ĐẠI HỌC BÁCH KHOA HÀ NỘI

TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

──────── \* ───────

**BÀI TẬP LỚN**

MÔN: PROJECT III

**Phân tích dữ liệu và xây dựng mô hình dự đoán dữ liệu bán hàng**

Mã lớp học : 709167

Giáo viên hướng dẫn : Th.S.**Nguyễn Hồng Phương**

Sinh viên thực hiện : Nguyễn Minh Tuấn

MSSV : 20183657

***Hà Nội, tháng 01 năm 2022***

MỤC LỤC

[MỤC LỤC 2](#_Toc93848773)

[LỜI NÓI ĐẦU 4](#_Toc93848774)

[CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI 5](#_Toc93848775)

[1.1. Mô tả yêu cầu bài toán 5](#_Toc93848776)

[1.2. Tập dữ liệu sử dụng 5](#_Toc93848777)

[1.3. Đầu ra của bài toán 6](#_Toc93848778)

[CHƯƠNG 2. KHAI PHÁ DỮ LIỆU 7](#_Toc93848779)

[2.1. Tích hợp và làm sạch dữ liệu 7](#_Toc93848780)

[2.2. Phân tích và khám phá dữ liệu 7](#_Toc93848781)

[2.2.1. Phân tích dữ liệu 7](#_Toc93848782)

[2.2.2. Khám phá dữ liệu 9](#_Toc93848783)

[2.3. Tiền xử lý dữ liệu 12](#_Toc93848784)

[CHƯƠNG 3. CƠ SỞ LÝ THUYẾT CÁC KĨ THUẬT HỌC MÁY SỬ DỤNG 14](#_Toc93848785)

[3.1. Thuật toán Random Forest 14](#_Toc93848786)

[3.1.1. Định Nghĩa 14](#_Toc93848787)

[3.1.2. Phương Pháp 14](#_Toc93848788)

[3.2. Thuật toán SVM 16](#_Toc93848789)

[3.2.1. Định Nghĩa 16](#_Toc93848790)

[3.2.2. Phương Pháp 16](#_Toc93848791)

[3.2.3. Các Bước Chính 18](#_Toc93848792)

[CHƯƠNG 4. THỬ NGHIỆM VÀ ĐÁNH GIÁ MÔ HÌNH 20](#_Toc93848793)

[4.1. Phương thức đánh giá 20](#_Toc93848794)

[4.1.1. Mean Absolute Error - MAE 20](#_Toc93848795)

[4.1.2. Mean squared error - MSE 20](#_Toc93848796)

[4.1.3. Root Mean Square Error (RMSE) 20](#_Toc93848797)

[4.1.4. R2 Square 20](#_Toc93848798)

[4.2. Thử nghiệm mô hình 21](#_Toc93848799)

[4.2.1. Thử nghiệm các mô hình khác nhau 21](#_Toc93848800)

[4.2.2. Thử nghiệm mô hình Random Forest 21](#_Toc93848801)

[CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 23](#_Toc93848802)

[5.1. Kết Luận 23](#_Toc93848803)

[5.2. Hướng Phát Triển 23](#_Toc93848804)

LỜI NÓI ĐẦU

Lời đầu tiên, em xin cảm ơn thầy đã đồng ý và giúp đỡ em hoàn thiện đề tài này một cách tốt nhất. Nhờ làm đề tài này, đã giúp em nâng cao được kỹ năng cá nhân về lĩnh vực mà mình đang theo đuổi và phát triển. Đồng thời trau dồi thêm các yêu cầu về chuyên môn mà mình cần có để phát triển.

Đề tài mà em chọn lựa là đề tài về phân tích dữ liệu và xây dựng mô hình dự đoán giá của sản phẩm rượu. Data Analyst là một lĩnh vực đang khá được ưa chuộng vào thời điểm bùng nổ của việc bán hàng hiện nay. Khi chúng ta bán hàng và thu thập dữ liệu ta cần phân tích và đánh giá lại kết quả để so sánh với kế hoạch đề ra ban đầu. Từ đó, biết được điểm mạnh, điểm yếu để cải thiện kế hoạch giúp tăng hiệu suất kinh doanh lên đáng kể. Ngoài ra việc phân tích và dự đoán được tương lai cũng giúp ta hoạch định chiến lược một cách rõ ràng hơn nhằm nâng cao khả năng bán hàng đúng đối tượng, đúng người dùng. “*Cung, Cầu*” chắc hẳn ai cũng biết cụm từ này trong kinh doanh, vậy để biết được cung và cầu phát triển như thế nào để đưa được sự cung ứng hợp lý tới đúng đối tượng “Cầu” đó chính là dựa vào công việc phân tích dữ liệu.

Trong quá trình tiếp cận đề tài, em còn gặp nhiều khó khăn bởi đây không phải là định hướng ban đầu. Tất cả nguồn dữ liệu và tài liệu đều do em tìm hiểu trên nguồn Internet nên hi vọng có gì chưa hoàn chỉnh được thầy chỉnh sửa và đánh giá góp ý giúp em. Ngoài ra trong quá trình xây dựng mô hình em cũng còn gặp nhiều khó khăn khi mới xây dựng lần đầu nên mất nhiều thời gian để thử nghiệm và chọn lựa . Nhưng cuối cùng sau quá trình thử nghiệm em cũng đã rút ra và chọn được phương pháp thích hợp nhất để thực hiện. Hi vọng em sẽ được thầy chia sẽ thêm kinh nghiệm để huấn luyện và đánh giá mô hình một cách tốt hơn nữa.

Em xin cảm ơn thầy đã giúp đỡ và mong chờ sự đánh giá tích cực từ thầy !

# GIỚI THIỆU ĐỀ TÀI

## Mô tả yêu cầu bài toán

* Bài toán : Phân tích dữ liệu và xây dựng mô hình dự đoán dữ liệu bán hàng về sản phẩm rượu trên thị trường Mỹ.
* Yêu cầu :

+ Bài toán đặt ra vấn đề cần phân tích được dữ liệu thu thập được, làm rõ các insight có giá trị của tập dữ liệu qua đó đánh giá được hiệu quả, chất lượng bán hàng của sản phẩm

+ Sau khi phân tích dữ liệu xây dựng mô hình dự đoán giá bán của mặt hàng rượu trong tương lại , cụ thể trong 3 tháng cuối năm 2017

+ Xây dựng được một bản kết hoạch sơ lược đúc rút được sau khi phân tích và dữ đoán từ tập dữ liệu

* Kết quả :

+ Phân tích dữ liệu : Các tri thức từ tập dữ liệu

+ Mô hình Predict : Đánh giá kết quả mô hình => Giá bán trong tương lai

* Đối tượng tác động : Các sản phẩm mặt hàng là rượu trên thị trường

## Tập dữ liệu sử dụng

Dataset : Nguồn : <https://www.kaggle.com/residentmario/iowa-liquor-sales>

**Mô tả :** Tập dữ liệu về bán hàng - Dòng sản phẩm : Rượu

**Thời gian :** Từ ngày 1.1.2012 – 31.12.2017

**Thuộc tính :** ( 24 columns x 12591077 rows)

**Ưu điểm :**

* Tập dữ liệu rất lớn
* Chi tiết, đầy đủ các trường dữ liệu

**Nhược điểm :**

* Thời gian tập dữ liệu năm 2012-2017 ⇒ Không còn phù hợp với xu hướng hiện tại

## Đầu ra của bài toán

Gồm có :

* EDA dữ liệu : các loại biểu đồ cần thiết để trả lời cho các câu hỏi về dữ liệu
* Bảng kết quả đánh giá mô hình
* Kết quả dự đoán giá của các sản phẩm : Lấy 5 sản phẩm cao nhất trong năm 2017
* Đề xuất tóm tắt cần thực hiện sau khi kết thúc

# KHAI PHÁ DỮ LIỆU

## Tích hợp và làm sạch dữ liệu

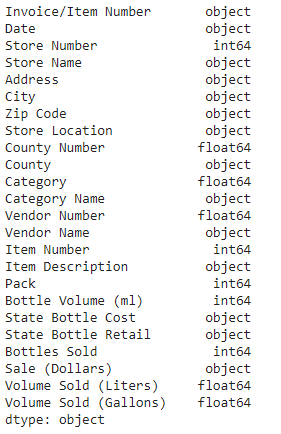
Bước này thực hiện làm sạch dữ liệu, phân tách label và tổng hợp dữ liệu cho quá trình training và kiểm tra model

* Loại bỏ các giá trị null
* Chuyển trường **Datetime** => **Year, Month, Date**
* Loại bỏ các kí tự đặc biệt : $,/
* Chuyển kiểu dữ liệu string sang int và float

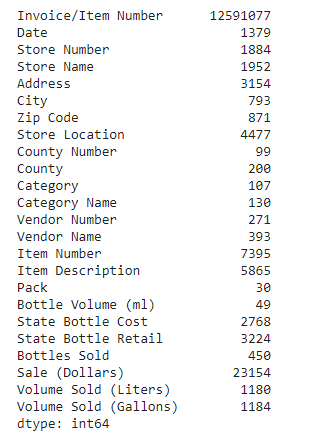
## Phân tích và khám phá dữ liệu

### Phân tích dữ liệu

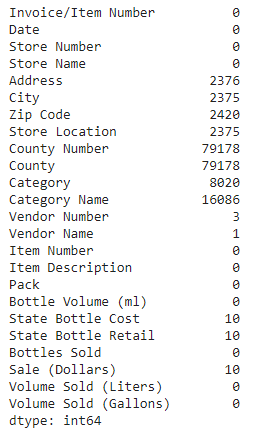
* Kiểm tra kiểu dữ liệu của các features



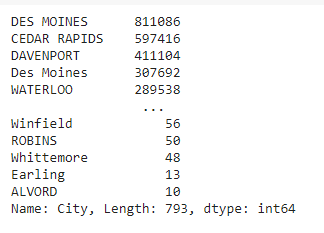
* Kiểm tra số lượng dữ liệu duy nhất của từng features



* Kiểm tra số lượng giá trị null của từng features

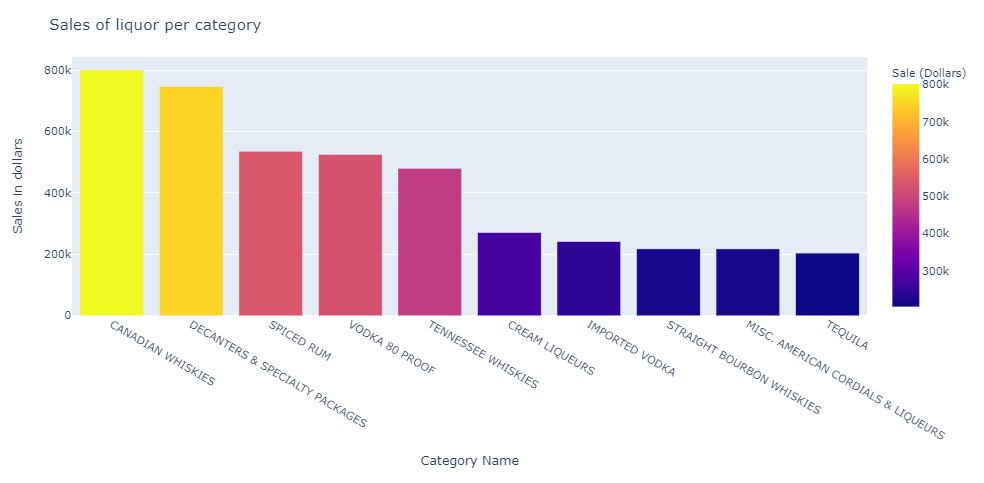


* Số lượng sản phẩm của các thành phố khác nhau

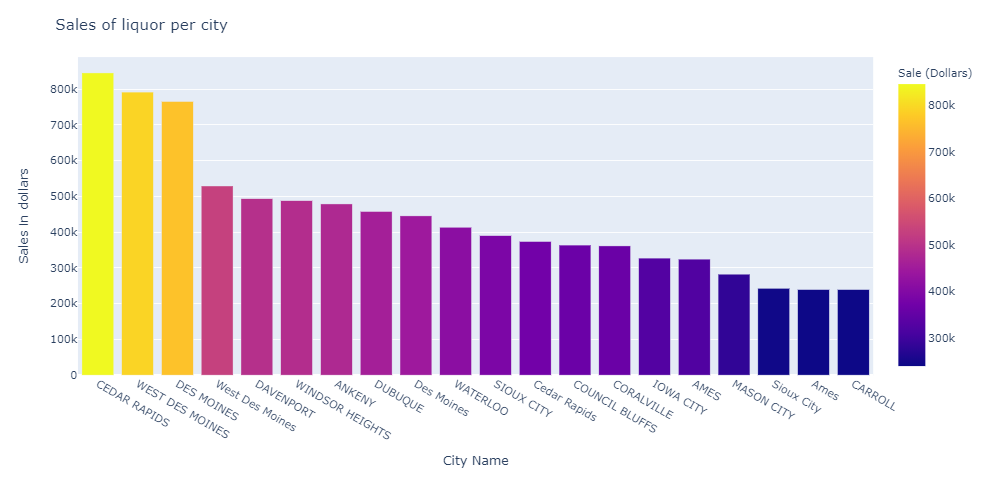


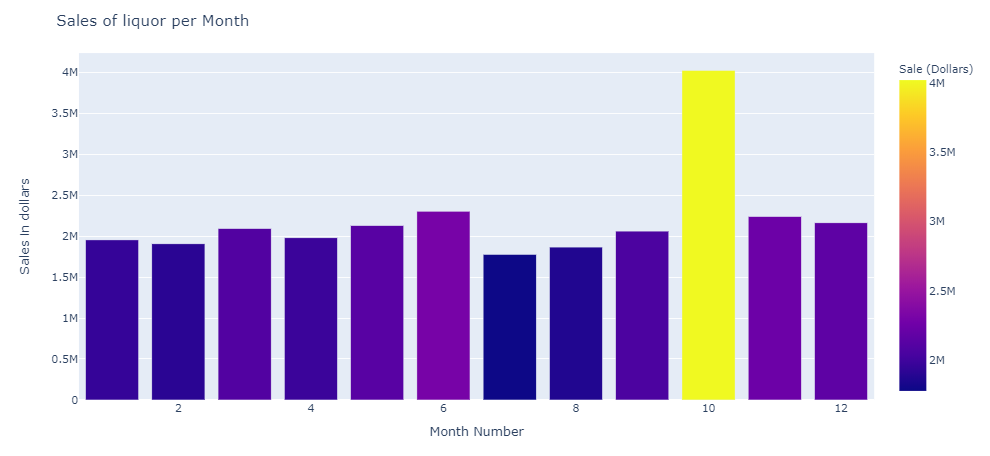
### Khám phá dữ liệu

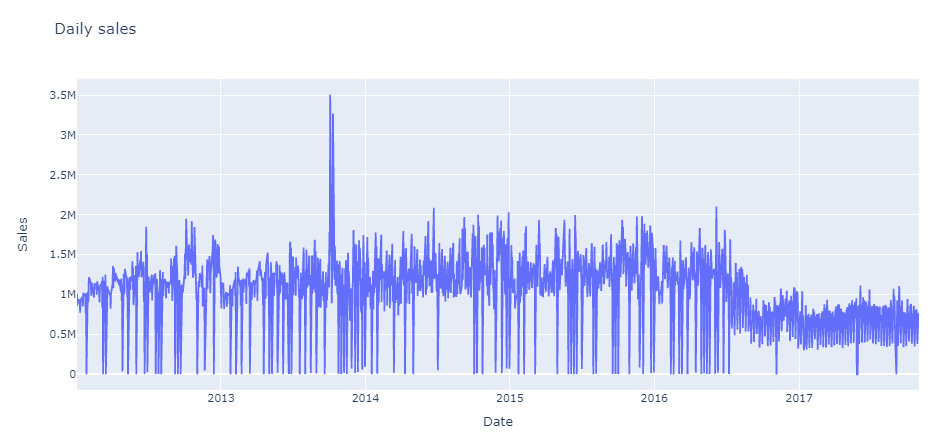
* Top 10 Categories of liquors that has higest sale



* Sale of liquor by city name (top 20)

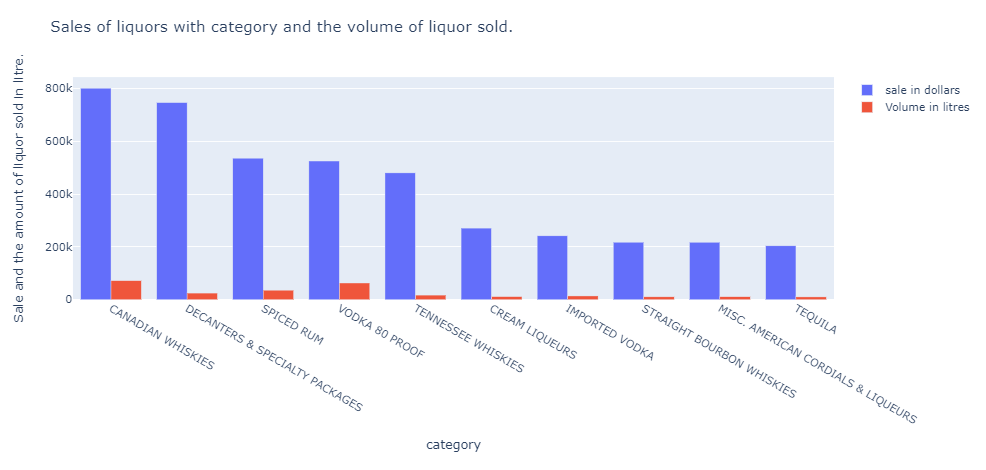




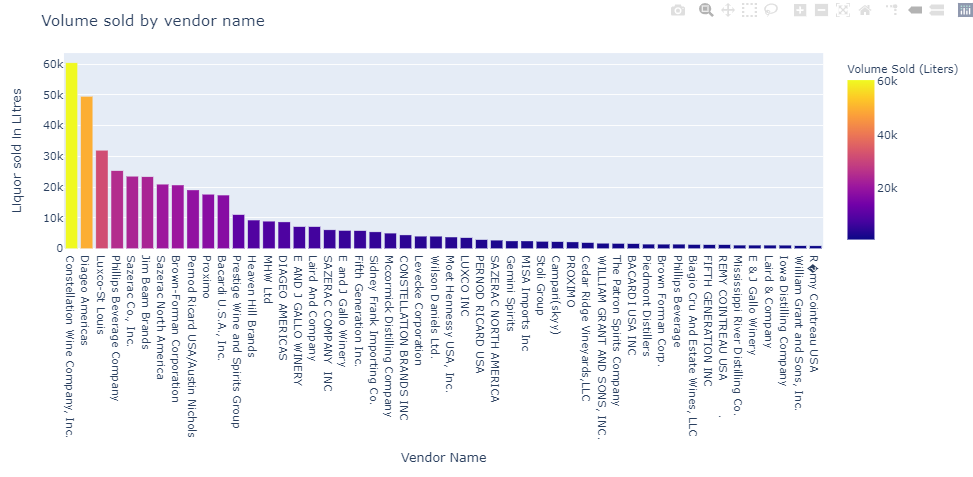


* Sales of liquors with category and the volumne of liquor sold

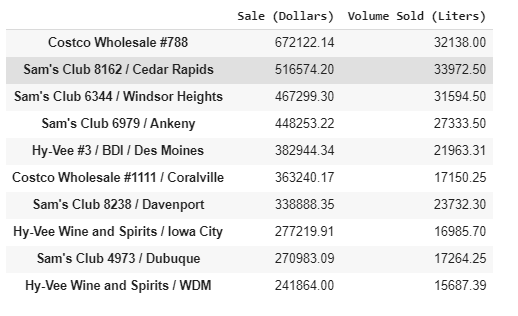




* Volume sold in litres by vendor name



* Top 10 Stores of liquors that has highest sale



## Tiền xử lý dữ liệu

Bước này ta loại bỏ các features không sử dụng nhiều tới trong lúc train mô hình và chuẩn hóa dữ liệu. Các trường loại bỏ bao gồm : *'Invoice/Item Number', 'Date' , 'Store Name','City','Category Name', 'Address', 'Zip Code' , 'Store Location', 'County' , 'Vendor Name','Vendor Number', 'Item Description'.*

Ngoài ra, tách bộ dữ liệu thành các tập category riêng biệt để đảm bảo sự hồi quy của dữ liệu trong mô hình predict. Ta tách tập dữ liệu thành train và test.

Trong đó :

* Train : Các features trừ *‘Sales Dollars’*
* Test : Chứa collumn *‘Sales Dollars’*

Tách thành các bộ nhỏ theo Category Name :

0 AMERICAN VODKAS

1 CANADIAN WHISKIES

2 AMERICAN VODKA

3 STRAIGHT BOURBON WHISKIES

4 SPICED RUM

5 TEMPORARY & SPECIALTY PACKAGES

6 IMPORTED VODKAS

7 WHISKEY LIQUEUR

8 IMPORTED BRANDIES

9 TENNESSEE WHISKIES

10 AGAVE TEQUILA

11 AMERICAN FLAVORED VODKA

12 BLENDED WHISKIES

13 IMPORTED VODKA

14 IMPORTED CORDIALS & LIQUEURS

15 IMPORTED FLAVORED VODKA

16 MIXTO TEQUILA

17 FLAVORED RUM

18 CREAM LIQUEURS

19 AMERICAN SCHNAPPS

20 IRISH WHISKIES

21 WHITE RUM

22 AMERICAN BRANDIES

23 SCOTCH WHISKIES

24 COCKTAILS /RTD

25 SINGLE MALT SCOTCH

26 IMPORTED SCHNAPPS

27 SPECIAL ORDER ITEMS

28 IMPORTED DRY GINS

29 AMERICAN DRY GINS

30 AMERICAN CORDIALS & LIQUEUR

31 HOLIDAY VAP

32 COCKTAILS RTD

33 AMERICAN CORDIALS & LIQUEURS

34 COFFEE LIQUEURS

35 STRAIGHT RYE WHISKIES

36 GOLD RUM

37 NEUTRAL GRAIN SPIRITS

38 CORN WHISKIES

39 AGED DARK RUM

40 SINGLE BARREL BOURBON WHISKIES

41 TRIPLE SEC

42 BOTTLED IN BOND BOURBON

43 IMPORTED DISTILLED SPIRIT SPECIALTY

44 AMERICAN DISTILLED SPIRIT SPECIALTY

45 IMPORTED CORDIALS & LIQUEUR

46 MEZCAL

47 IOWA DISTILLERIES

48 IMPORTED DISTILLED SPIRITS SPECIALTY

49 AMERICAN DISTILLED SPIRITS SPECIALTY

50 IOWA DISTILLERY WHISKIES

51 NEUTRAL GRAIN SPIRITS FLAVORED

52 FLAVORED GIN

53 AMERICAN SLOE GINS

54 TEMPORARY & SPECIALTY PACKAGES

55 DELISTED ITEMS

56 AMERICAN GINS

# CƠ SỞ LÝ THUYẾT CÁC KĨ THUẬT HỌC MÁY SỬ DỤNG

## Thuật toán Random Forest

Random forest là thuật toán supervised learning, có thể giải quyết cả bài toán regression và classification.

### Định Nghĩa

Random forest là là một phương pháp học tập tổng hợp để phân loại, hồi quy và các nhiệm vụ khác hoạt động bằng cách xây dựng vô số cây quyết định tại thời điểm đào tạo. Đối với các nhiệm vụ phân loại, đầu ra của rừng ngẫu nhiên là loại được chọn bởi hầu hết các cây. Đối với các nhiệm vụ hồi quy, giá trị trung bình hoặc dự đoán trung bình của các cây riêng lẻ được trả về.

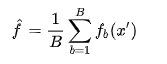
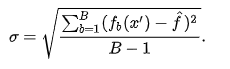
Random forest nhìn chung tốt hơn cây quyết định, nhưng độ chính xác của chúng thấp hơn cây tăng cường độ dốc. Tuy nhiên, đặc điểm dữ liệu có thể ảnh hưởng đến hiệu suất của chúng.

### Phương Pháp

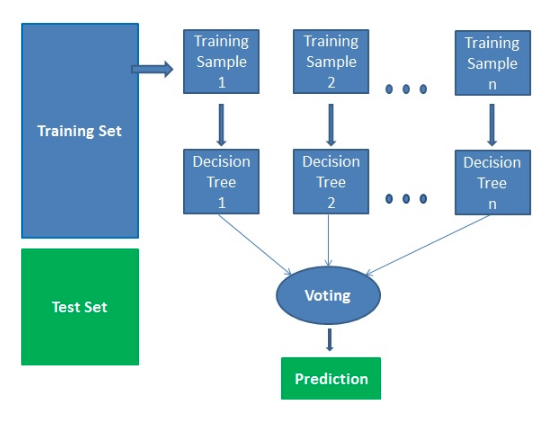
Thuật toán hoạt động theo bốn bước:

1. Chọn các mẫu ngẫu nhiên từ tập dữ liệu đã cho.
2. Thiết lập cây quyết định cho từng mẫu và nhận kết quả dự đoán từ mỗi quyết định cây.

Giá trị dự đoán được tính theo công thức :

1. Bỏ phiếu cho mỗi kết quả dự đoán.
2. Chọn kết quả được dự đoán nhiều nhất là dự đoán cuối cùng.



**Ưu điểm:** Random forests được coi là một phương pháp chính xác và mạnh mẽ vì số cây quyết định tham gia vào quá trình này. Nó không bị vấn đề overfitting. Lý do chính là nó mất trung bình của tất cả các dự đoán.

**Nhược điểm:** Random forests chậm tạo dự đoán bởi vì nó có nhiều cây quyết định. Bất cứ khi nào nó đưa ra dự đoán, tất cả các cây trong rừng phải đưa ra dự đoán cho cùng một đầu vào cho trước và sau đó thực hiện bỏ phiếu trên đó. Toàn bộ quá trình này tốn thời gian. Mô hình khó hiểu hơn so với cây quyết định.

## Thuật toán SVM

Thuật toán được sử dụng trong bài toán phân lớp (Classification) và dự đoán (Prediction) là hai bài toán cơ bản và có rất nhiều ứng dụng trong tất cả các lĩnh vực như: học máy, nhận dạng, trí tuệ nhân tạo, .v.v .

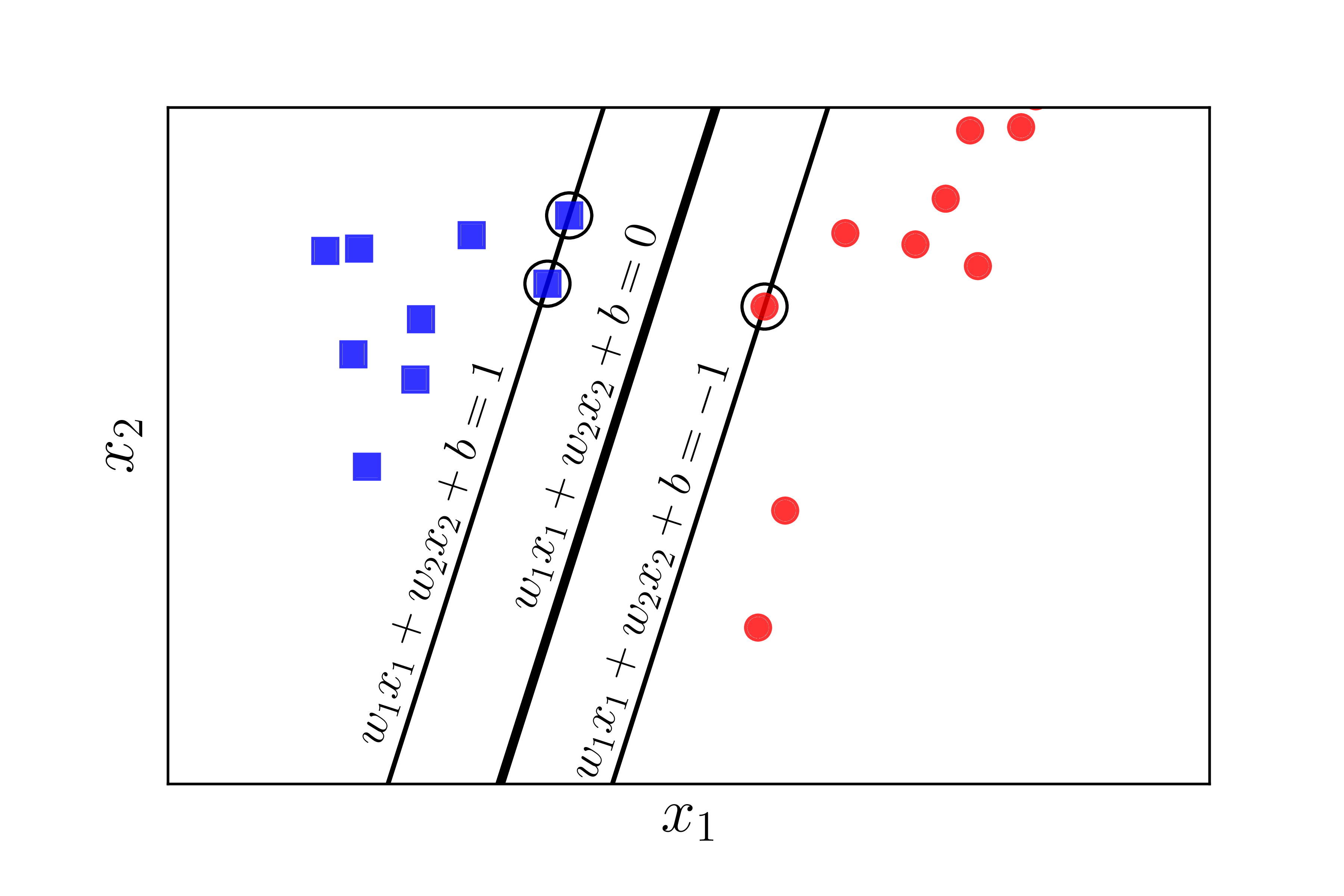
Phương pháp SVM được coi là công cụ mạnh cho những bài toán phân lớp phi tuyến tính được các tác giả Vapnik và Chervonenkis phát triển mạnh mẽ năm 1995. Phương pháp này thực hiện phân lớp dựa trên nguyên lý Cực tiểu hóa Rủi ro có Cấu trúc SRM (Structural Risk Minimization), được xem là một trong các phương pháp phân lớp giám sát không tham số tinh vi nhất cho đến nay. Các hàm công cụ đa dạng của SVM cho phép tạo không gian chuyển đổi để xây dựng mặt phẳng phân lớp.

### Định Nghĩa

Là phương pháp dựa trên nền tảng của lý thuyết thống kê nên có một nền tảng toán học chặt chẽ để đảm bảo rằng kết quả tìm được là chính xác. Là thuật toán học giám sát (supervied learning) được sử dụng cho phân lớp dữ liệu. Là 1 phương pháp thử nghiệm, dưa ra 1 trong những phương pháp mạnh và chính xác nhất trong số các thuật toán nổi tiếng về phân lớp dữ liệu SVM là một phương pháp có tính tổng quát cao nên có thể được áp dụng cho nhiều loại bài toán nhận dạng và phân loại.

### Phương Pháp

Cho trước một tập huấn luyện, được biểu diễn trong không gian vector, trong đó mỗi bản ghi là một điểm dữ liệu , phương pháp này tìm ra một siêu phẳng quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng là lớp “+” và lớp “-”. Chất lượng của siêu phẳng này được quyết định bởi khoảng cách (gọi là biên) của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng này. Khi đó, khoảng cách biên càng lớn thì mặt phẳng quyết định càng tốt, đồng thời việc phân loại càng chính xác.



SVM thực chất là một bài toán tối ưu, mục tiêu của thuật toán này là tìm được một không gian F và siêu phẳng quyết định f trên F sao cho sai số phân loại là thấp nhất. Cho tập mẫu {(x1, y1), (x2, y2), … (xf , yf )} với xi ∈ Rn , thuộc vào hai lớp nhãn: yi ∈ {-1,1}.

Ta có, phương trình siêu phẳng chứa vectơ   trong không gian:

Như vậy f( ) biểu diễn sự phân lớp của vào hai lớp như đã nêu. Ta nói yi= +1 nếu € lớp I và yi = -1 nếu € lớp II . Khi đó, để có siêu phẳng f ta sẽ phải giải bài toán sau:

Tìm  với W thỏa mãn điều kiện sau:

Bài toán SVM có thể giải bằng kỹ thuật sử dụng toán tử Lagrange để biến đổi về thành dạng đẳng thức. Một đặc điểm thú vị của SVM là mặt phẳng quyết định chỉ phụ thuộc các Support Vector và nó có khoảng cách đến mặt phẳng quyết định

là 1/ . Cho dù các điểm khác bị xóa đi thì thuật toán vẫn cho kết quả giống nhau ban đầu. Đây chính là điểm nổi bật của phương pháp SVM so với các phương pháp khác vì tất cả các dữ liệu trong tập huấn luyện đều được đùng để tối ưu hóa kết quả.

Bài toán đặt ra là: Xác định hàm phân lớp để phân lớp các mẫu trong tương lai, nghĩa là với một mẫu dữ liệu mới xi thì cần phải xác định xi được phân vào lớp +1 hay lớp -1 Để xác định hàm phân lớp dựa trên phương pháp SVM, ta sẽ tiến hành tìm hai siêu phẳng song song sao cho khoảng cách y giữa chúng là lớn nhất có thể để phân tách hai lớp này ra làm hai phía. Hàm phân tách tương ứng với phương trình siêu phẳng nằm giữa hai siêu phẳng tìm được. Các điểm mà nằm trên hai siêu phẳng phân tách được gọi là các Support Vector. Các điểm này sẽ quyết định đến hàm phân tách dữ liệu.

### Các Bước Chính

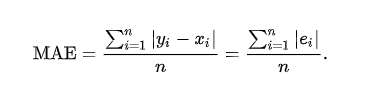
* Phương pháp SVM yêu cầu dữ liệu được diễn tả như các vector của các số thực. Như vậy nếu đầu vào chưa phải là số thì ta cần phải tìm cách chuyển chúng về dạng số của SVM
* Tiền xử lý dữ liệu: Thực hiện biến đổi dữ liệu phù hợp cho quá trình tính toán, tránh các số quá lớn mô tả các thuộc tính. Thường nên co giãn (scaling) dữ liệu để chuyển về đoạn [-1, 1] hoặc [0, 1].
* Chọn hàm hạt nhân: Lựa chọn hàm hạt nhân phù hợp tương ứng cho từng bài toán cụ thể để đạt được độ chính xác cao trong quá trình phân lớp.
* Thực hiện việc kiểm tra chéo để xác định các thám số cho ứng đụng. Điều này cũng quyết định đến tính chính xác của quá trình phân lớp.
* Sử dụng các tham số cho việc huấn luyện với tập mẫu. Trong quá trình huấn luyện sẽ sử dụng thuật toán tối ưu hóa khoảng cách giữa các siêu phẳng trong quá trình phân lớp, xác định hàm phân lớp trong không gian đặc trưng nhờ việc ánh xạ dữ liệu vào không gian đặc trưng bằng cách mô tả hạt nhân, giải quyết cho cả hai trường hợp dữ liệu là phân tách và không phân tách tuyến tính trong không gian đặc trưng.
* Kiểm thử tập dữ liệu Test

# THỬ NGHIỆM VÀ ĐÁNH GIÁ MÔ HÌNH

## Phương thức đánh giá

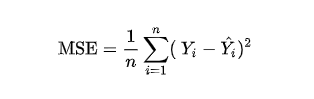
### Mean Absolute Error - MAE

**MAE** là một phương pháp đo lường sự khác biệt giữa hai biến liên tục. Giả sử rằng X và Y là hai biến liên tục thể hiện kết quả dự đoán của mô hình và kết quả thực tế. chúng ta có độ đo **MAE** được tính theo công thức sau:



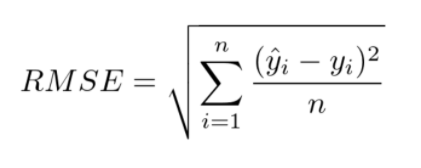
### Mean squared error - MSE

**MSE** của một phép ước lượng là trung bình của **bình phương của sai số**, tức là sự khác biệt giữa các giá trị được mô hình dự đoán và gía trị thực. MSE là một **hàm rủi ro**, tương ứng với giá trị kỳ vọng của sự mất mát sai số bình phương hoặc mất mát bậc hai. **MSE** là **moment bậc hai** (về nguồn gốc) của sai số là moment bậc hai (về nguồn gốc) của sai số



### ****Root Mean Square Error (RMSE)****

RMSE cho biết mức độ phân tán các giá trị dự đoán từ các giá trị thực tế. Công thức tính RMSE là



### R2 Square

R bình phương cho biết mô hình đó hợp với dữ liệu ở mức bao nhiêu %.



## Thử nghiệm mô hình

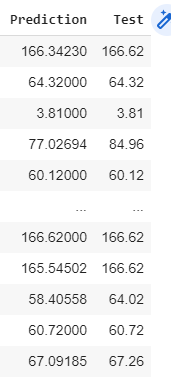
### Thử nghiệm các mô hình khác nhau



### Thử nghiệm mô hình Random Forest



* Kết quả so sánh giữa giá trị predict và giá trị thử nghiệm



# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

## Kết Luận

* Sau khi thử nghiệm các mô hình khác nhau, để chọn mô hình tốt nhất cho bài toán predict ta thấy mô hình Random Forest cho ta kết quả tốt nhất dựa trên các tiêu chí so sánh. Chính vì thế ta lựa chọn Random Forest làm phương pháp chính cho bài toán dự đoán giá .
* Với các tập dữ liệu khác nhau, thuật toán Random Forest cho kết quả dự đoán khá tốt.
* Sử dụng Random Forest để trainning mô hình và dự đoán giá bán của các sản phẩm

## Hướng Phát Triển

Dựa vào giá cả dự đoán ta có thể thấy giá cả của mặt hàng rượu các loại sẽ gia tăng từ 0-10% giá trị sản phẩm trong các tháng tiếp theo. Chính vì thế ta vẫn sẽ tiếp tục nhập các mặt hàng này và bày bán chúng bình thường.

Với một số sản phẩm ta nên tạo theo combo để bán sẽ giúp gia tăng doanh thu hơn nữa.

Để tiếp tục phát triển ta sẽ tiếp tục thu thập thêm dữ liệu và chạy định kì model để cải tiến khả năng học của model. Ngưỡng đảm bảo hiệu xuất dự đoán 80% của model sẽ được kiểm tra hàng quý nhằm cải tiến và phát triển.

TÀI LIỆU THAM KHẢO

[1] <https://machinelearningcoban.com/>

[2] <https://phamdinhkhanh.github.io/2019/09/16/VisualizationPython.html>