dữ liệu sử dụng các hình ảnh, biểu tượng và đồ thị để truyền tải thông tin từ dữ liệu.Trực quan hóa dữ liệu là một công cụ mạnh mẽ có thể được sử dụng cho nhiều mục đích khác nhau. Bằng cách sử dụng trực quan hóa dữ liệu một cách hiệu quả, chúng ta có thể giúp mọi người hiểu rõ hơn về dữ liệu và đưa ra quyết định sáng suốt hơn.

Trực quan hóa dữ liệu có thể được sử dụng cho nhiều mục đích khác nhau, bao gồm:

+ **Hiểu dữ liệu:** Trực quan hóa dữ liệu có thể giúp mọi người hiểu rõ hơn về dữ liệu bằng cách cung cấp cho họ một cách trực quan để xem dữ liệu. Điều này có thể hữu ích cho việc xác định các xu hướng, mẫu và mối tương quan trong dữ liệu.

**+Trực quan hóa dữ liệu cho người khác:** Trực quan hóa dữ liệu có thể được sử dụng để truyền đạt thông tin dữ liệu cho người khác. Điều này có thể hữu ích cho việc trình bày dữ liệu cho các nhà lãnh đạo doanh nghiệp, khách hàng hoặc các bên liên quan khác.

**+Làm cho dữ liệu hấp dẫn hơn:** Trực quan hóa dữ liệu có thể giúp làm cho dữ liệu hấp dẫn và dễ tiếp cận hơn. Điều này có thể hữu ích cho việc thu hút sự chú ý của mọi người và giúp họ hiểu rõ hơn về dữ liệu.

**Đối tượng nghiên cứu của trực quan hóa dữ liệu bằng Orange:** là các dữ liệu ở dạng số, văn bản hoặc hình ảnh. Dữ liệu này có thể được thu thập từ nhiều nguồn khác nhau, chẳng hạn như bảng tính, cơ sở dữ liệu, khảo sát hoặc mạng xã hội.

**Phạm vi của trực quan hóa dữ liệu bằng Orange:** bao gồm việc tạo ra các biểu đồ và đồ thị để trực quan hóa dữ liệu. Orange cung cấp một loạt các loại biểu đồ và đồ thị phổ biến, cũng như các tính năng nâng cao như phân tích dữ liệu lớn và học máy.

Việc xác định đối tượng và phạm vi của trực quan hóa dữ liệu bằng Orange phụ thuộc vào mục tiêu của nghiên cứu. Nếu mục tiêu của nghiên cứu là hiểu rõ hơn về dữ liệu, thì đối tượng nghiên cứu có thể là tất cả các dữ liệu có liên quan đến nghiên cứu. Phạm vi của nghiên cứu có thể bao gồm việc tạo ra các biểu đồ và đồ thị để hiển thị các xu hướng, mẫu và mối tương quan trong dữ liệu.

Nếu mục tiêu của nghiên cứu là truyền đạt thông tin dữ liệu cho người khác, thì đối tượng nghiên cứu có thể là dữ liệu cần được truyền đạt. Phạm vi của nghiên cứu có thể bao gồm việc tạo ra các biểu đồ và đồ thị dễ hiểu và dễ tiếp cận cho đối tượng người xem cụ thể.

**CHƯƠNG 1: TỔNG QUAN**

1. **Tại sao lại cần khai phá dữ liệu.**

Khoảng hơn một thập kỷ trở lại đây, lượng thông tin lưu trữ trên các thiết bị điện tử không ngừng tăng lên. Sự tích lũy dữ liệu này xảy ra với một tốc độ bùng nổ. Người ta ước đoán rằng lượng thông tin trên toàn cầu tăng gấp đôi sau khoảng hai năm và theo đó số lượng cũng như kích cỡ của các cơ sở dữ liệu cũng tăng lên một cách nhanh chóng. Câu hỏi đặt ra là liệu chúng ta có thể khai thác được những “núi” dữ liệu tưởng chừng bỏ đi ấy không.

1. **Khai phá dữ liệu.**

Khai phá dữ liệu (Data Mining) được định nghĩa như là một quá trình chắt lọc hay khai phá tri thức từ một lượng lớn dữ liệu. Thuật ngữ Data Mining ám chỉ việc tìm kiếm một tập hợp nhỏ có giá trị từ một số lượng lớn các dữ liệu. Có nhiều thuật ngữ hiện được dùng cũng có nghĩa tương tự với từ Data Mining như: Knowledge Mining (khai phá tri thức), Knowledge extraction (chắt lọc tri thức), Data archaeoloog (khóa cổ dữ liệu),...

**Định nghĩa :** Khai phá dữ liệu là một tập hợp các kĩ thuật được sử dụng để tự động khai thác và tìm ra các mối quan hệ lẫn nhau của dữ liệu trong một tập hợp dữ liệu khổng lồ và phức tạp, đồng thời cũng tìm ra các mẫu tìm ẩn trong tập dữ liệu đó.

1. **Các chức năng chính của khai phá dữ liệu**

Data Mining được chia nhỏ thành một số hướng chính như sau:

* Mô tả khái niệm: thiên về mô tả, tổng hợp và tóm tắt khái niệm.
* Luật kết hợp: là dạng luật biểu diễn tri thức ở dạng khá đơn giản . Ví dụ 60% nam giới vào siêu thị mua bia thì có 80% trong số họ sẽ mua thêm thịt bò. Luật kết hợp được ứng dụng nhiều trong lĩnh vực kinh doanh, y học, tài chính...
* Phân lớp và dự đoán: xếp một đối tượng vào một trong những lớp đã biết trước, hướng tiếp cận này thường được sử dụng một số kĩ thuật của Machine Learning như cây quyết định, mạng nơ ron nhân tạo,... Người ta còn gọi phân lớp là học có giám sát.
* Phân cụm: xếp các đối tượng theo từng cụm. Người ta còn gọi phân cụm là học không giám sát.
* Khai phá chuỗi: tương tự như khai phá luận kết hợp nhưng có thêm tính thứ tự và tính thời gian. Hướng tiếp cận này được ứng dụng nhiều lĩnh vực tài chính và thị trường chứng khoán vì nó có tính dự báo cao.

1. **Ứng dụng của khai phá dữ liệu**

Data Mining tuy là một hướng tiếp cận mới nhưng thu hút được rất nhiều sự quan tâm của các nhà nghiên cứu và phát triển nhờ vào những ứng dụng thực tiễn của nó. Một số ứng dụng điển hình:

* Phân tích dữ liệu và hỗ trợ ra quyết định.
* Điều trị y học.
* Text mining & Web mining.
* Tin - sinh.
* Tài chính và thị trường chứng khoán.
* Bảo hiểm.
* Nhận dạng.
* ...

**CHƯƠNG 2: NGHIÊN CỨU LÝ THUYẾT**

1. **Trực quan hóa dữ liệu**
2. **Khái niệm**

**Trực quan hóa dữ liệu** là quá trình sử dụng các yếu tố hình ảnh như biểu đồ, đồ thị hoặc bản đồ để trình bày dữ liệu. Quá trình này chuyển đổi dữ liệu phức tạp, có dung lượng lớn hoặc dữ liệu số thành hình ảnh trình bày trực quan có thể xử lý dễ dàng hơn.

Trực quan hóa dữ liệu có nhiều lợi ích, bao gồm:

* **Giúp người dùng hiểu dữ liệu dễ dàng hơn:** Trực quan hóa dữ liệu có thể giúp người dùng hiểu dữ liệu phức tạp một cách nhanh chóng và dễ dàng hơn. Điều này là do não bộ con người có thể xử lý thông tin hình ảnh nhanh hơn nhiều so với thông tin văn bản.
* **Kích thích sự khám phá:** Trực quan hóa dữ liệu có thể kích thích người dùng khám phá dữ liệu và tìm thấy các mối quan hệ và xu hướng mà họ có thể không nhận thấy bằng cách chỉ nhìn vào dữ liệu ở dạng văn bản.
* **Thúc đẩy giao tiếp hiệu quả:** Trực quan hóa dữ liệu có thể giúp giao tiếp dữ liệu một cách hiệu quả hơn. Điều này có thể hữu ích cho việc chia sẻ dữ liệu với người khác, chẳng hạn như các nhà ra quyết định, khách hàng hoặc đồng nghiệp.

1. **Ứng dụng**

Trực quan hóa dữ liệu được ứng dụng trong nhiều lĩnh vực, bao gồm:

* **Kinh doanh**: Trực quan hóa dữ liệu được sử dụng để phân tích dữ liệu kinh doanh, giúp doanh nghiệp hiểu rõ hơn về tình hình hoạt động, đưa ra các quyết định kinh doanh hiệu quả.
* **Khoa học**: Trực quan hóa dữ liệu được sử dụng để trình bày dữ liệu khoa học, giúp các nhà khoa học dễ dàng truyền đạt kết quả nghiên cứu của mình.
* **Giáo dục**: Trực quan hóa dữ liệu được sử dụng để giảng dạy, giúp học sinh dễ dàng tiếp thu kiến thức.
* **Báo chí**: Trực quan hóa dữ liệu được sử dụng để truyền tải tin tức, giúp người đọc dễ dàng hiểu và ghi nhớ thông tin.

1. **Những điều cần lưu ý**

Để tạo ra một biểu đồ hoặc đồ thị trực quan hóa dữ liệu hiệu quả, cần lưu ý các nguyên tắc sau:

* Sử dụng các yếu tố trực quan rõ ràng, dễ hiểu.
* Chọn loại biểu đồ hoặc đồ thị phù hợp với dữ liệu và mục đích sử dụng.
* Sử dụng màu sắc và kích thước phù hợp để thu hút sự chú ý của người dùng.
* Thêm chú thích và thông tin bổ sung để giúp người dùng hiểu rõ hơn về dữ liệu.

Trực quan hóa dữ liệu là một kỹ năng quan trọng trong thời đại dữ liệu bùng nổ. Việc nắm vững các kiến thức và kỹ năng trực quan hóa dữ liệu sẽ giúp chúng ta dễ dàng hiểu và phân tích dữ liệu, đưa ra các quyết định hiệu quả hơn.

1. **Orange**
2. **Khái niệm**

Phần mềm Orange được biết đến như một sự kết hợp giữa các công cụ khai thác dữ liệu khác nhau và là một máy học đơn giản, thông minh được lập trình bằng Python với sự tương tác dễ dàng và một giao diện trực quan. Qua nhiều chức năng đã được tích hợp, phần mềm này có thể dễ dàng phân tích các tài liệu, dữ liệu được cho từ đơn giản tới phức tạp; tạo ra các đồ họa biểu lộ rõ ràng, phong phú các khía cạnh thông tin mà người sử dụng muốn biểu hiện trực tiếp đồng thời giảm mức độ khó khăn trong việc khai thác dữ liệu và học máy cho chuyên gia cũng như người mới học.

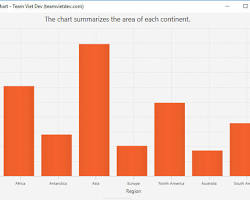
Trực quan hóa dữ liệu bằng Orange là một quá trình sử dụng các công cụ và chức năng của Orange để tạo ra các biểu đồ và đồ thị trực quan hóa dữ liệu. Orange cung cấp một loạt các công cụ trực quan hóa dữ liệu, bao gồm:

* Line chart: Dùng để thể hiện xu hướng thay đổi của dữ liệu theo thời gian.



**Hình 2. 1. Biểu đồ đường**

* Bar chart: Dùng để so sánh các giá trị của dữ liệu giữa các nhóm khác nhau.



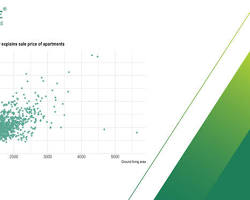
**Hình 2. 2. Biểu đồ cột**

* Pie chart: Dùng để thể hiện tỷ lệ của các thành phần trong tổng thể.



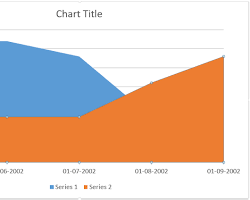
**Hình 2. 3. Biểu đồ tròn**

* Scatter plot: Dùng để thể hiện mối quan hệ giữa hai biến số.



**Hình 2. 4. Biểu đồ tán xạ**

* Area chart: Dùng để thể hiện tổng của các giá trị dữ liệu theo thời gian



**Hình 2. 5. Biểu đồ miền**

1. **Các bước tạo ra một biểu đồ**

Để tạo ra một biểu đồ hoặc đồ thị trực quan hóa dữ liệu bằng Orange, bạn có thể thực hiện theo các bước sau:

B1: Mở tập dữ liệu trong Orange.

B2: Nhấp vào tab "Visualize".

B3: Chọn loại biểu đồ hoặc đồ thị mà chúng ta muốn tạo.

B4: Chọn các biến dữ liệu mà bạn muốn sử dụng trong biểu đồ hoặc đồ thị.

B5: Thiết lập các tùy chọn của biểu đồ hoặc đồ thị.

Orange cũng cung cấp một số công cụ để giúp bạn tùy chỉnh biểu đồ và đồ thị của mình. Bạn có thể thay đổi màu sắc, kích thước, kiểu dáng và các thuộc tính khác của biểu đồ và đồ thị để tạo ra một biểu đồ hoặc đồ thị trực quan hóa dữ liệu hiệu quả và hấp dẫn.

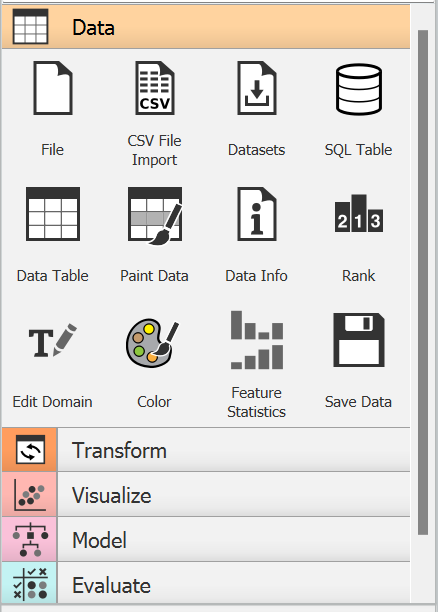
**3. Ứng dụng**

Ứng dụng Orange tạo điều kiện thuận lợi cho các hoạt động thống kê, xây dựng mô hình, thực nghiệm, dự đoán nhằm cung cấp các thông tin cần thiết cho việc đưa ra quyết định trên các lĩnh vực như:

* Trong kinh tế, Orange giúp đỡ các nhà kinh tế trong việc xác định các biến chủ yếu ảnh hưởng đến sự phát triển kinh tế của quốc gia trong một khoảng thời gian cụ thể, từ đó các nhà kinh tế có tạo dựng các báo cáo kinh tế hợp lý, gần sát với thực tế hơn.
* Trong y học, Orange được xem như một nền tảng nơi mà các triệu chứng bệnh được thiết lập mối quan hệ với nhau, từ đó cho ra chẩm đoán về bệnh và phương hướng trị liệu bệnh đó.
* Trong mạng viễn thông, phần mềm được sử dụng để phân tích các cuộc gọi điện thoại nhằm phân tích hành vi người tiêu dùng với mục đích nâng cao chất lượng dịch vụ.
* Trong giáo dục, Orange được sử dụng như một công cụ giảng dạy của giáo viên về phương pháp học máy và khai phá dữ liệu cho học sinh.

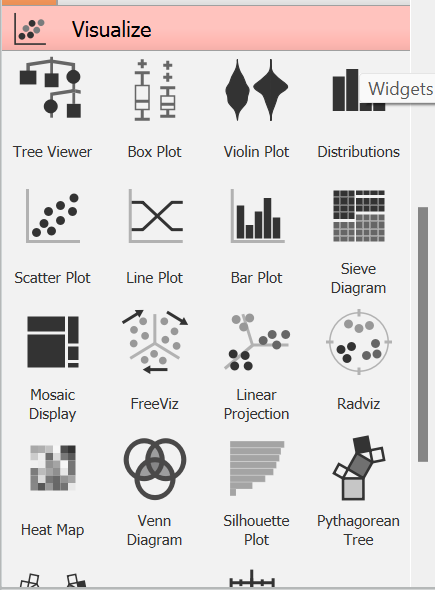
**4. Các tính năng của Orange**

* Chức năng Data:dùng để nhập dữ liệu, thay đổi cũng như trích xuất dữ liệu.



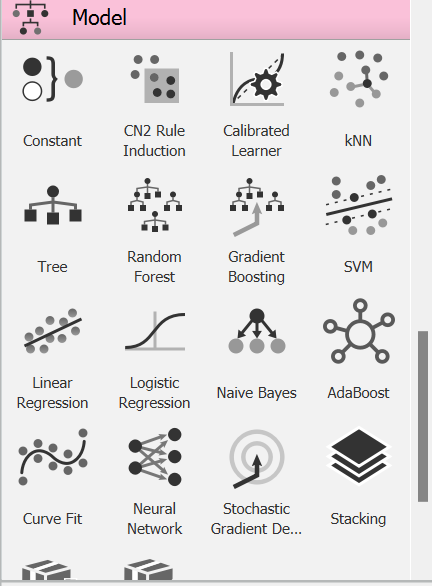
**Hình 2. 6. Chức năng của Data**

* Chức năng Visualize: Dùng để thể hiện dữ liệu bằng các biểu đồ (chart) cho thuận tiện quan sát dữ liệu.



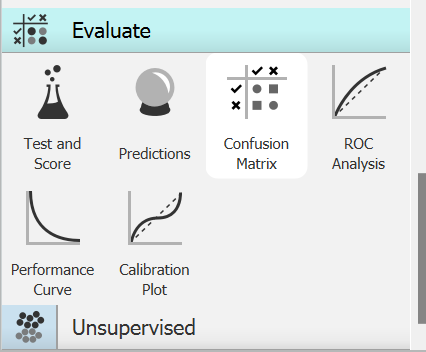
**Hình 2. 7. Chức năng của Visualize**

* Chức năng Model: Bao gồm phương pháp quan trọng trong mô hình phân lớp như SVM, logistic regression, Tree,... và các hàm học máy (machine learning).



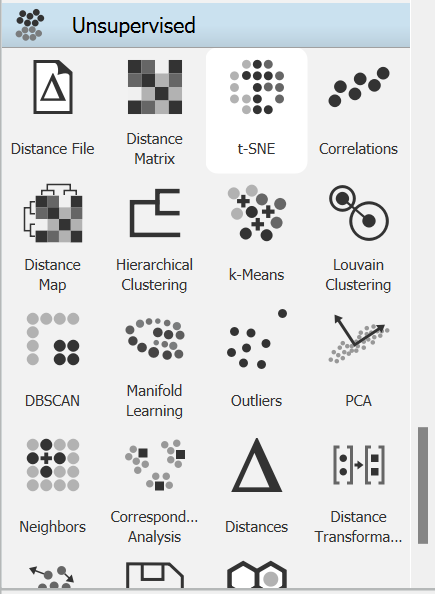
**Hình 2. 8. Chức năng của Model**

* Chức năng Evaluate: Tổng hợp các cách kiểm định mô hình như: Confusion, Test & Score, Prediction,..



**Hình 2. 9. Chức năng của Evaluate**

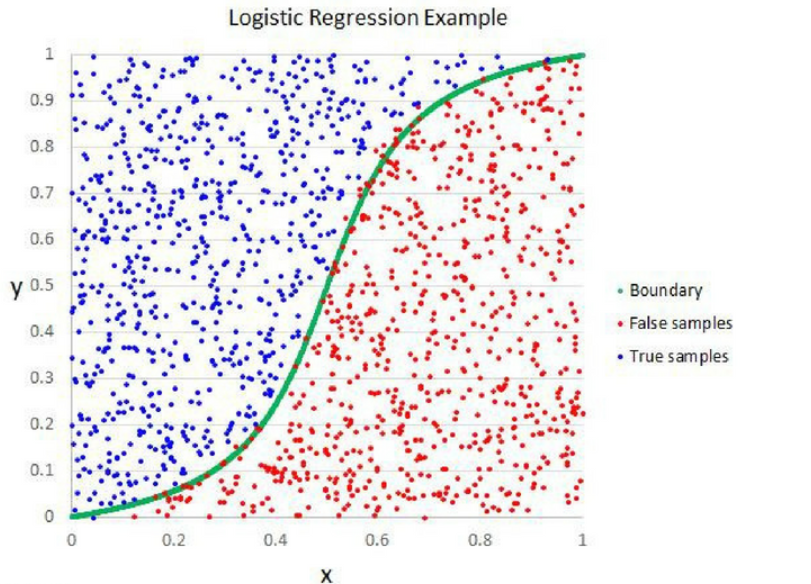
* Chức năng Unsupervised: Gồm các hàm học máy(Machinelearning) gom nhóm dữ liệu như: Distance, K-means, …



**Hình 2. 10. Chức năng của Unsuperived**

**5. Các mô hình phân lớp dữ liệu**

5.1 Mô hình Logistic Regression (hồi quy logistic):



**Hình 2. 11. Mô hình Logistic Regression**

Thuật toán phân loại hồi quy logistic với chính quy hóa LASSO (LI) hoặc sườn núi (L.2).

Đầu vào.

- Dữ liệu: dữ liệu đầu vào.

- Tiền xử lý: phương pháp tiền xử lý.

Đầu ra

- Học viện: thuật toán học hồi quy logistic.

- Mô hình: mô hình được đào tạo Các hệ số: hệ số hồi quy logistic.”

5.2 Mô hình Decision tree (cây quyết định):



**Hình 2. 12. Mô hình Decision tree**

1. Name: Người học có thể được đặt tên theo ý muốn. Tên mặc định là Tree

2. Thông số:

- Induce binary tree: xây dựng cây nhị phân (chia thành hai nút con)

-Min. number of instances in leaves (số tối thiểu các ví dụ lá): nếu được chọn, thuật toán sẽ không bao giờ đặt số nút ít hơn số dữ liệu tham khảo”

- Do not split subsets smaller than (Không phân chia các tập hợp nhỏ hơn): cấm thuật toán phân chia các nút có ít hơn số lượng ví dụ đã cho.”

-Limit the maximal tree depth (Giới hạn độ sâu cây tối đa): giới hạn độ sâu của cây phân loại ở số cấp nút được chỉ định”

3.Tạo một báo cáo.

4. Nhấp vào Apply sau khi thay đổi. Nếu chúng ta đánh dấu vào ô bên trái nút Apply, các thay đổi sẽ được tự động.

5.3 Mô hình Support Vector Machine



**Hình 2. 13. Mô hình Support Vector Machine**

Đầu vào

- Dữ liệu: dữ liệu đầu vào

- Tiền xử lý: phương pháp tiền xử lý

Đầu ra

- Learner: thuật toán học hồi quy tuyến tính”

- Mô hình: mô hình được đào tạo Vecto hỗ trợ: thể hiện được sử dụng làm vectơ hỗ trợ.

- Máy vectơ hỗ trợ (SVM) là một kỹ thuật máy học phân tách không gian thuộc tính với một siêu phẳng, do đó tối đa hóa các điểm dữ liệu của tất cả các lớp. Kỹ thuật này thường mang lại kết quả dự đoán tối cao.

- Đối với hồi quy, SVM thực hiện hồi quy tuyến tính trong không gian tính năng với kích thước cao bằng cách sử dụng ε-insensitive. Độ chính xác ước tính của nó phụ thuộc tốt vào các tham số C, ε và kernel.

- Hoạt động cho cả nhiệm vụ phân loại và hồi quy.”

5.4 Mô hình Neural Network

Một thuật toán perceptron (MLP) nhiều lớp với backpropagation.

Đầu vào

- Dữ liệu: tập dữ liệu đầu vào

- Tiền xử lý: phương pháp tiền xử lý

Kết quả

- Người học: thuật toán học perception nhiều lớp

- Mô hình: mô hình được đào tạo.

Tiện ích Neural Network sử dụng thuật toán Perceptron nhiều lớp của sklearn có thể học các mô hình phi tuyến tính cũng như tuyến tính.

**6. Quy trình phân lớp dữ liệu**

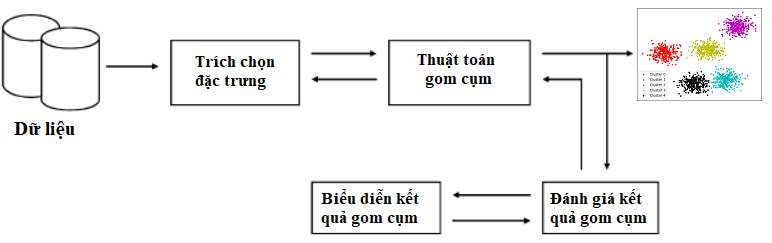
6.1 Phân cụm dữ liệu

Phân cụm dữ liệu là quá trình gom cụm/nhóm các đối tượng/dữ liệu có các đặc điểm tương đồng vào những cụm hoặc nhóm tương ứng.

Trong đó:

- Các đối tượng có các tính chất tương đồng thuộc chung một cụm.

- Các đối tượng có các tính chất khác nhau thuộc cụm hoặc nhóm khác nhau.



**Hình 2. 14. Phân cụm dữ liệu**

6.2 Phân cụm phân cấp (Hierarchical Clustering) bằng phương pháp Diana

*a) Phân cụm phân cấp (Hierarchical clustering)*

Xây dựng một cây phân cấp cho dữ liệu cần gom cụm dựa trên:

- Ma trận khoảng cách giữa các phần tử (similarity matrix hoặc dissimilarity matrix).

* Similarity matrix là số đo mức giống nhau của hai đối tượng.
* Dissimilarity matrix là số đo mức độ khác nhau của hai đối tượng.

- Độ đo khoảng cách giữa các cụm (single link, complete link…)

Phương pháp này không cần xác định trước số cụm nhưng cần xác định điều kiện dừng.

Các phương pháp điển hình: Agnes, Diana…

*b) Phương pháp phân cụm phân cấp Diana*

Theo chiến lược top down:

- Bắt đầu với 1 cụm gồm tất cả phần tử

- Ở mỗi bước, chia cụm ban đầu thành 2 cụm: Khoảng cách giữa 2 cụm là khoảng cách giữa 2 điểm gần nhất từ hai cụm, hoặc khoảng cách trung bình.

- Thực hiện đệ quy trên các cụm mới được tách ra và lặp lại cho đến khi mỗi phần tử là 1 cụm.

- Kết quả phát sinh cây phân cấp (dendrogram)

6.3 Phân cụm phân cấp bằng phương pháp Diana và kết quả

*a. Dùng phần mềm Orange để phân cụm phân cấp bằng phương pháp Diana*

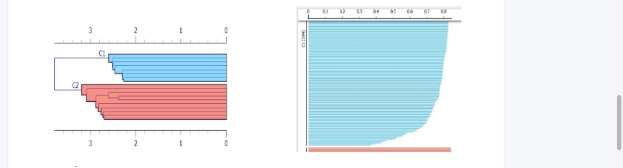


**Hình 2. 15. Phân cụm phân cấp bằng phương pháp Diana**

- File input: Sử dụng dữ liệu cho dự án, bỏ qua (skip) biến mục tiêu và subscriber

- Distances: Sử dụng độ đo khoảng cách Euclidean (Euclid=Minkowski (n=2))

- Hierarchical Clustering: Sử dụng phương pháp tính khoảng cách Single-link và chia làm 2 cụm để được chỉ số Sihouette Cluster dương nhiều nhất



**Hình 2. 16. Kết quả phân cụm bằng Diana**

*b. Kết quả :* Số cụm tốt nhất là 2 (vì chia làm 2 cụm để được chỉ số Sihouette Cluster dương nhiều nhất).

6.4 Phân cụm phân hoạch bằng thuật toán K-means và kết quả

*a) Dùng phần mềm Orange để phân cụm phân hoạch bằng thuật toán K-means và*

*chọn số cụm tốt nhất:*



**Hình 2. 17. Phân cụm phân hoạch bằng thuật toán K-means**

- File input: Bộ dữ liệu sử dụng cho dự án nhưng bỏ qua (skip) biến mục tiêu là subscribed.

- K-means: chạy phân cụm từ 2-5 cụm, so sánh Sihouette Scores được tính ra, chọn chia thành 3 cụm (vì 3 là số cụm có Sihouette Scores cao nhất).

*b)Kết quả:*

Số cụm tốt nhất là 3 (vì 3 là số cụm có Sihouette Scores cao nhất)

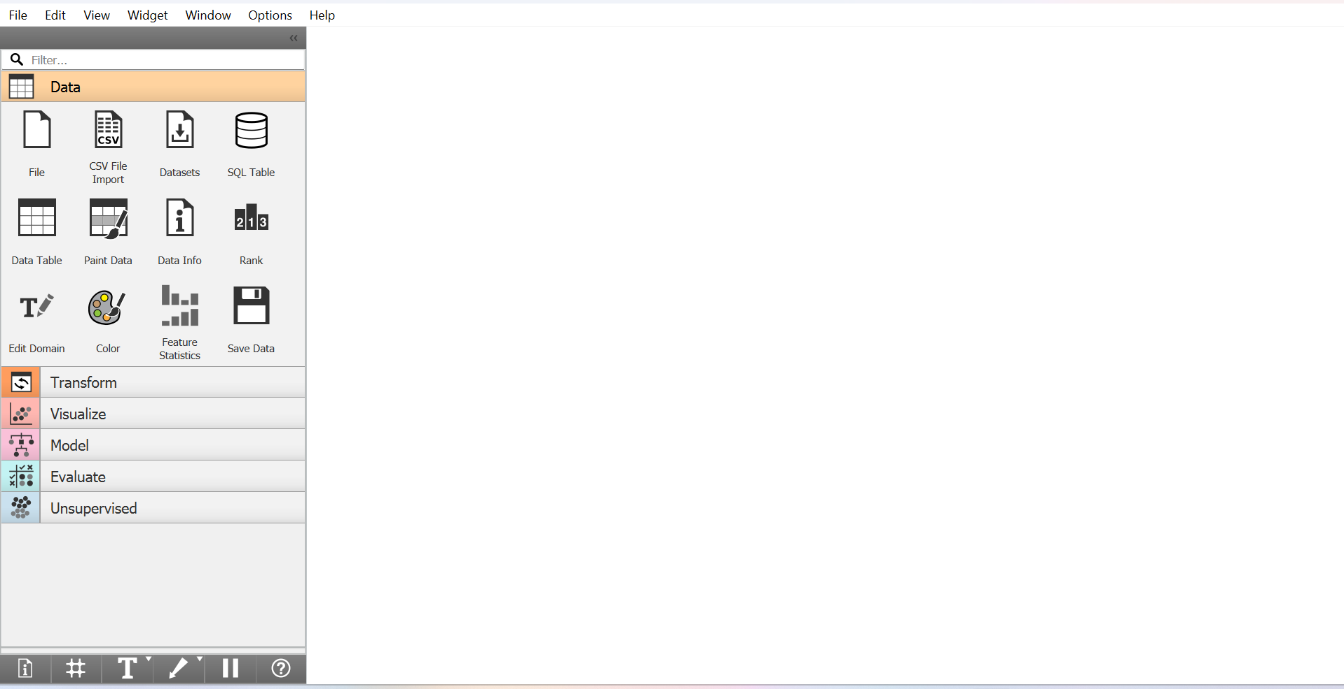
**CHƯƠNG 3: HIỆN THỰC HÓA NGHIÊN CỨU**

1. **Giao diện chính của phần mềm**

Logo phần mềm Orange:



Sau khi nháy đúp chuột vào biểu tượng thì chúng ta sẽ thấy giao diện chính của Orange như sau:



**Hình 3. 1. Giao diện chính của Orange**

1. **Thực hiện trực quan hóa**

Bước 1: Trước hết chúng ta cần phải có một file dữ liệu, Orange hỗ trợ trực quan hóa dữ liệu từ nhiều loại file khác nhau, bao gồm:

* **File csv** (comma-separated values): Đây là loại file dữ liệu phổ biến nhất, được định dạng bằng các dấu phẩy để phân tách các giá trị dữ liệu.
* **File xlsx** (Excel): Đây là loại file dữ liệu được tạo bởi Microsoft Excel.
* **File txt** (text): Đây là loại file văn bản thuần túy, không có định dạng đặc biệt.
* **File json** (JavaScript Object Notation): Đây là loại file dữ liệu được định dạng bằng JSON.
* **File yaml** (YAML Ain't Markup Language): Đây là loại file dữ liệu được định dạng bằng YAML.

Để mở một file dữ liệu bằng Orange, bạn có thể chọn **File > Open** và chọn file dữ liệu cần mở.

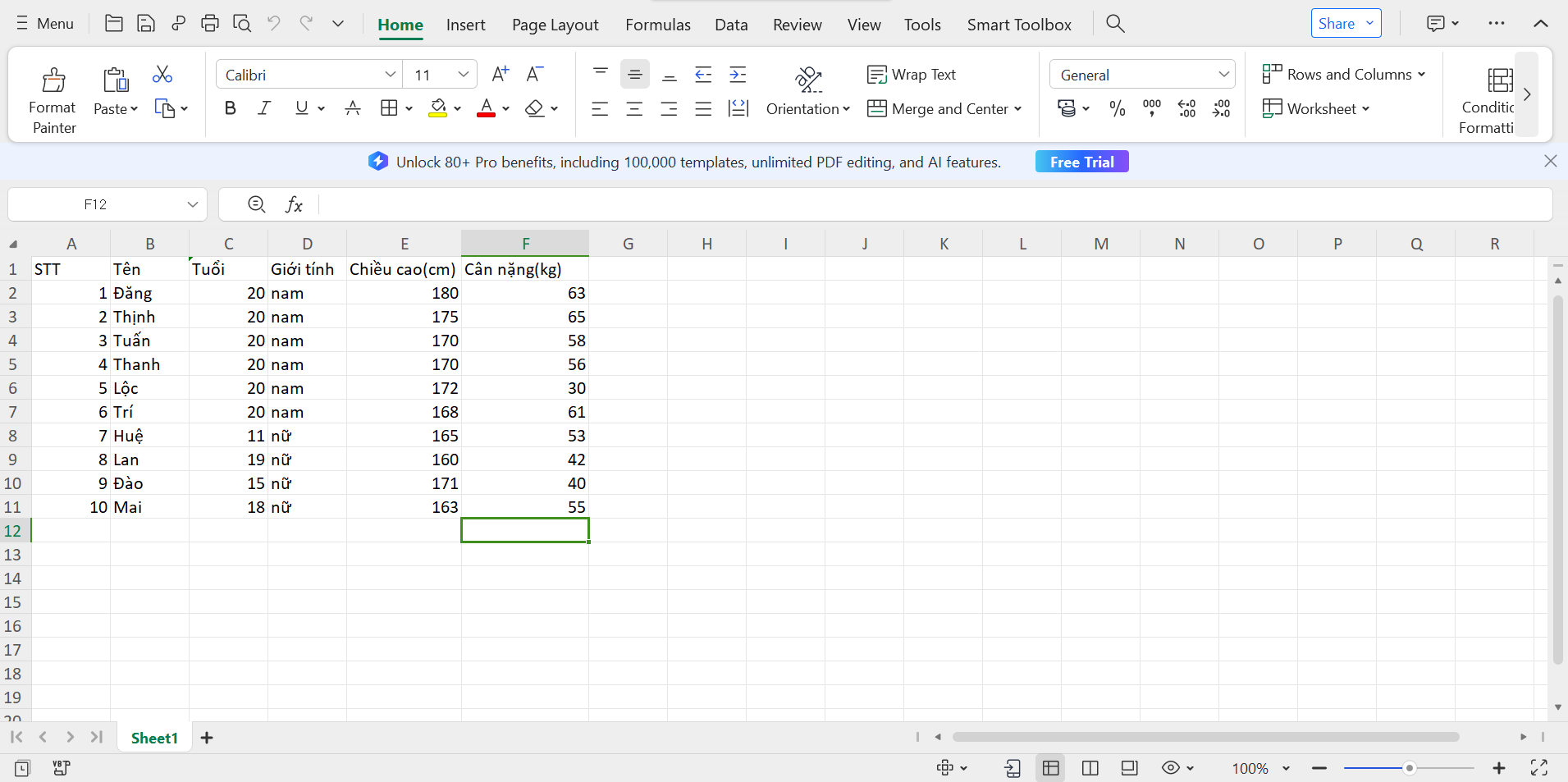
Orange cũng hỗ trợ nhập dữ liệu từ các nguồn khác nhau, chẳng hạn như:

* **Các cơ sở dữ liệu**, chẳng hạn như MySQL, PostgreSQL, v.v.
* **Các API**, chẳng hạn như API của Google Analytics, v.v.
* **Các dịch vụ web**, chẳng hạn như dịch vụ web của Google Maps, v.v.

Để nhập dữ liệu từ một nguồn khác, bạn có thể chọn **File > Import** và chọn nguồn dữ liệu cần nhập.

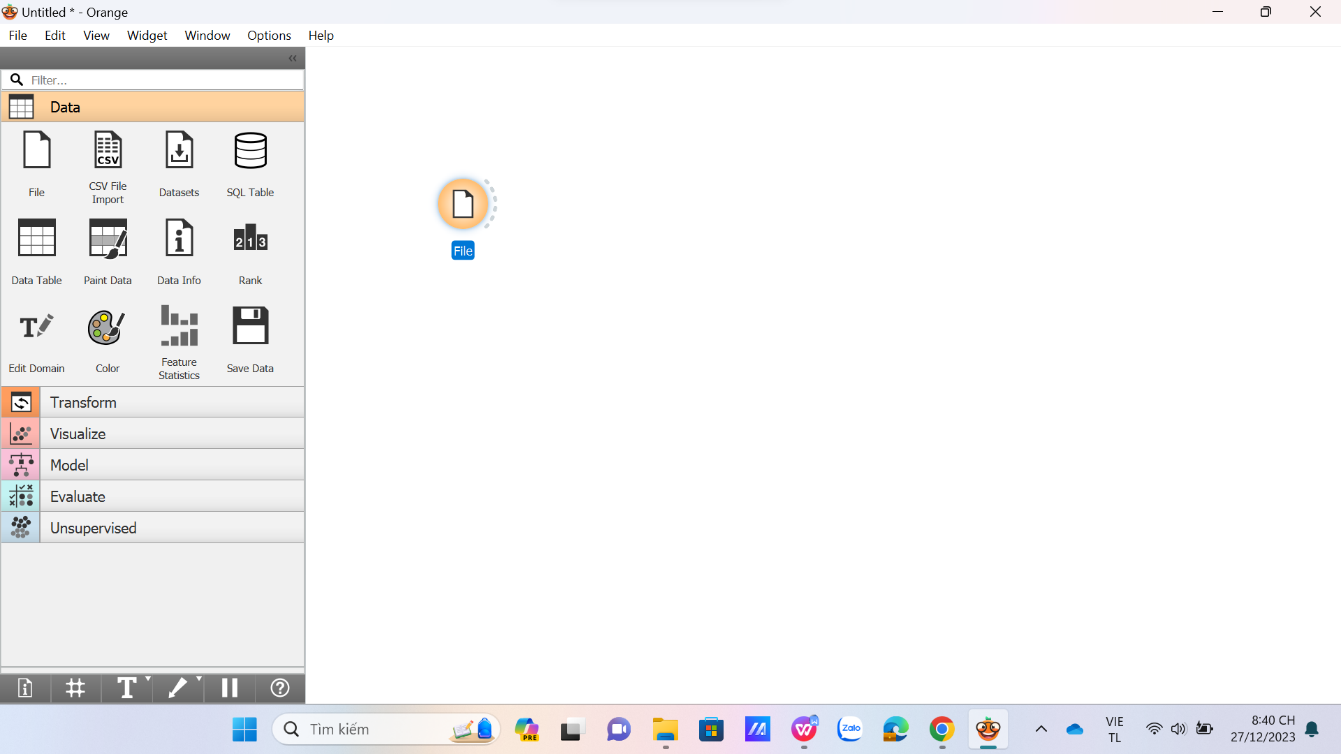
Sau khi mở hoặc nhập dữ liệu, bạn có thể sử dụng Orange để tạo các biểu đồ và đồ thị trực quan hóa dữ liệu.

Ví dụ như có file **xlsx có nội dung như thế này:**



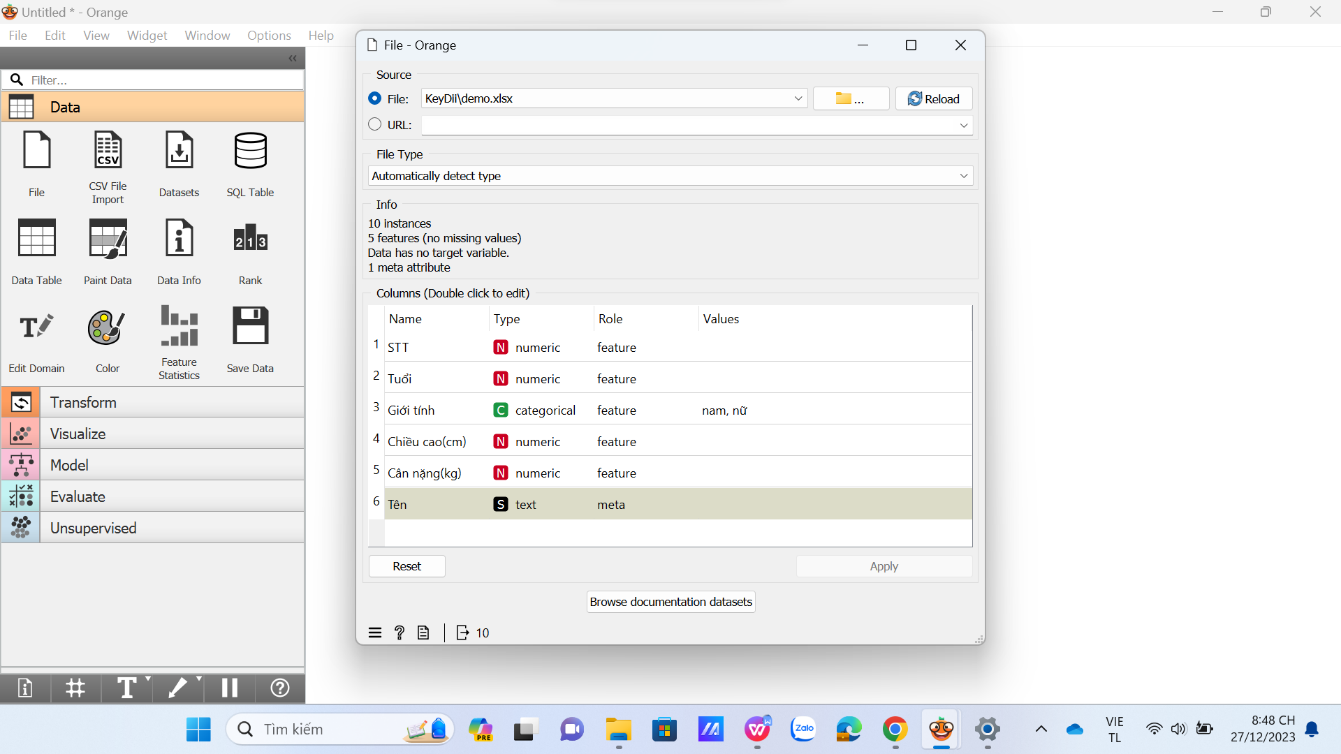
**Hình 3. 2. File Excel**

Bước 2: Tiếp theo sẽ mở tab Data và chọn lệnh File. Lệnh file sự tự động hiện lên trên vùng trống bên phải thanh công cụ như hình:



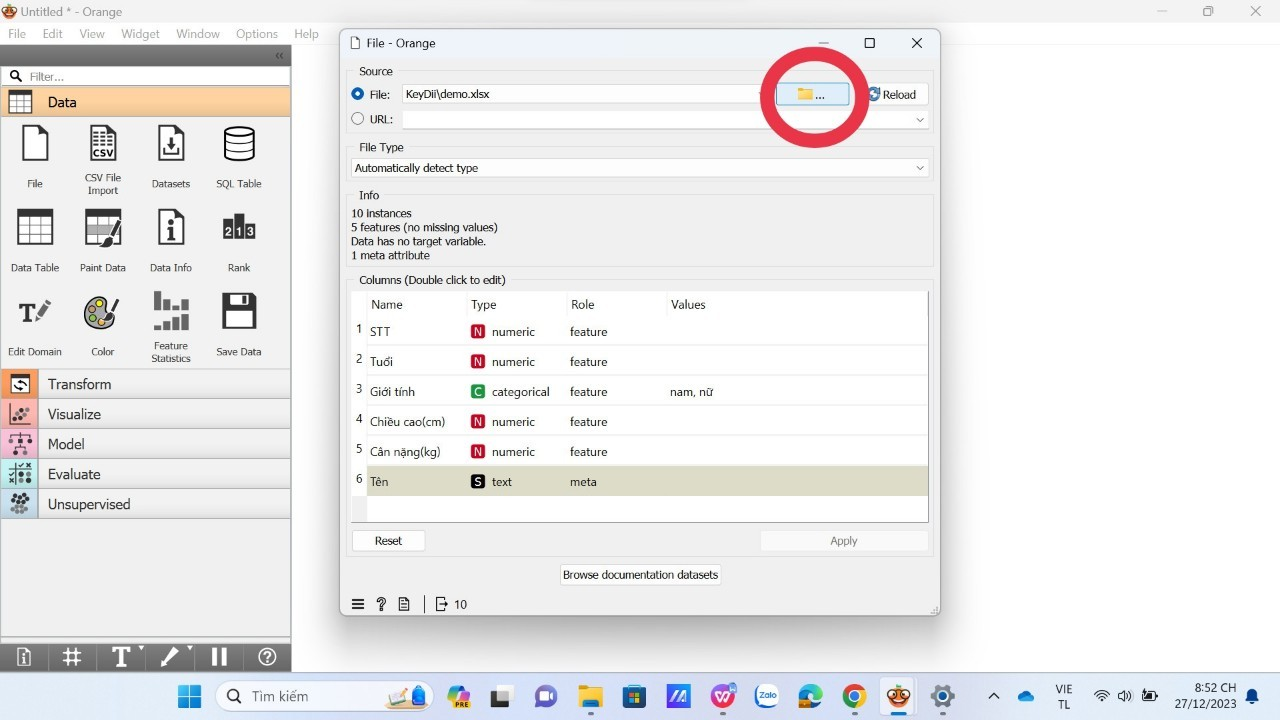
**Hình 3. 3. Lệnh file**

Bước 3: Nháy đúp chuột vào lệnh file bên vùng trống sẽ xuất hiện hộp thoại file:



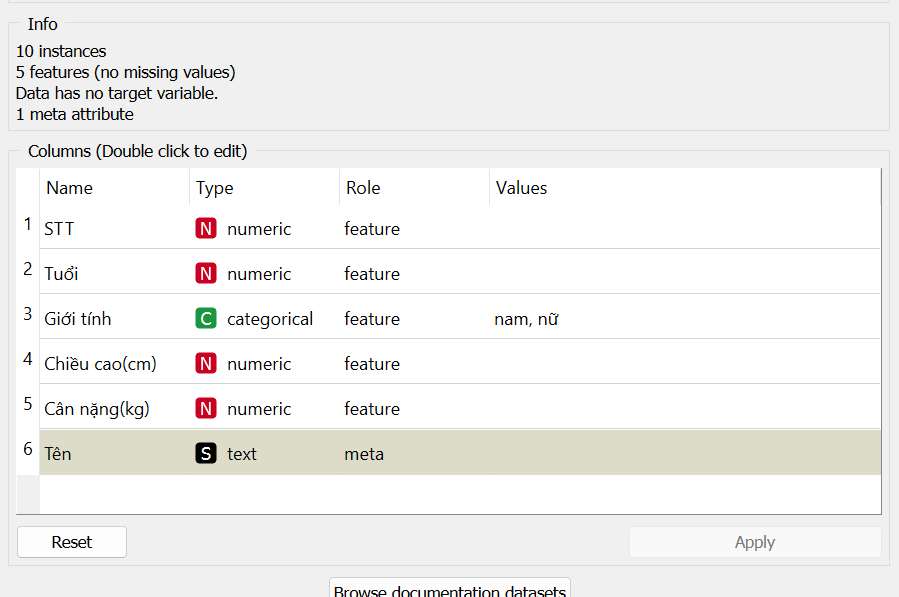
**Hình 3. 4. Hộp thoại file**

Bước 4: Chúng ta sẽ chọn file thích hợp để trực quan hóa dữ liệu tại đây:



**Hình 3. 5. Chọn file**

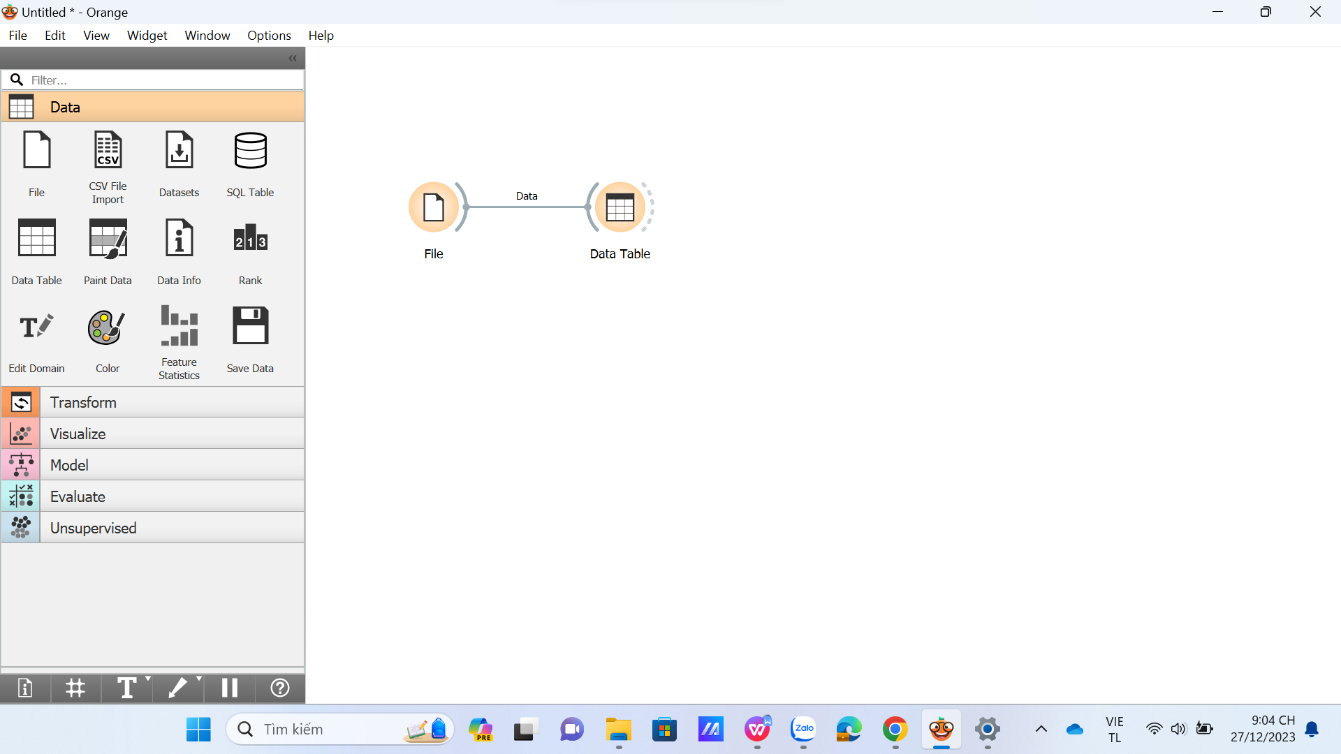
Số cột trong file đã chọn sẽ thể hiện đầy đủ phía dưới:



**Hình 3. 6. Hiện cột của file**

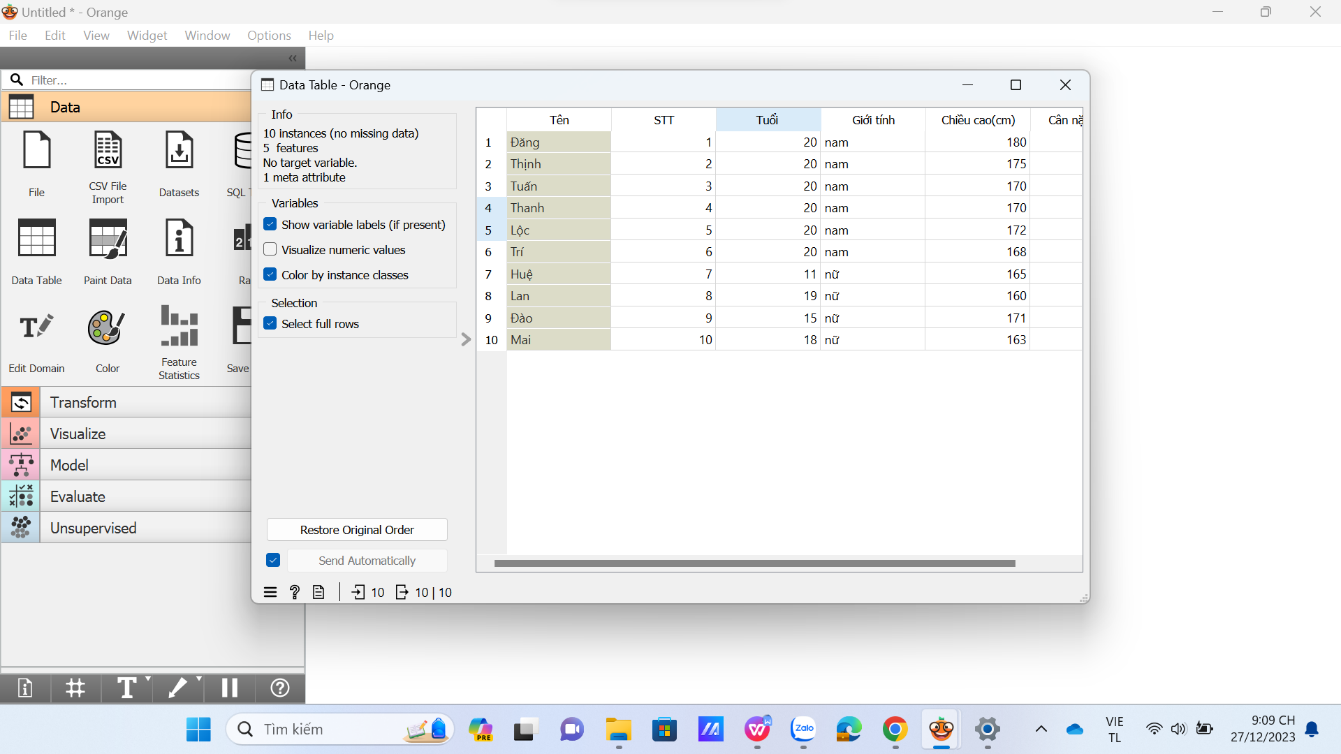
Tiếp theo ta sẽ đóng hộp thoại lại, chọn X bên phải phía trên hợp thoại

Bước 5: Vào tab Data chọn Data Table liên kết lệnh file lại với Data Table bằng cách nắm giữ chuột từ lệnh file kéo tới lệnh Data Table:



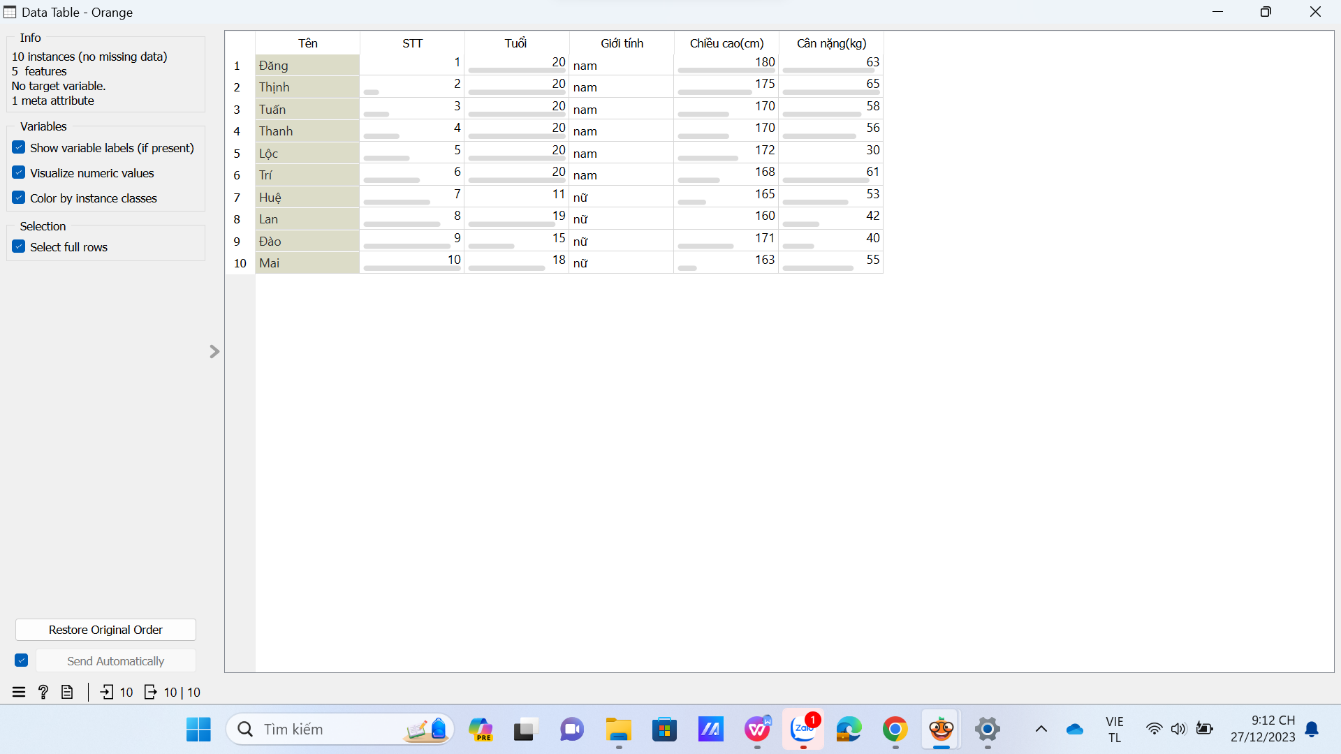
**Hình 3. 7 Liên kết file với Data Table**

Nháy đúp vào Data Table để xem chi tiết các cột



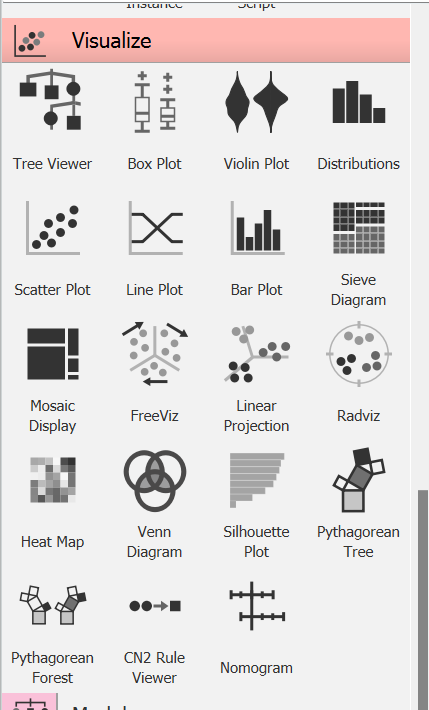
**Hình 3. 8. Mở Data Table**

Nếu muốn trực quan hóa các giá trị số thì chúng ta sẽ chọn vào Visualize numeric values:



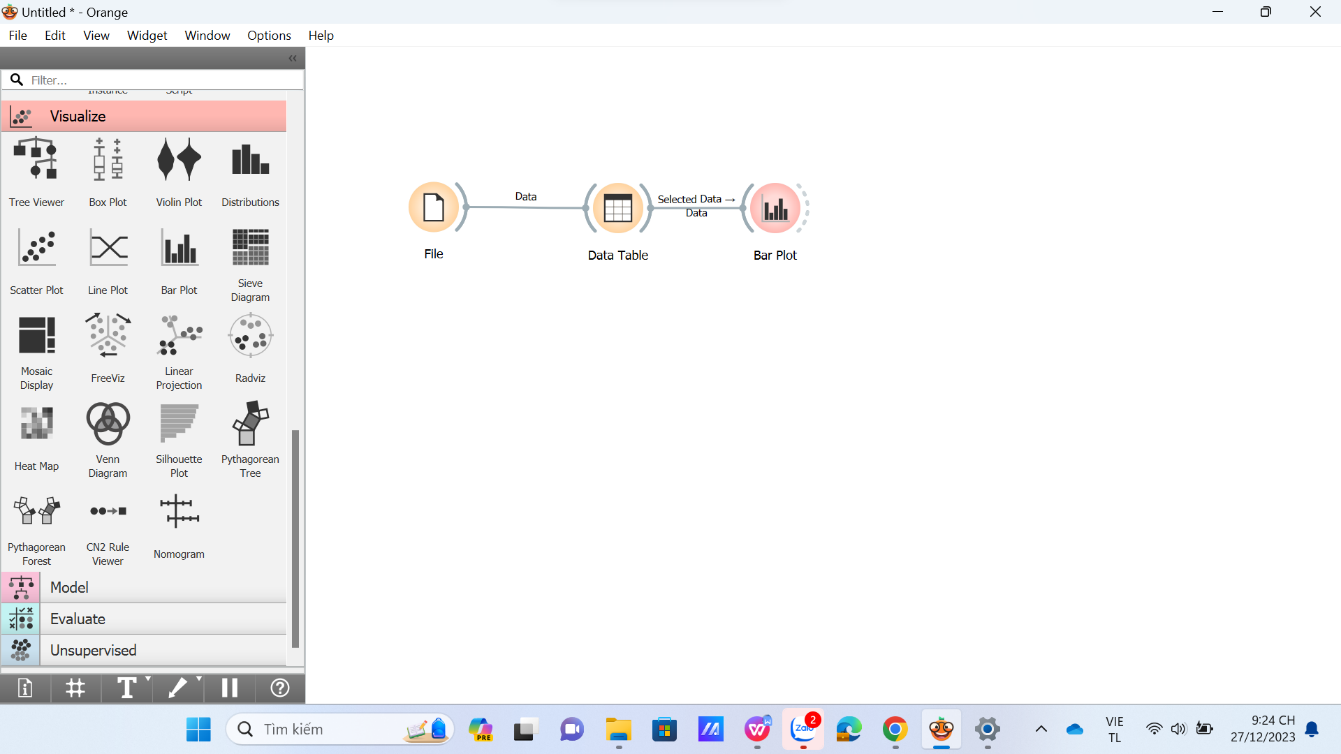
**Hình 3. 9. Trực quan hóa các giá trị số**

Bước 6: Tiếp theo chúng ta chúng trực quan dữ liệu bằng biểu đồ nào thì hãy chọn tab Visualize:



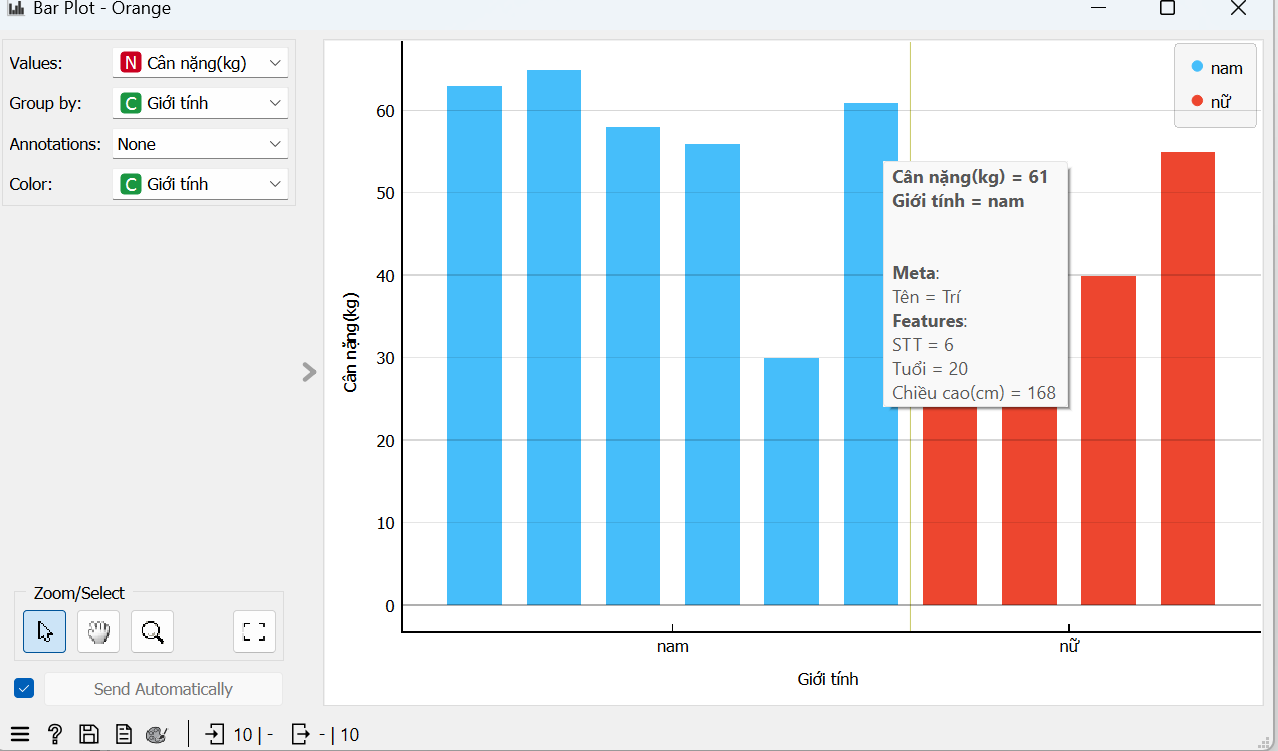
**Hình 3. 10. Thanh công cụ Visualize**

Chẳng hạn chúng ta chọn biểu đồ cột thì nháy vào lệnh Bar Plot sau đó nối từ Data Table với Bar Plot:



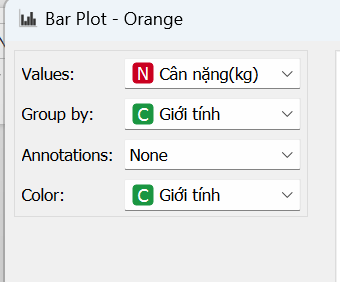
**Hình 3. 11. Kếtt nối Bar Plot**

Nháy đúp vào Bar Plot bên vùng trống sẽ thấy được biểu đồ trực quan hóa:



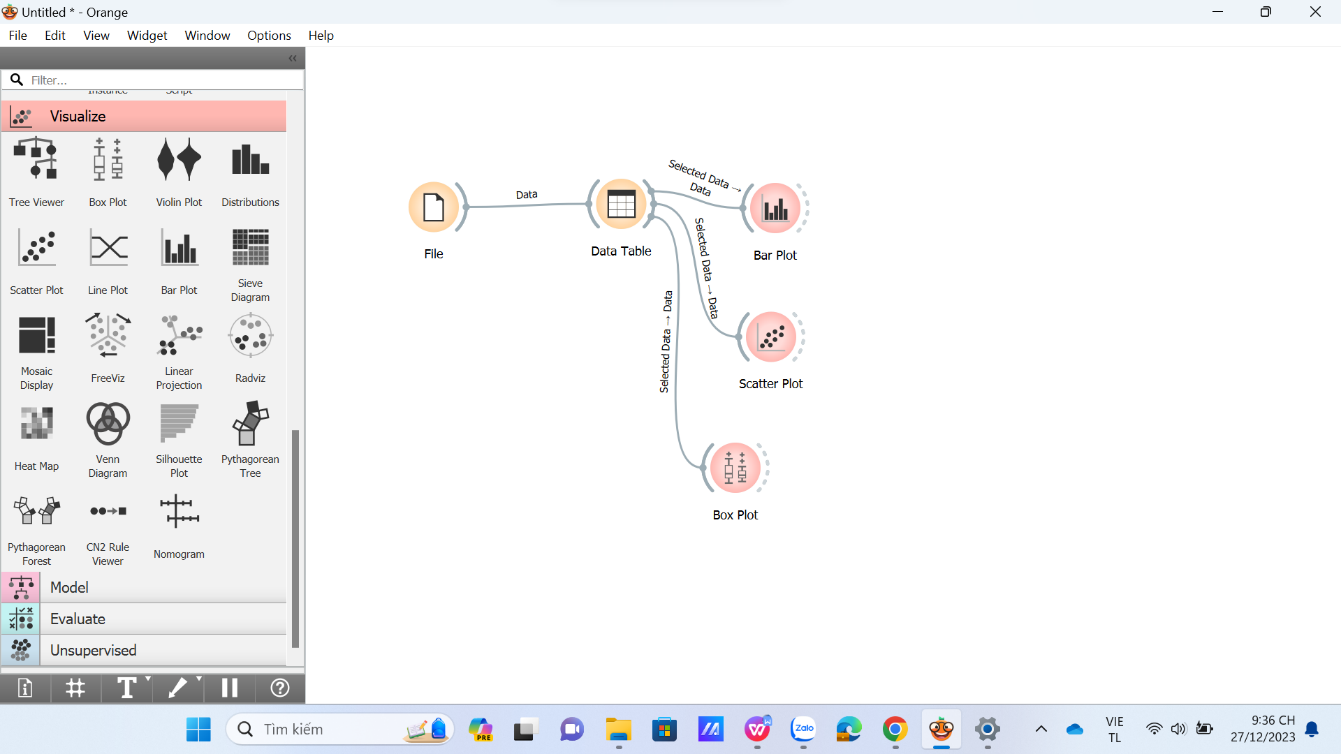
**Hình 3. 12. Biểu đồ được trực quan hóa**

Chúng ta có thể thay đổi các thuộc tính gốc trái phía trên



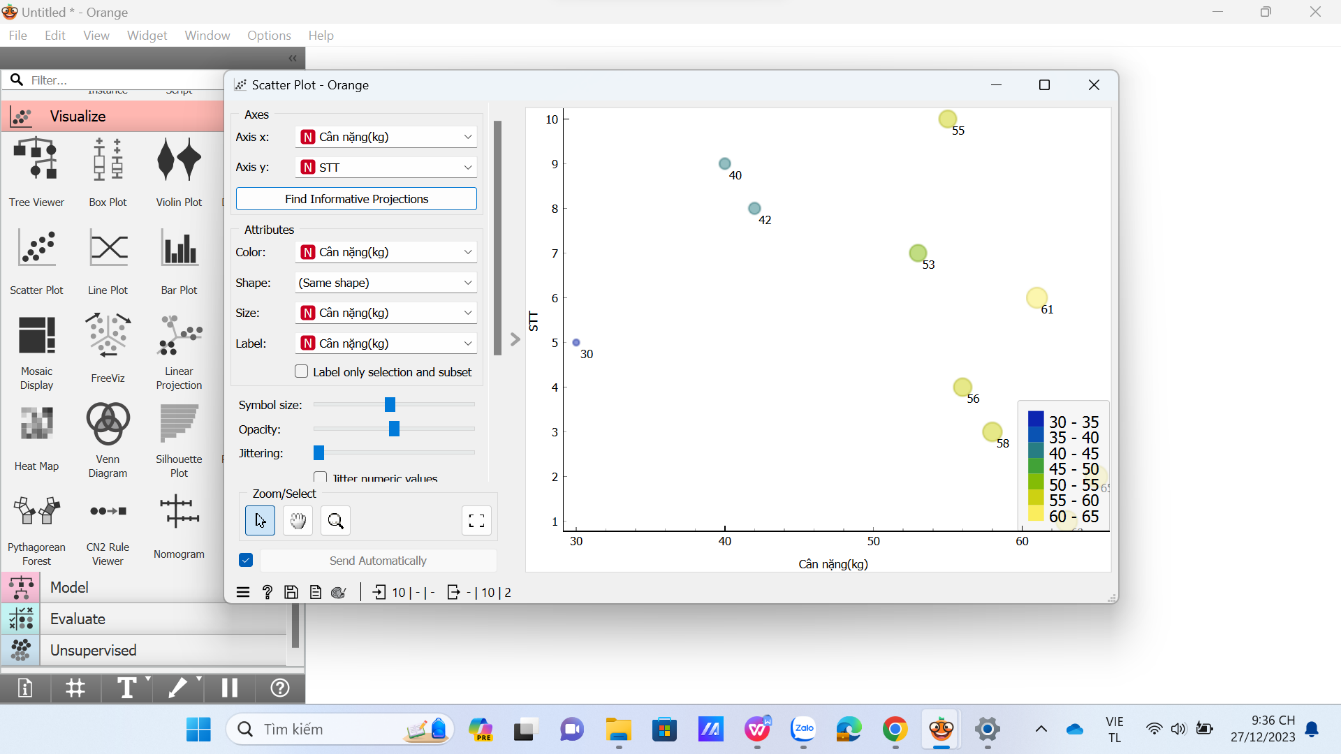
**Hình 3. 13. Thay đổi thuộc tính**

Tương tự như vậy chúng ta sẽ tạo ra những biểu đồ khác nhau:



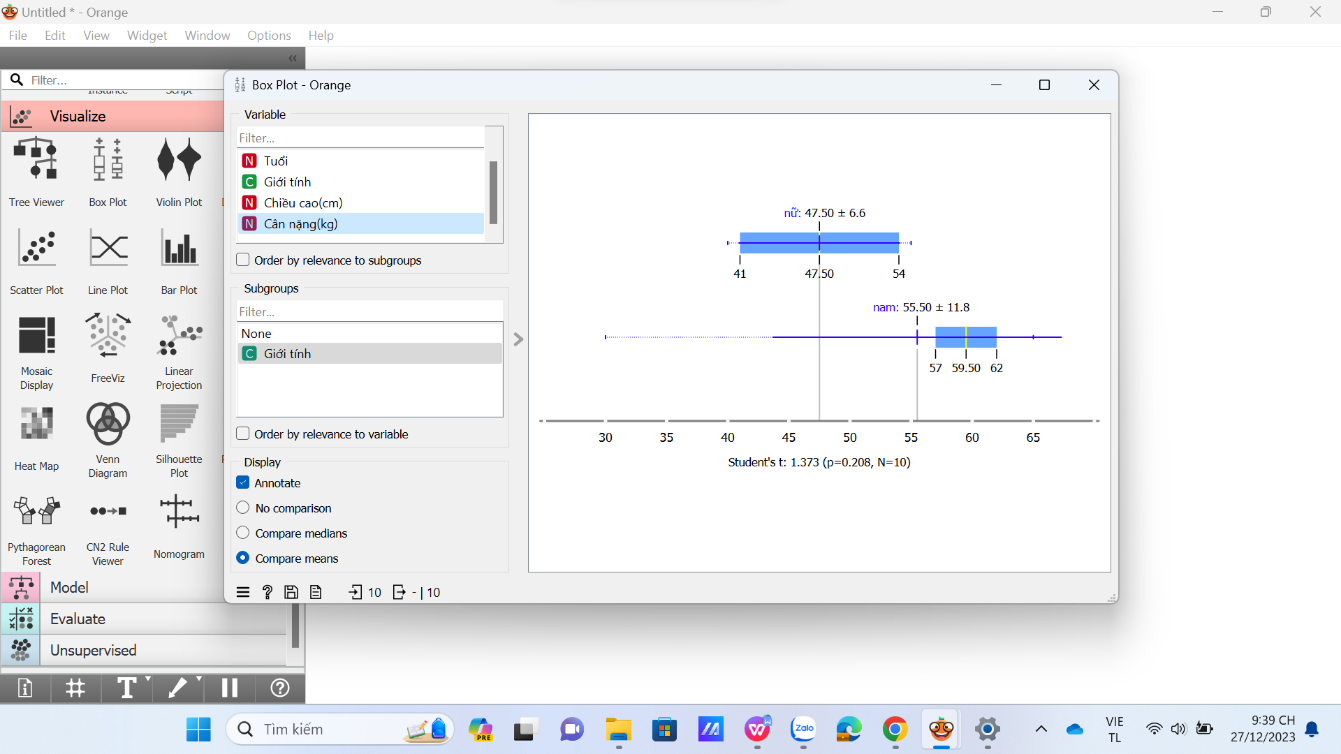
**Hình 3. 14. Kết nối nhiều lại biểu đồ**

Scatter Plot:



**Hình 3. 15. Scatter Plot**

Box Plot:



**Hình 3. 16. Box Plot**

**CHƯƠNG 4: KẾT QUẢ NGHIÊN CỨU**

**Orange đã được sử dụng để phát triển các công cụ trực quan hóa dữ liệu cho nhiều ứng dụng khác nhau: Trực quan hóa dữ liệu thời gian thực, trực quan hóa dữ liệu không gian, trực quan hóa dữ liệu đa chiều, trực quan hóa dữ liệu phi định lượng:**

**Orange đã được sử dụng để cải thiện hiệu quả của việc trực quan hóa dữ liệu bằng cách sử dụng các kỹ thuật như: phân tích dữ liệu, học máy:**

**Orange đã được sử dụng để tăng cường khả năng tương tác của việc trực quan hóa dữ liệu bằng cách sử dụng các công nghệ như: Web, cảm ứng**

Nhìn chung, nghiên cứu trực quan hóa dữ liệu bằng Orange đã đạt được những kết quả đáng kể, góp phần thúc đẩy sự phát triển của lĩnh vực trực quan hóa dữ liệu.

**CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

1. **Kết Luận**

Trực quan hóa dữ liệu là một kỹ thuật quan trọng trong việc phân tích và hiểu dữ liệu. Nó cho phép chúng ta biến dữ liệu thành các hình ảnh và biểu đồ trực quan, giúp chúng ta dễ dàng nắm bắt các xu hướng, mối quan hệ và thông tin quan trọng khác trong dữ liệu.

Tuy nhiên, cũng cần lưu ý rằng trực quan hóa dữ liệu chỉ là một công cụ, và nó không thể thay thế cho phân tích dữ liệu kỹ lưỡng. Chúng ta cần hiểu rõ về dữ liệu và các kỹ thuật trực quan hóa dữ liệu để có thể sử dụng trực quan hóa dữ liệu một cách hiệu quả.

1. **Hướng phát triển**

Trực quan hóa dữ liệu bằng Orange là một lĩnh vực nghiên cứu đang phát triển nhanh chóng. Dưới đây là một số hướng phát triển tiềm năng của lĩnh vực này:

+ Phát triển các công cụ và kĩ thuật trực quan hóa dữ liệu mới.

+ Cải thiện hiệu quả của công việc trực quan hóa.

+ Tăng cường khả năng tương tác của việc trực quan hóa dữ liệu.

**DANH MỤC TÀI LIỆU THAM KHẢO**

[**https://websitehcm.com/data-mining-su-dung-orange/**](https://websitehcm.com/data-mining-su-dung-orange/)

[**https://www.studocu.com/vn/document/truong-dai-hoc-kinh-te-dai-hoc-da-nang/quan-tri-hoc/tong-quan-ve-orange-trong-khai-pha-du-lieu/18445899**](https://www.studocu.com/vn/document/truong-dai-hoc-kinh-te-dai-hoc-da-nang/quan-tri-hoc/tong-quan-ve-orange-trong-khai-pha-du-lieu/18445899)

[**https://ongxuanhong.wordpress.com/2018/02/03/voc-thu-orange-phan-mem-data-mining/**](https://ongxuanhong.wordpress.com/2018/02/03/voc-thu-orange-phan-mem-data-mining/)

[**https://fr.scribd.com/document/627667751/Ti%E1%BB%83u-lu%E1%BA%ADn-khoa-h%E1%BB%8Dc-d%E1%BB%AF-li%E1%BB%87u**](https://fr.scribd.com/document/627667751/Ti%E1%BB%83u-lu%E1%BA%ADn-khoa-h%E1%BB%8Dc-d%E1%BB%AF-li%E1%BB%87u)

**PHỤ LỤC**