# **Yêu cầu với BTL NHẬP MÔN KHAI PHÁ DỮ LIỆU VÀ MÁY HỌC**

1. Tên đề tài có dạng  
   “**Ứng dụng khai phá dữ liệu và máy học tích hợp vào hệ thống <tên hệ thống> [giải quyết bài toán <tên bài toán>]**”
2. Sử dụng dữ liệu có nguồn gốc rõ ràng, tin cậy, đủ lớn và có liên quan đến bài toán định xử lý; được cung cấp sẵn trên các cộng đồng DMML hoặc SV tự thu thập.
3. Áp dụng Khai phá luật kết hợp (1) hoặc Phân lớp (2) hoặc Phân cụm (3) để xây dựng giải pháp ứng dụng cho bài toán cụ thể
4. Mô hình được lựa chọn sau huấn luyện (/khai phá) phải căn cứ trên kết quả đánh giá tin cậy, chính xác và có căn cứ khoa học.
5. Mô hình cần được tích hợp vào hệ thống cụ thể, ở chức năng liên quan tiêu biểu để chứng minh tính ứng dụng - giải quyết bài toán đã xác định (các chức năng khác không phải là mục tiêu của học phần này)
6. Đáp ứng đúng mục tiêu theo đề cương của học phần NHẬP MÔN KHAI PHÁ DỮ LIỆU VÀ MÁY HỌC
7. Từng SV cần thực hiện nhiệm vụ để chứng minh đạt được Chuẩn đầu ra của học phần.

## 

***Báo cáo BTL học phần này cần trình bày theo cấu trúc dưới đây (SV soạn thảo***

***trực tiếp vào file này), tuân thủ Quy định trình bày của khoa***

# 

# **Tên đề tài:** Ứng dụng khai phá dữ liệu và máy học vào tích hợp hệ thống dự đoán nhu cầu sản phẩm của khách hàng tại các cửa hàng thời trang bán lẻ

## **Chương 1. Tổng quan về đề tài**

### **Giới thiệu bài toán**

Bài toán dự đoán nhu cầu sản phẩm trong tương lai trong một cửa hàng thời trang bán lẻ trực tuyến là một yếu tố quan trọng đối với quản lý tồn kho, kế hoạch sản xuất và vận chuyển, cũng như chiến lược tiếp thị và quảng cáo của doanh nghiệp.

Vấn đề chính mà bài toán này đặt ra là làm thế nào để dự đoán một cách chính xác và hiệu quả nhu cầu của khách hàng đối với các sản phẩm trong tương lai, từ đó giúp cửa hàng tối ưu hóa tồn kho, cải thiện trải nghiệm mua sắm và tăng doanh số bán hàng. Điều này đòi hỏi sự kết hợp giữa dữ liệu lịch sử bán hàng, thông tin về sản phẩm và các yếu tố môi trường để đưa ra dự đoán chính xác về nhu cầu sản phẩm.

### **Đánh giá đề tài liên quan (vấn đề hiện đã được giải quyết như thế nào, hiệu quả đến đâu)**

**Dự đoán xu hướng mua hàng:** Bài toán này tập trung vào việc dự đoán xu hướng mua hàng của khách hàng dựa trên dữ liệu lịch sử bán hàng và các yếu tố bên ngoài như mùa vụ, sự kiện đặc biệt, thời tiết. Các phương pháp máy học như học sâu và học tăng cường đã được áp dụng để dự đoán xu hướng mua hàng với độ chính xác cao.

**Nguồn:**[SuNT's Blog | AI in Practical (tiensu.github.io)](https://tiensu.github.io/blog/109_ml_in_demand_prediction/)

**Tối ưu hóa tồn kho:** Bài toán này tập trung vào việc quản lý tồn kho sao cho phù hợp với nhu cầu thực tế của khách hàng và giảm thiểu chi phí lưu kho. Các mô hình dự đoán nhu cầu sản phẩm có thể giúp cải thiện quản lý tồn kho bằng cách dự đoán mức độ cần thiết của từng sản phẩm trong tương lai.

**Nguồn:**[4 Mô hình quản lý hàng tồn kho EOQ, POQ, PDM, ABC (vti-solutions.vn)](https://vti-solutions.vn/mo-hinh-quan-ly-hang-ton-kho/)

**Quản lý chuỗi cung ứng:** Bài toán này liên quan đến việc dự đoán nhu cầu của cả chuỗi cung ứng từ nguồn cung đến điểm bán hàng. Dự đoán chính xác nhu cầu sản phẩm giúp cải thiện quá trình lập kế hoạch sản xuất và vận chuyển trong chuỗi cung ứng.

**Nguồn:**[Mô hình Just In Time: Vai trò, Ứng dụng và Quy trình triển khai (atalink.com)](https://vietnam.atalink.com/blog/tim-hieu-ve-mo-hinh-just-in-time/)

### **Mục đích đề tài**

**Dự đoán nhu cầu về sản phẩm trong tương lai** của khách hàng tại các cửa hàng thời trang bán lẻ nhằm tối ưu hóa tồn kho và điều chỉnh chiến lược kinh doanh hiệu quả trong năm tiếp theo, giúp cửa hàng thời trang bán lẻ hoạt động một cách hiệu quả, linh hoạt và có thể vững vàng hơn trong môi trường kinh doanh ngày càng cạnh tranh.

### **Mục tiêu đề tài**

* Phát triển một mô hình học máy để dự đoán nhu cầu sản phẩm của khách hàng dựa trên dữ liệu lịch sử mua sắm, xu hướng thị trường và hành vi người dùng.
* Tích hợp mô hình vào hệ thống quản lý cửa hàng thời trang bán lẻ để cung cấp thông tin dự đoán về nhu cầu sản phẩm.

### **Phân công và Kế hoạch thực hiện**

| **Nội dung (Khoảng thời gian thực hiện)** | **Công việc chi tiết** | **Thời gian hoàn thành** | **Chu Quỳnh Anh** | **Lê Văn Đạt** | **Nguyễn Thị Thu Hằng** |
| --- | --- | --- | --- | --- | --- |
| Chương 1  (2/4 - 4/4) | - Giới thiệu bài toán  - Đánh giá đề tài liên quan | 3/4/2024 |  | **x** |  |
|  | - Mục đích đề tài | 4/4/2024 | **x** |  |  |
|  | - Mục tiêu đề tài | 3/4/2024 |  |  | **x** |
|  | - Phân công và lên kế hoạch | 3/4/2024 | **x** | **x** | **x** |
| Chương 2  (10/4 - 11/4) | - Tìm Dataset | 10/4/2024 | **x** |  |  |
|  | - Thực hiện phân tích dữ liệu | 11/4/2024 |  |  | **x** |
|  | - Tiến hành các bước tiền xử lý dữ liệu | 11/4/2024 |  | **x** |  |
|  | - Biểu diễn trực quan hóa dữ liệu | 11/4/2024 | **x** |  |  |
|  | - Làm slide báo cáo | 11/4/2024 | **x** |  |  |
|  | - Thuyết trình | 11/4/2024 | **x** | **x** | **x** |
| (11/4 - 16/4) | - Chỉnh sửa bài tập lớn theo các góp ý của giáo viên |  | **x** | **x** | **x** |

#### 

## **Chương 2. Tiền xử lý dữ liệu**

### 2.1 Thực hiện phân tích về dữ liệu

* Sử dụng bộ ClothesDataSet.csv cho đề tài (nguồn Kaggle.com)

(<https://www.kaggle.com/datasets/nishchay331/retail-store/data>)

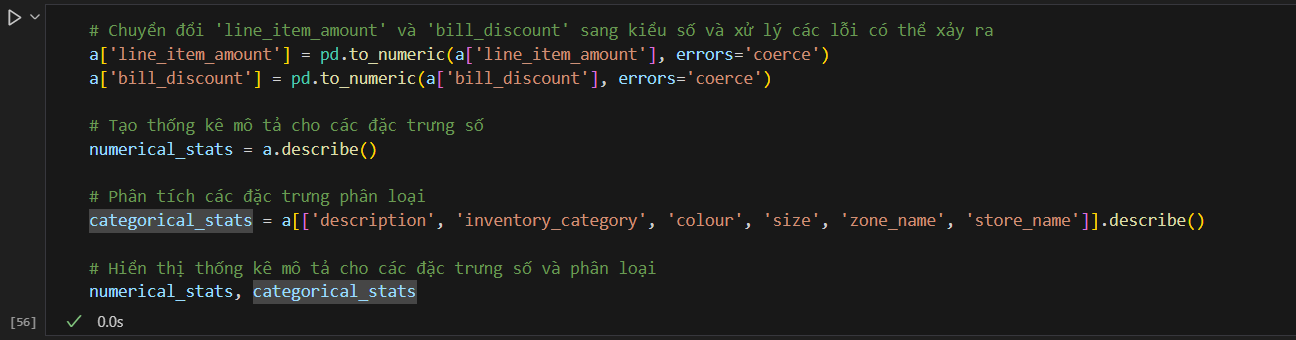
* Tập dữ liệu được lựa chọn gồm 2999 bản ghi lịch sử mua hàng của khách hàng tại một cửa hàng thời trang bán lẻ trong năm 2022, bao gồm 12 cột như sau : user\_id, bill\_id, line\_item\_amount, bill\_discount ,transaction\_date ,description,

inventory\_category ,colour ,size, zone\_name ,store\_name,year

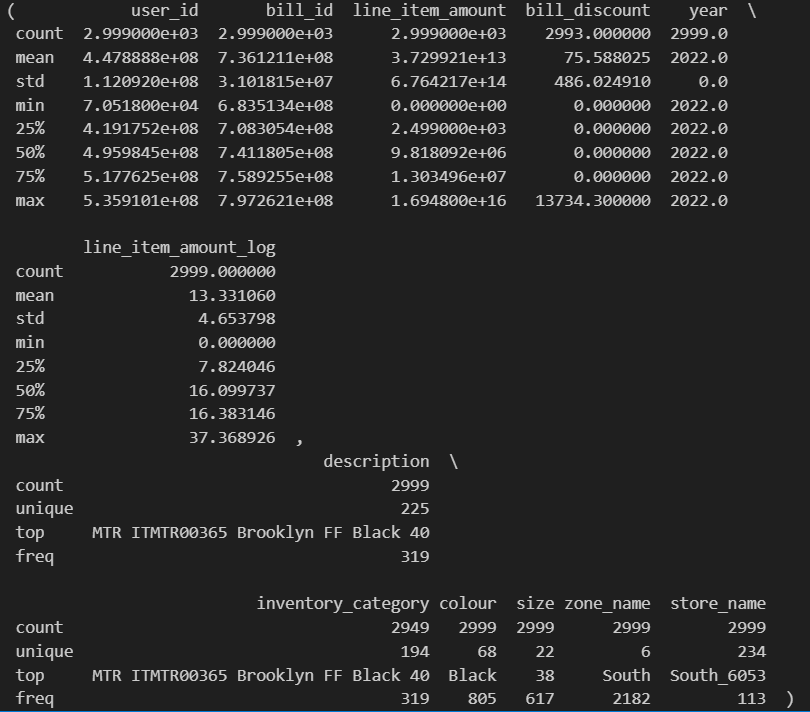
* Nguồn lưu trữ dataset được sử dụng trong đề tài:

<https://drive.google.com/file/d/1tvu5qsEP4cHSQW_s6uuT-Sc4Z3k5-uNH/view?usp=drive_link>

### 2.2 Phân tích đặc trưng:



Kết quả thu được:



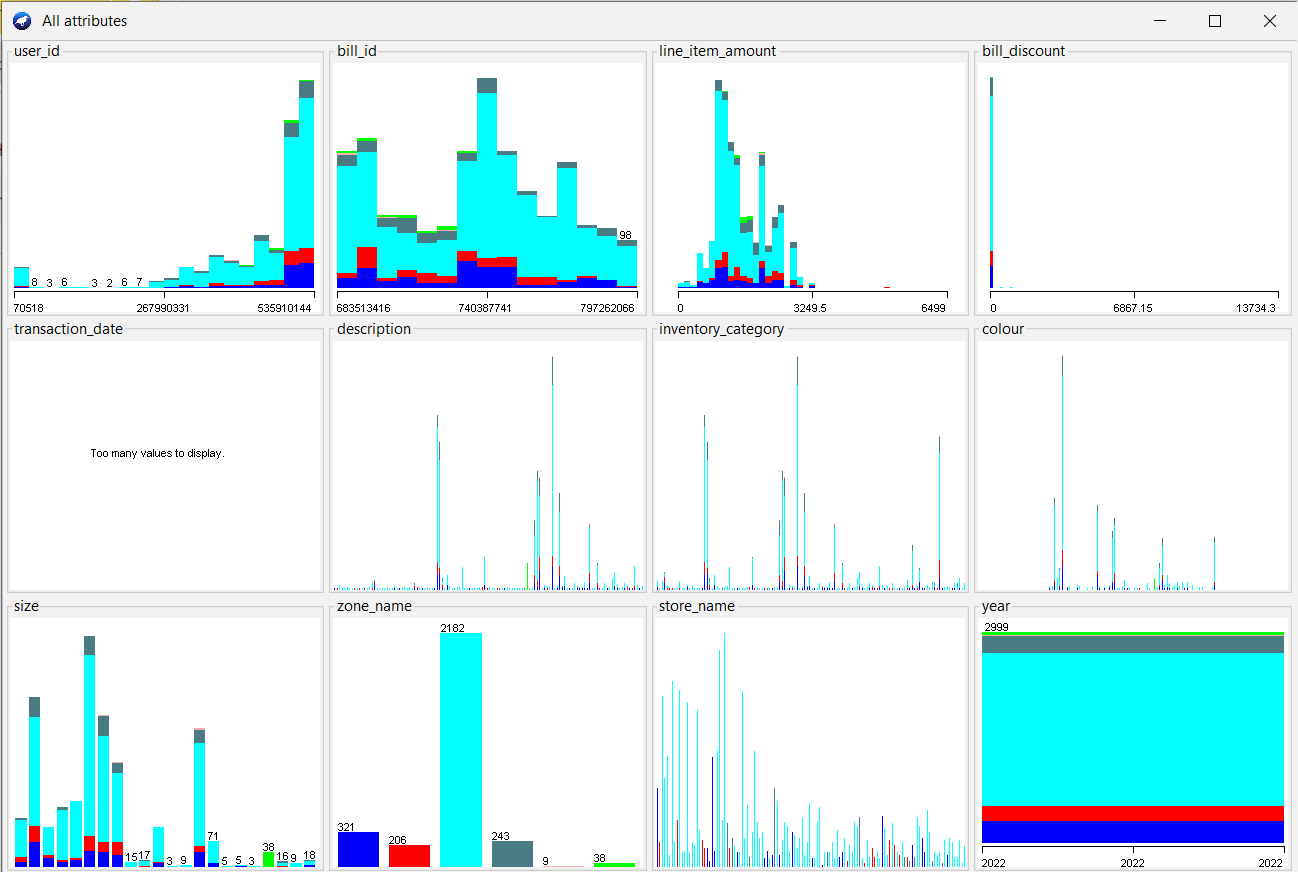
\* Phân tích số liệu:

* line\_item\_amount: Có 1025 giá trị hợp lệ với giá trị trung bình là 1919.71, phân bố từ 0 đến 6499. Có sự chênh lệch lớn giữa giá trị tối đa và trung bình, cho thấy có sự biến động lớn về giá trong các mặt hàng.
* bill\_discount: Có 2993 giá trị hợp lệ, phần lớn là 0, chỉ ra rằng hầu hết các hóa đơn không có chiết khấu.
* year: Tất cả các giao dịch đều diễn ra trong năm 2022.

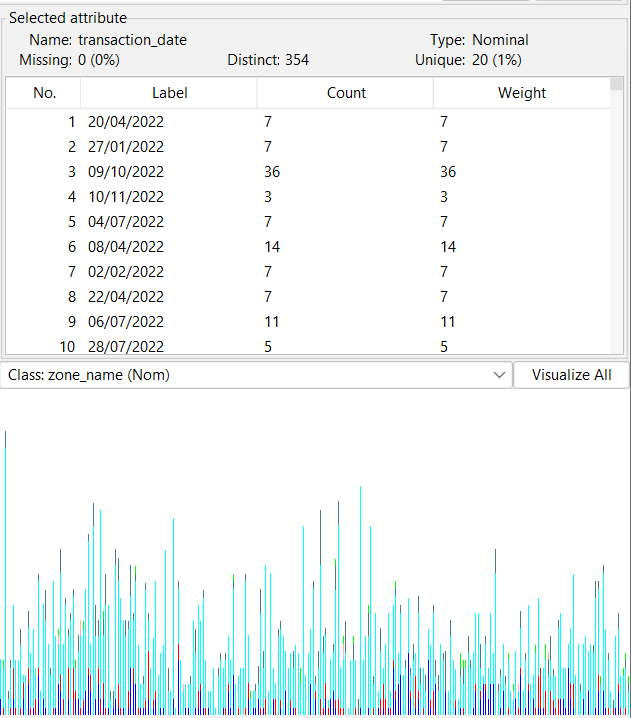
\* Phân tích mô tả:

* description: Có 225 mô tả sản phẩm khác nhau, trong đó mặt hàng "MTR ITMTR00365 Brooklyn FF Black 40" xuất hiện nhiều nhất với 319 lần.
* inventory\_category: 194 danh mục hàng tồn kho khác nhau, cũng có mặt hàng "MTR ITMTR00365 Brooklyn FF Black 40" dẫn đầu về số lần xuất hiện.
* colour: 68 màu sắc khác nhau, với "Black" là màu phổ biến nhất.
* size: 22 kích thước khác nhau, trong đó kích thước '38' là phổ biến nhất.
* zone\_name: Dữ liệu bao gồm 6 khu vực khác nhau, với "South" là khu vực có nhiều giao dịch nhất.
* store\_name: 234 cửa hàng khác nhau, trong đó cửa hàng "South\_6053" có nhiều giao dịch nhất.

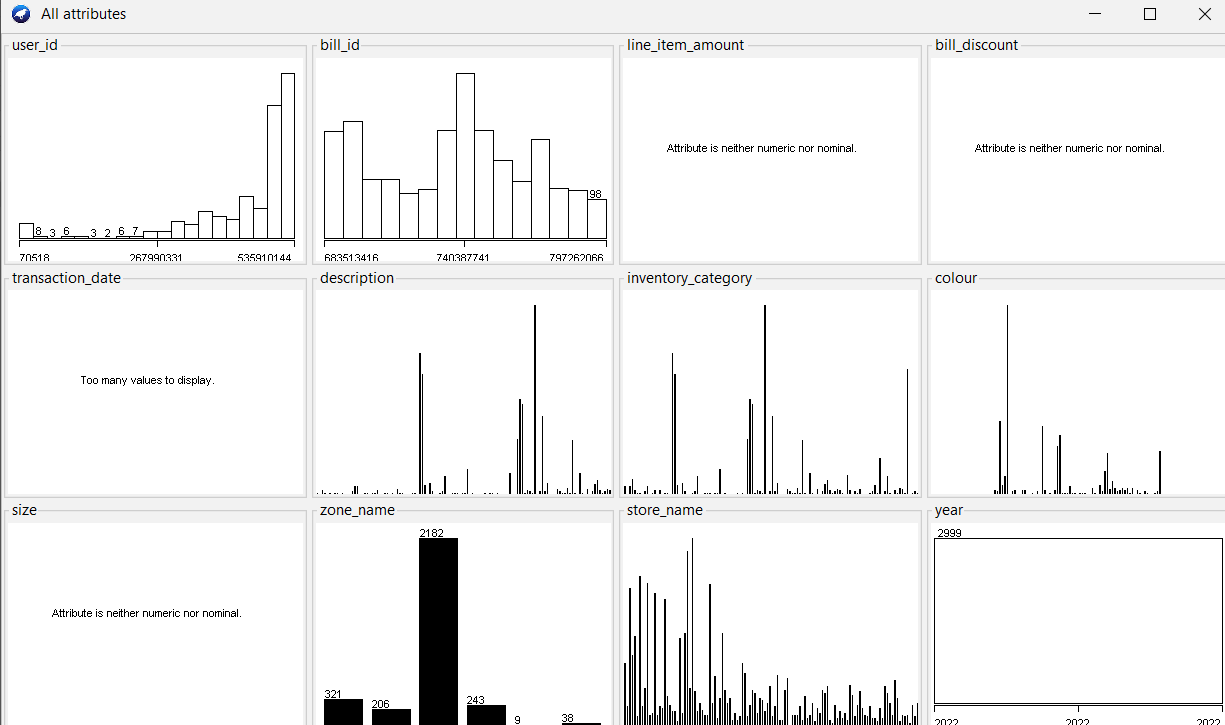
### \* Trực quan hóa dữ liệu trước khi tiền xử lý dữ liệu



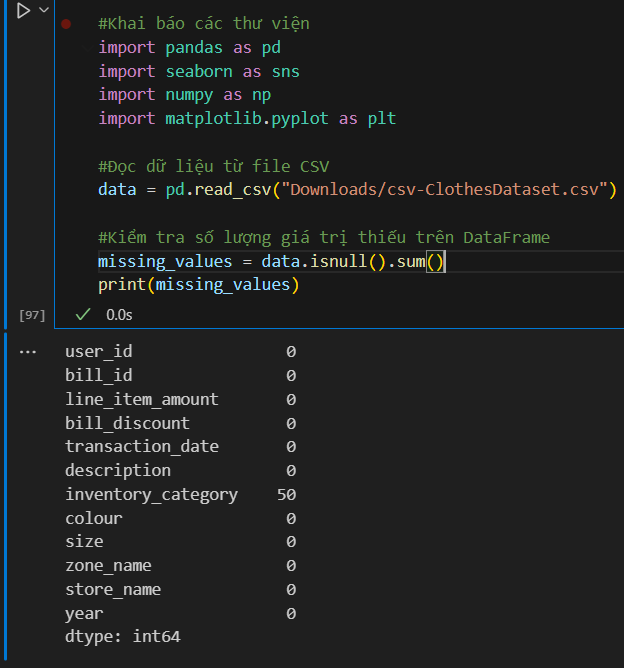
*Ảnh trực quan hóa dữ liệu của tất cả các đặc trưng*



*Ảnh trực quan hóa dữ liệu của đặc trưng transaction\_date (ngày giao dịch)*



Đánh giá dữ liệu trước khi tiền xử lý:



* Có dữ liệu thiếu ở trường:
* inventory\_category : thiếu 50 dữ liệu
* Không xuất hiện dữ liệu ngoại lai

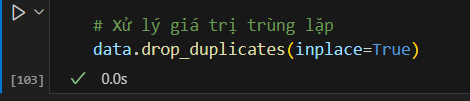
### 

### 

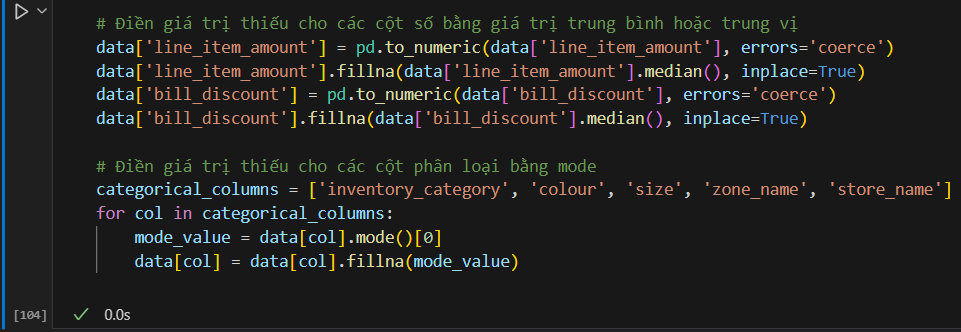
### 

### 2.3 Tiến hành các bước tiền xử lý dữ liệu

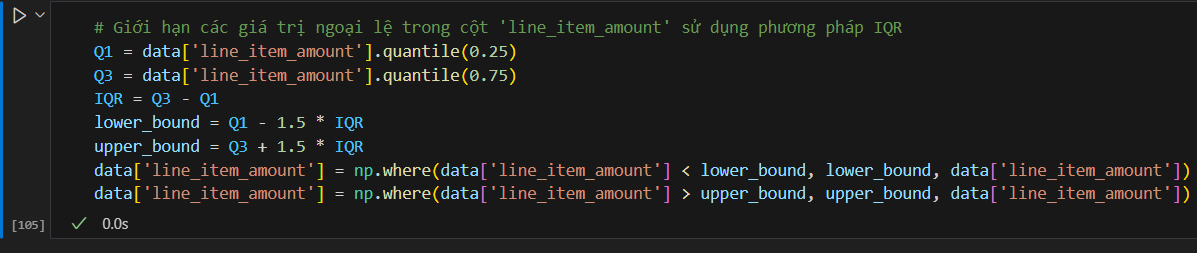
* Xử lý giá trị trùng lặp:

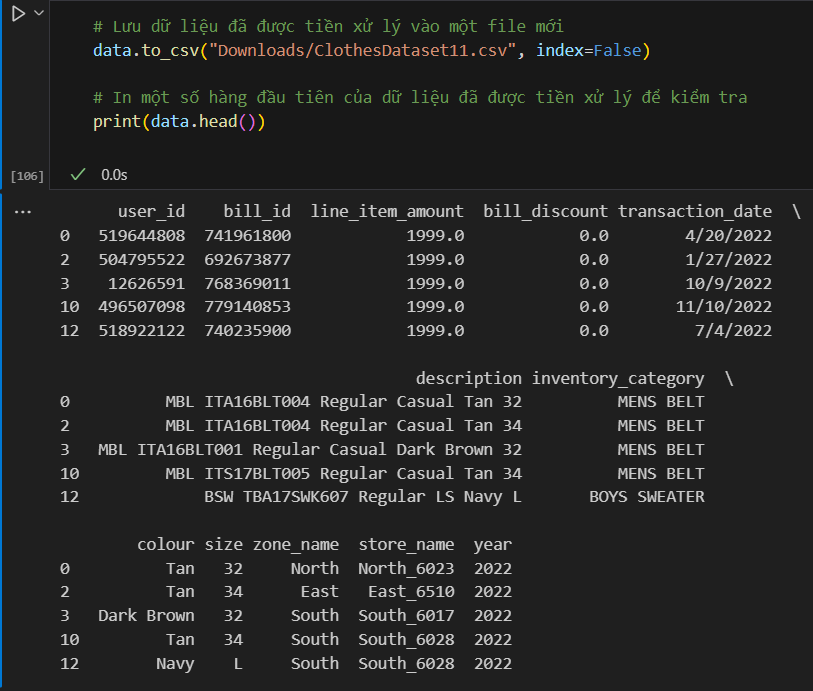


* Xử lý giá trị thiếu:

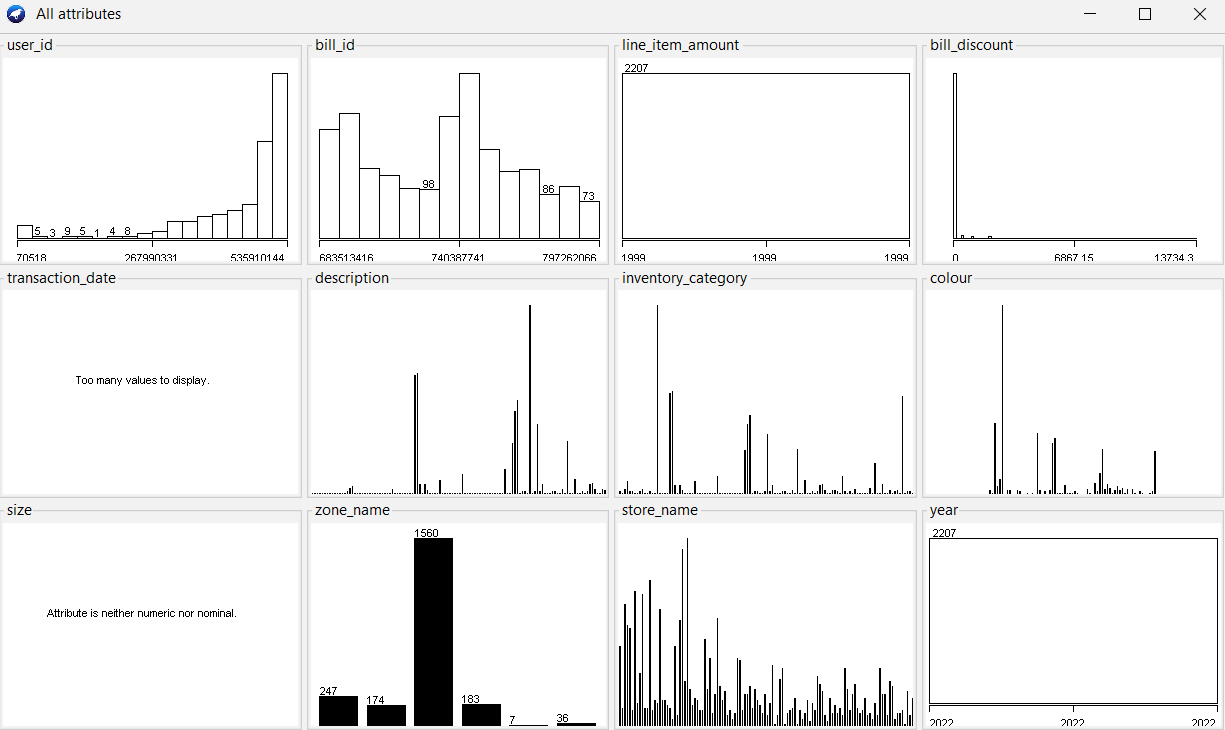


* Xử lý giá trị nhiễu:





### 2.4 Trực quan hóa sau khi đã tiền xử lý dữ liệu:



## **Chương 3. Xây dựng mô hình**

Lựa chọn hướng giải quyết, lựa chọn thuật toán, công nghệ sử dụng cần căn cứ theo tính chất của dữ liệu và mục tiêu của đề tài, kết hợp với việc tham khảo hướng giải quyết từ các đề tài liên quan.

### **1.** **Thuật toán Decision Trees, Support Vector Machines(SVM) và mô hình Bagging Classifier**

**Giới thiệu về thuật toán cây quyết định (Decision Tree)**

- Thuật toán cây quyết định là một phương pháp học máy không tham số, được sử dụng cho các bài toán phân loại và hồi quy.

- Thuật toán này tạo ra một cây quyết định bằng cách tách dữ liệu thành các tập con dựa trên các đặc trưng của dữ liệu.

- Cách hoạt động:

+ Bắt đầu với một nút gốc đại diện cho toàn bộ dữ liệu.

+ Tại mỗi nút, thuật toán chọn một đặc trưng để phân chia dữ liệu thành các tập con. Việc phân chia này được thực hiện bằng cách tối ưu hóa một hàm mục tiêu như Entropy hoặc Gini Impurity.

+ Quá trình phân chia tiếp tục cho đến khi mỗi nút lá chứa một lớp dữ liệu hoặc đạt đến một điều kiện dừng nhất định.

- Ứng dụng trong đề tài:

+ Trong đề tài về đánh giá nguồn nước, cây quyết định có thể được sử dụng để xây dựng một mô hình đánh giá chất lượng nước dựa trên các yếu tố như pH, hàm lượng Sunfat, hàm lượng Organic Carbon,...

+ Mô hình cây quyết định sẽ giúp xác định các yếu tố quan trọng nhất ảnh hưởng đến chất lượng nước và tạo ra các quy tắc quyết định để phân loại các mẫu nước vào các nhóm chất lượng khác nhau.

- Ưu điểm:

+ Dễ hiểu và diễn giải.

+ Có khả năng xử lý cả dữ liệu số và dữ liệu phân loại.

+ Khả năng xử lý các biến số và đặc trưng phức tạp.

- Nhược điểm:

+ Dễ bị overfitting, đặc biệt là khi cây quá sâu.

+ Không linh hoạt trong việc xử lý các mối quan hệ phức tạp trong dữ liệu.

**Thuật toán Support Vector Machines(SVM)**

- SVM là một phương pháp học máy được sử dụng rộng rãi cho các bài toán phân loại và hồi quy. SVM là tìm ra ranh giới phân chia tốt nhất giữa các lớp dữ liệu bằng cách tối đa hóa khoảng cách giữa các điểm dữ liệu gần nhất thuộc hai lớp (margin).

- Cách hoạt động:

+ SVM tìm ra một ranh giới phân chia bằng cách tìm ra siêu phẳng (hyperplane) tốt nhất phân chia hai lớp dữ liệu sao cho khoảng cách từ mỗi điểm dữ liệu gần nhất đến siêu phẳng là lớn nhất.

+ Trong trường hợp dữ liệu không thể phân chia tuyến tính, SVM sử dụng một hàm kernel để ánh xạ dữ liệu vào một không gian cao chiều (high-dimensional space) để tạo ra ranh giới phân chia phi tuyến.

- Ứng dụng trong đề tài:

+ Trong đề tài về đánh giá nguồn nước, SVM có thể được sử dụng để xây dựng một mô hình phân loại chất lượng nước dựa trên các yếu tố như pH, hàm lượng Sunfat, hàm lượng Organic Carbon,...

+ Mô hình SVM sẽ tìm ra một ranh giới phân chia tối ưu giữa các lớp dữ liệu, giúp phân loại các mẫu nước vào các nhóm chất lượng khác nhau.

- Ưu điểm:

+ Hiệu suất cao trong các không gian chiều cao.

+ Hiệu quả khi số lượng chiều dữ liệu lớn hơn số lượng mẫu.

+ Linh hoạt với việc sử dụng các hàm kernel khác nhau để xử lý các vấn đề phi tuyến.

- Nhược điểm:

* Đòi hỏi khá nhiều tài nguyên tính toán khi có một lượng lớn dữ liệu.
* Cần lựa chọn kernel và tinh chỉnh siêu tham số một cách cẩn thận để tránh overfitting.

### **2.Công nghệ áp dụng**

**Thư viện sử dụng:**

● **PANDAS:**

- Pandas là một thư viện Python cung cấp các cấu trúc dữ liệu và công cụ phân tích dữ liệu mạnh mẽ.

- Nó tập trung vào việc cung cấp các đối tượng như DataFrame (bảng dữ liệu hai chiều), Series (danh sách dữ liệu một chiều), và các phương pháp để thực hiện các thao tác như chọn, lọc, sắp xếp và biến đổi dữ liệu.

- Pandas thường được sử dụng để tiền xử lý dữ liệu, làm sạch dữ liệu, thống kê mô tả và chuẩn bị dữ liệu cho việc huấn luyện mô hình máy học.

● **Scikit-Learn (sklearn):**

- Scikit-learn là một thư viện Python mã nguồn mở phổ biến được sử dụng rộng rãi trong Machine Learning và Data Mining

- Ưu điểm:

+ Dễ sử dụng: Scikit-learn cung cấp một giao diện dễ sử dụng và đồng nhất cho việc triển khai và sử dụng các thuật toán Machine Learning. Điều này làm giảm thời gian và công sức cần thiết cho việc học và triển khai mô hình.

+ Tích hợp tốt: Scikit-learn tích hợp tốt với các thư viện Python khác như NumPy, SciPy và pandas, cung cấp một hệ sinh thái mạnh mẽ cho việc xử lý dữ liệu và triển khai mô hình.

+ Đa dạng thuật toán: Scikit-learn cung cấp một loạt các thuật toán Machine Learning cho các nhiệm vụ phân loại, hồi quy, gom cụm và giảm chiều dữ liệu, cung cấp sự linh hoạt trong việc lựa chọn thuật toán phù hợp với bài toán cụ thể.

+ Hiệu suất cao: Scikit-learn cung cấp các triển khai hiệu suất cao của các thuật toán Machine Learning, cho phép xử lý các tập dữ liệu lớn và phức tạp một cách hiệu quả.

● **Joblib:**

Joblib là một thư viện mã nguồn mở trong ngôn ngữ lập trình Python được sử dụng để lưu trữ và tải lại các đối tượng Python, chẳng hạn như mô hình machine learning, một cách hiệu quả.

Mục đích chính của Joblib là giúp lưu trữ các đối tượng lớn trong Python một cách nhanh chóng và hiệu quả mà không cần phải chuyển đổi chúng thành các chuỗi hoặc các định dạng dữ liệu khác. Điều này rất hữu ích trong machine learning khi muốn lưu trữ một mô hình đã được huấn luyện để sử dụng lại sau này mà không cần phải huấn luyện lại từ đầu.

Một số tính năng chính của Joblib bao gồm:

+ Hiệu suất cao: Joblib được thiết kế để đạt được hiệu suất cao khi lưu trữ và tải lại các đối tượng lớn trong Python.

+ Đơn giản và dễ sử dụng: Joblib cung cấp một giao diện đơn giản và dễ sử dụng cho việc lưu trữ và tải lại các đối tượng Python, không yêu cầu người dùng phải thực hiện các bước phức tạp.

+ Tích hợp với machine learning: Joblib thường được sử dụng trong machine learning để lưu trữ và tải lại các mô hình đã được huấn luyện, giúp tiết kiệm thời gian và công sức khi triển khai các ứng dụng machine learning.

* **Tkinter**

Tkinter là một thư viện giao diện người dùng (GUI) được tích hợp sẵn trong Python để tạo ra các ứng dụng có giao diện người dùng đồ họa. Tên gọi "Tkinter" xuất phát từ việc kết hợp giữa "Tk" - một toolkit GUI được sử dụng rộng rãi, và "inter" - viết tắt của "interface".

Tkinter cung cấp một loạt các widget (đối tượng giao diện) như cửa sổ, nút, hộp văn bản, ô nhập liệu, danh sách, thanh cuộn, v.v., cho phép người lập trình tạo ra các ứng dụng có giao diện người dùng phong phú và tương tác.

Một số đặc điểm của Tkinter bao gồm:

- Dễ học và sử dụng: Tkinter có cú pháp đơn giản và dễ hiểu, làm cho việc tạo ra các ứng dụng GUI trở nên dễ dàng đối với người mới bắt đầu.

- Tích hợp sẵn: Tkinter là một phần của thư viện chuẩn của Python, không cần phải cài đặt thêm. Điều này có nghĩa là có thể bắt đầu làm việc ngay lập tức mà không cần phải tìm kiếm và cài đặt thư viện bổ sung.

- Đa nền tảng: Tkinter hoạt động trên nhiều nền tảng hệ điều hành khác nhau như Windows, macOS và Linux.

- Mạnh mẽ và linh hoạt: Mặc dù Tkinter có thể được coi là một thư viện GUI đơn giản, nhưng nó vẫn cung cấp đủ tính năng để tạo ra các ứng dụng GUI phức tạp với giao diện tương tác và đẹp mắt.

* **NumPy**

NumPy là một thư viện Python cung cấp hỗ trợ cho các phép toán trên mảng và ma trận, cùng với một loạt các chức năng cho các phép toán toán học trên các cấu trúc dữ liệu này. Nó là một trong những thư viện quan trọng nhất trong hệ sinh thái khoa học dữ liệu của Python và thường được sử dụng trong các ứng dụng liên quan đến tính toán số học và xử lý mảng dữ liệu lớn.

Các tính năng chính của NumPy bao gồm:

● Các đối tượng mảng nhiều chiều (ndarray) cho phép lưu trữ dữ liệu số học nhanh chóng và hiệu quả.

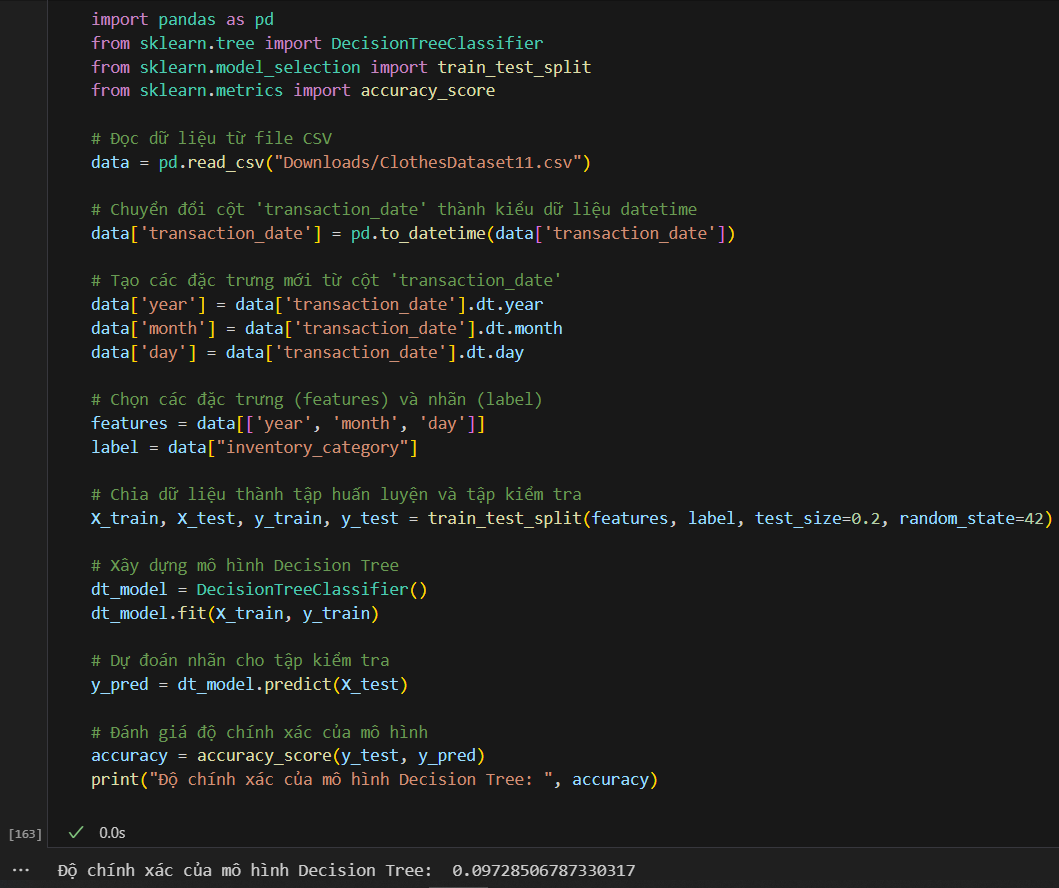
● Một bộ công cụ mạnh mẽ cho phép thực hiện các phép toán toán học cơ bản và phức tạp trên mảng, bao gồm cả phép toán phân phối, biến đổi và ma trận.

● Các chức năng cho phép tạo ra các mảng dữ liệu, tạo mảng mới từ dữ liệu hiện có, truy cập và thay đổi các phần tử của mảng, và nhiều hơn nữa.

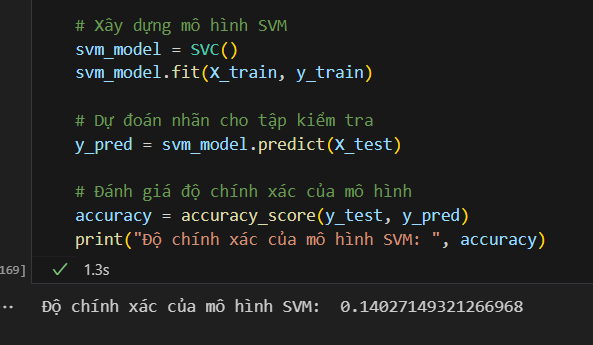
NumPy cũng là một phần quan trọng của nền tảng cho nhiều thư viện Python khác trong lĩnh vực khoa học dữ liệu và máy học, như SciPy, pandas và scikit-learn.

### **3.Xây dựng mô hình dự đoán**

* **Decision Tree**

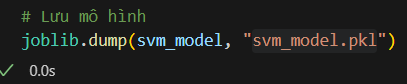
****

* **SVM**



→ Ta thấy được rằng mước độ chính xác cả SVM hơn Decision Tree nên sẽ chon SVM để tích hợp vào hệ thống

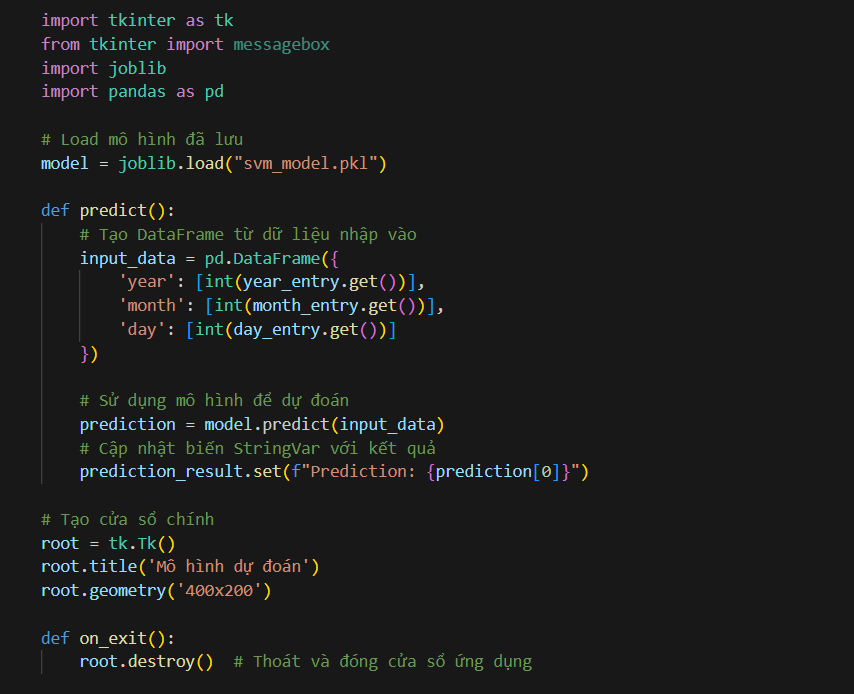
* Lưu mô hình

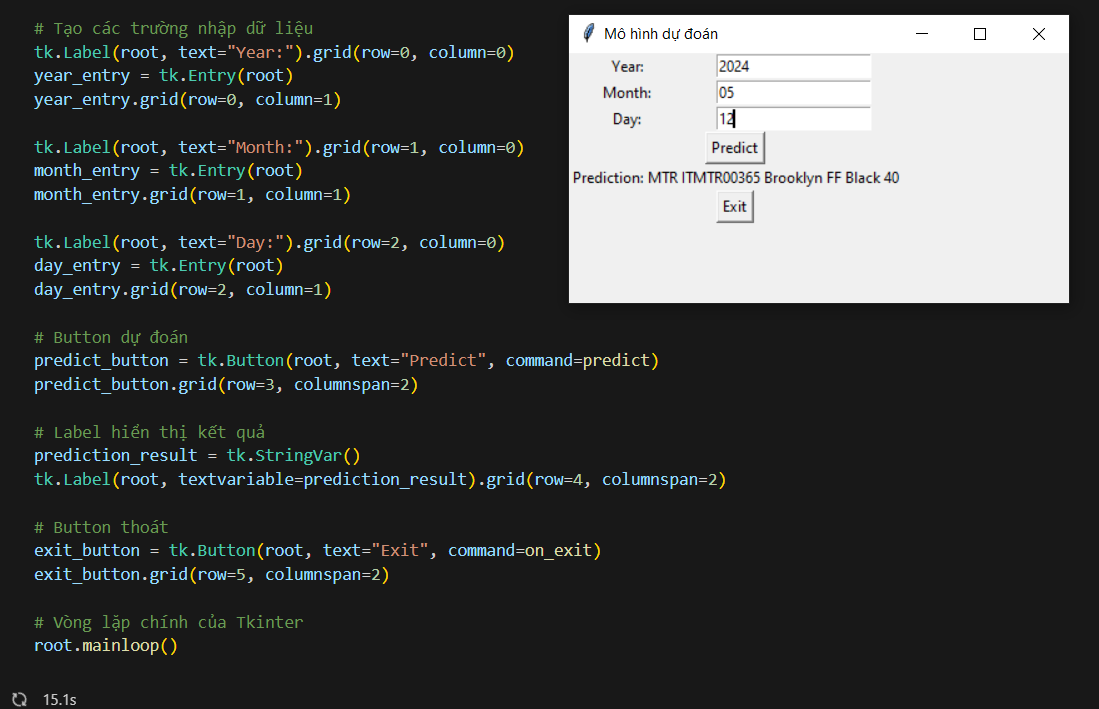


## **Chương 4. Tích hợp mô hình vào hệ thống**

### **Tạo giao diện cho chức năng**

* Sử dụng thư viện Tkinter được tích hợp sẵn trong Python để tạo ra các ứng dụng có giao diện người dùng đồ họa để xây dựng
* Tạo giao diện cho chức năng “dự đoán sản phẩm” thông qua các chỉ số nước được nhập vào từ bàn phím.





## **Kết luận**

Kết luận theo từng mục tiêu của đề tài

## **Tài liệu tham khảo**

**Phụ lục**