## TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN TP. HCM KHOA CÔNG NGHỆ THÔNG TIN

**DATA MINING**

**BÁO CÁO LAB2**

|  |  |
| --- | --- |
| **SINH VIÊN THỰC HIỆN:**  **1712919 Lê Văn Vũ**  **1712502 Trần Quang Huy** | **GV LÝ THUYẾT/ HD THỰC HÀNH:**  **Thầy Lê Hoài Bắc**  **Thầy Nguyễn Ngọc Đức**  Thầy **Dương Nguyễn Thái Bảo** Thầy **Hoàng Xuân Trường** |

**MỤC LỤC**

[I. Data: 3](#_Toc17556)

[II. Phân tích dữ liệu ( EDA ) và các tác cụ tiền xử lí. 4](#_Toc10595)

[1. Phân tích dữ liệu 4](#_Toc18990)

[a. Mức độ tương quan giữa các loại các thuộc tính 4](#_Toc29080)

[b. Mối liên hệ của những thuộc tính rời rạc với thuộc tính phân lớp ( churn ) 5](#_Toc25463)

[2. Các tác cụ tiền xử lí: 7](#_Toc3388)

[a. Loại bỏ các thuộc tính không mang giá trị sinh luật. 7](#_Toc21199)

*[b.](#_Toc6693)* [Phân lớp các thuộc tính liên tục: 8](#_Toc6693)

[III. Code: 9](#_Toc6630)

[IV. Experiments 9](#_Toc32647)

[1. Mục đích phân tích dữ liệu: 9](#_Toc11623)

[2. Thực nghiệm 9](#_Toc7983)

[V. Tóm tắt kết quả: 12](#_Toc22748)

[1. Cách đánh giá kết quả: 12](#_Toc32763)

[2. (Những) tập luật tốt nhất thu được: 13](#_Toc13450)

[3. Ưu và nhược điểm trong bài tập: 13](#_Toc30485)

[a. Ưu điểm 13](#_Toc15407)

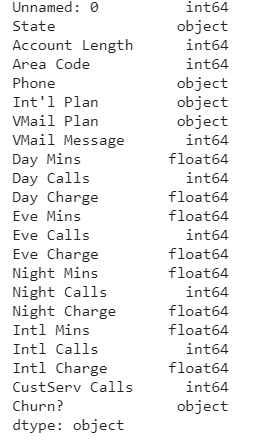
[b. Nhược điểm: 13](#_Toc2447)

[VI. Nguồn ham khảo: 14](#_Toc10258)

# **Data:**

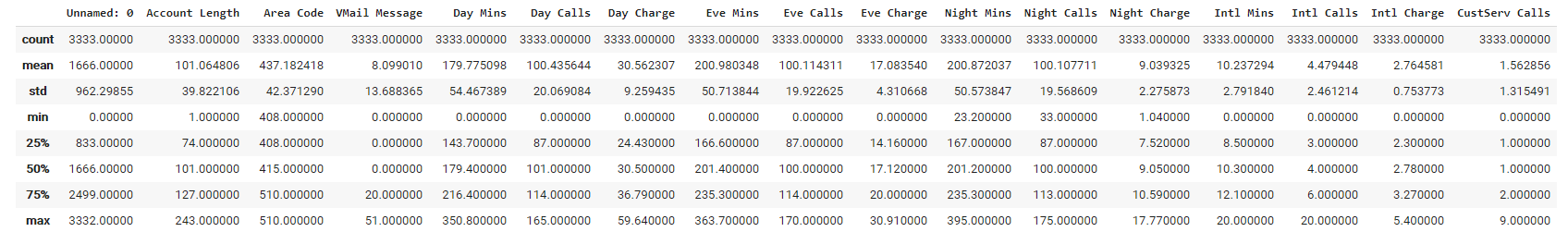
* Tập dữ liệu churn.txt có 3333 mẫu, 21 thuộc tính:
  + State: Các bang thuộc Columbia. Thuộc tính định danh.
  + Account length: Thời gian tính từ kích hoạt tài khoản. Số nguyên
  + Area Code: mã khu vực. Thuộc tính định danh.
  + Phone: Số điện thoại của khách hàng. Dùng để làm ID cho khách hành.
  + Vmail Plan: Có sử dụng dịch vụ Vmail hay không? Thuộc tính nhị phân.
  + Intl Plan: Có sử dụng dịch vụ quốc tế hay không? Thuộc tính nhị phân.
  + Vmail Message: Số tin nhắn. Số nguyên.
  + Day Mins: Số phút gọi trong ngày. Số nguyên
  + Day Calls: Số cuộc gọi trong ngày. Số nguyên.
  + Day Charge: Phí gọi trong ngày. Số nguyên.
  + Eve Mins: Số phút gọi vào buổi chiều. Số nguyên
  + Eve Calls: Số cuộc gọi vào buổi chiều. Số nguyên.
  + Eve Charge: Phí gọi vào buổi chiều. Số nguyên.
  + Night Mins: Số phút gọi vào ban đêm. Số nguyên
  + Night Calls: Số cuộc gọi vào ban đêm. Số nguyên.
  + Night Charge: Phí gọi vào ban đêm. Số nguyên.
  + Intl Mins: Số cuộc phút gọi quốc tế. Số nguyên.
  + Intl Calls: Số cuộc gọi quốc tế. Số nguyên.
  + Intl Charge: Phí gọi quốc tế. Số nguyên,
  + CustServ Calls: Cuộc gọi đến dịch vụ chăm sóc khách hàng. Số nguyên.

(kết quả khi dùng dtype:



🡺Thuộc tính lớp là : **Churn?**.là thuộc tính nhị phân: cho biết khách hàng đó có ngưng sử dụng dịch vụ hay không?

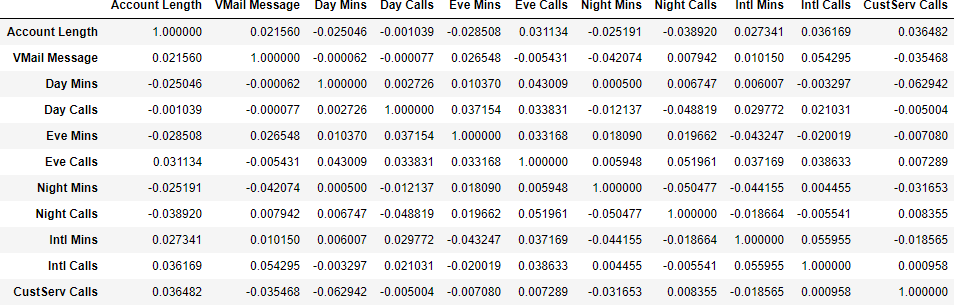
Mô tả dữ liệu:



# **Phân tích dữ liệu ( EDA ) và các tác cụ tiền xử lí.**

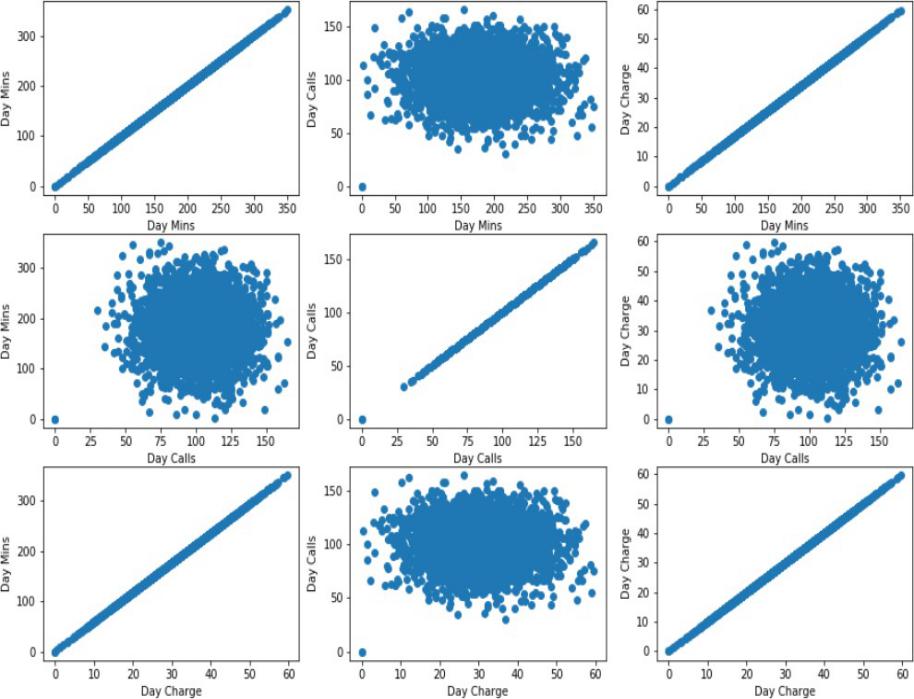
* 1. **Phân tích dữ liệu**
     1. Mức độ tương quan giữa các loại các thuộc tính

Giá trị từ nằm trong khoảng 0 -> 1 thể hiện 2 thuộc tính có mối tương quan tăng dần. Ngược lại trong khoảng -1 <- 0 thể hiện 2 thuộc tính không có mối tương quan tăng dần.



Mức độ tương quan giữa các thuộc tính

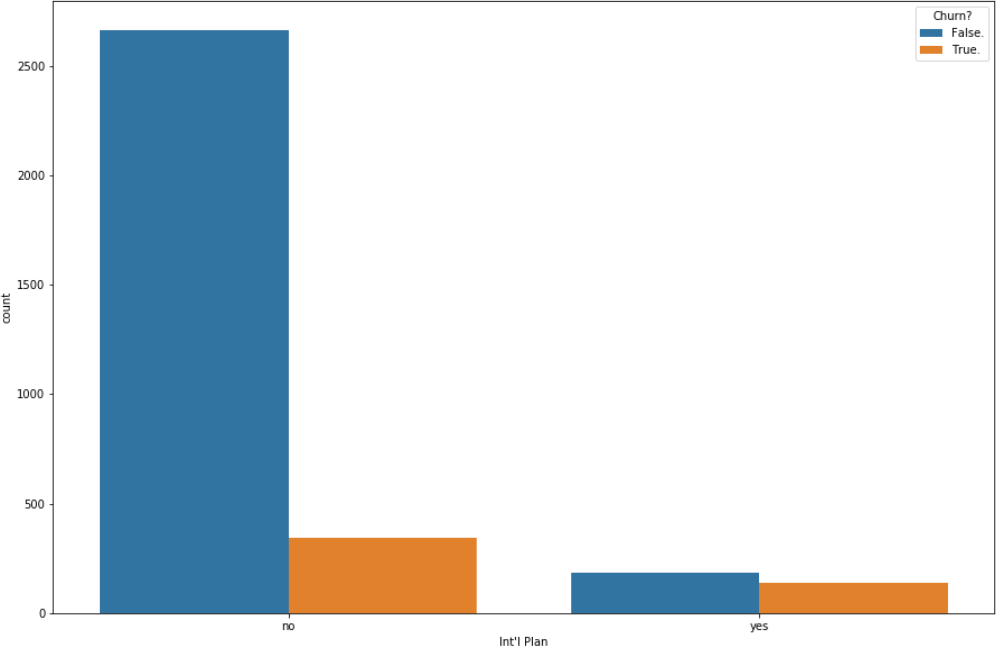
*Vd: T*ương quan giữa Day Mins, Day Calls, Day charge

Nhìn vào biểu đồ trên, ta thấy được mối liên hệ giữa Day Mins và DayCharge là tuyến tính. Số cuộc gọi trong ngày càng nhiều thì số tiền phải chi trả càng nhiều → Việc này là hiển nhiên và không mang lại giá trị ngầm nào cần khai thác từ tập dữ liệu.

Biểu đồ thể hiện mối quan hệ giữa Day Mins, Day Calls và Day Charge

## → Chúng ta loại bỏ những trường charge trong tập dữ liệu.

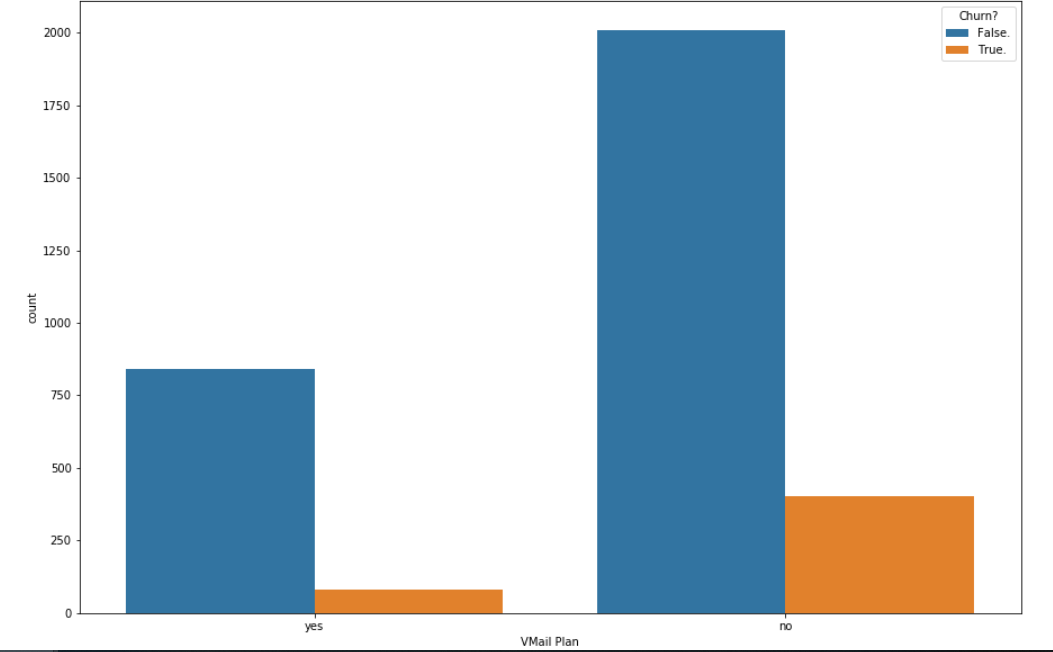
* + 1. Mối liên hệ của những thuộc tính rời rạc với thuộc tính phân lớp ( churn):
* **International Plan:**



Quan hệ giữa Int'l Plan với Churn?

Biểu đồ cho ta thấy đa phần những khách hàng có sử dụng International Plan có xu hướng rời khỏi dịch vụ của công ty cao hơn. Theo thống kê, Có 323 người dùng sử dụng International Length này thì có tới 137 người sẽ rời khỏi dịch vụ, chiếm 42,4%. Trong khi những khách hàng không sử dụng International Length thì tỉ lệ này chỉ là 14,5%.

Từ đó ta chú tập trung vào nhóm khách hàng có sử dụng International Plan để đưa ra những ưu đãi, khuyến mãi để giữ chân những khách hàng này.

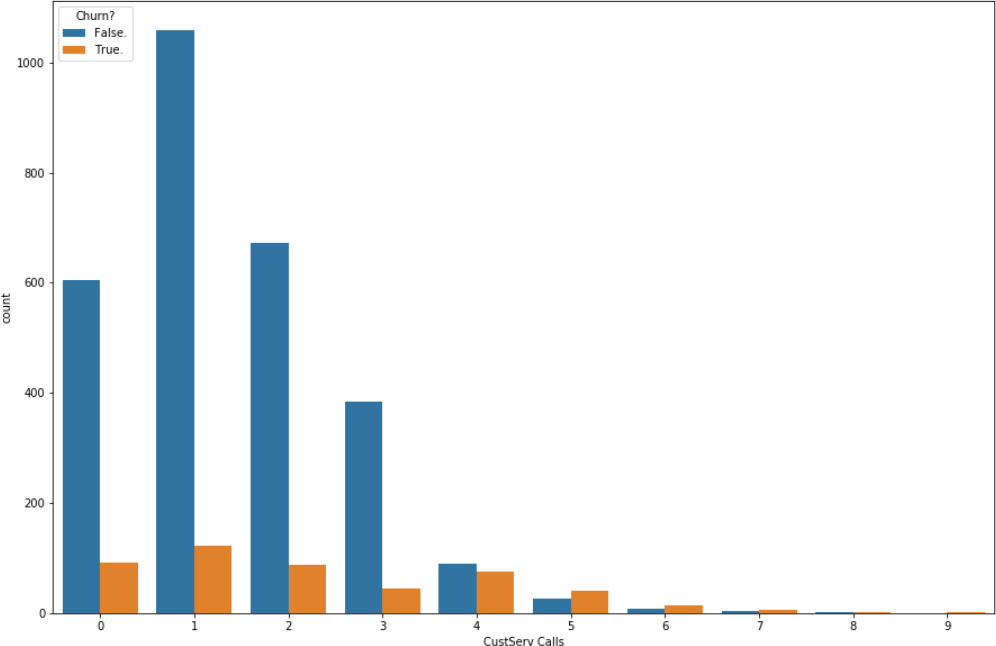
* **VoiceMail Plan:**

Biểu đồ để hiện quan hệ giữa Vmail Plan với Churn?

Biểu đồ trên cho ta thấy nhưng tỉ lệ khách hàng sử dụng VoiceMail Plan có rời khỏi dịch vụ thấp hơn so với tỉ lệ khách hàng không sử dụng VoiceMail Plan. Cụ thể 16,7% tỉ lệ khách hàng không sử dụng VoiceMail Plan là churn so với 8.7% của những khách hàng có sử dụng VoiceMail Plan.

Điều này chỉ ra rằng nếu chúng ta nâng cao, làm cho khách hàng dễ dàng tiếp cận với dịch vụ VoiceMail Plan thì có khả năng nâng cao lượng khách hàng trung thành.

* **Number of calls to customer service ( CustServ Calls )**

Biểu đồ thể hiện mối quan hệ giữa CustServ Calls với Churn

Biểu đồ cho chúng ta thấy là tỉ lệ khách hàng mà thực hiện số cuộc gọi tới dịch vụ khách hàng từ 4 lần trở lên thường là những người sẽ rời bỏ dịch vụ. Mục tiêu là chúng ta theo dõi được số cuộc tới của mỗi khách hàng ( có số cuộc gọi tới dưới 3 lần ) để có thể cung cấp những đãi ngộ thích hợp để giữ chân những khách hàng này tiếp tục sử dụng dịch vụ.

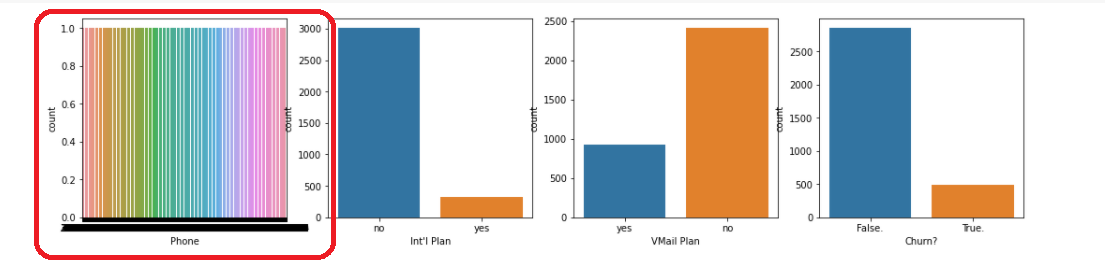
* 1. **Các tác cụ tiền xử lí:**
     1. Loại bỏ các thuộc tính không mang giá trị sinh luật.

Các thuộc tính không mang lại giá trị sinh luật:

* + - * **State:** Là trường dị thường. Không hiểu được ý nghĩa.
      * **Account Length:** không có mối mối quan hệ rõ ràng với churn.
      * **Area Code:** Là trường dị thường chỉ mang 3 giá trị 408, 415, 510.
      * **Phone:** Không mang ý nghĩa.
      * Như đã đề cặp ở trên các trường Charge không mang ý nghĩa nên ta loại đi, các trường này bao gồm:
        + **Day Charge**
        + **Eve Charge**
        + **Night Charge**
        + **Intl Charge**

***Ví dụ:***Thuộc tính Phone khi cho quá nhiều giá trị không có ý nghĩa so vớicac thuộc tính khác:

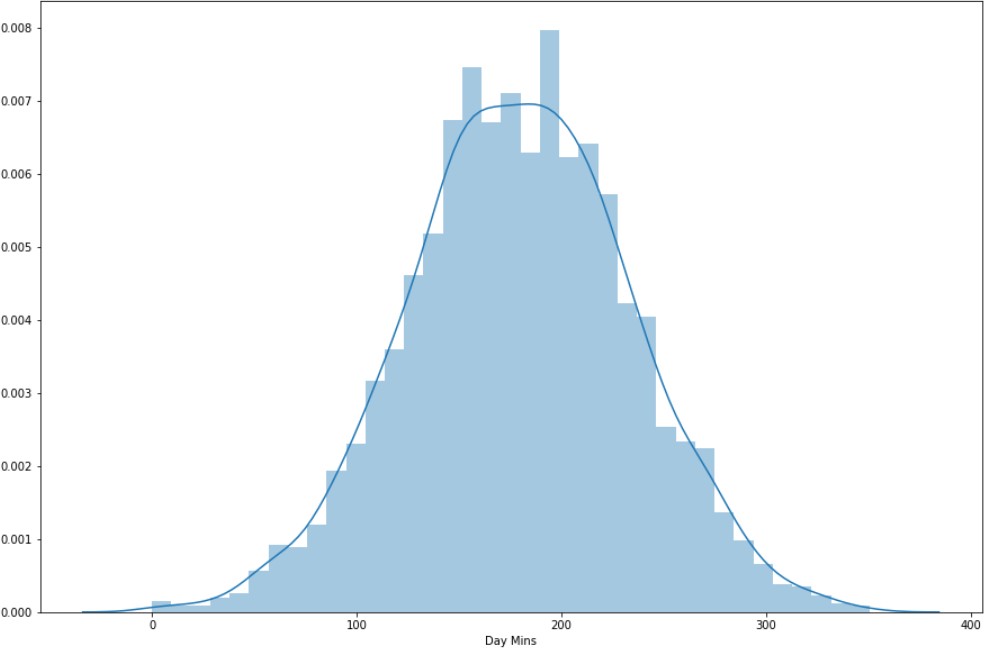
(hình dưới)



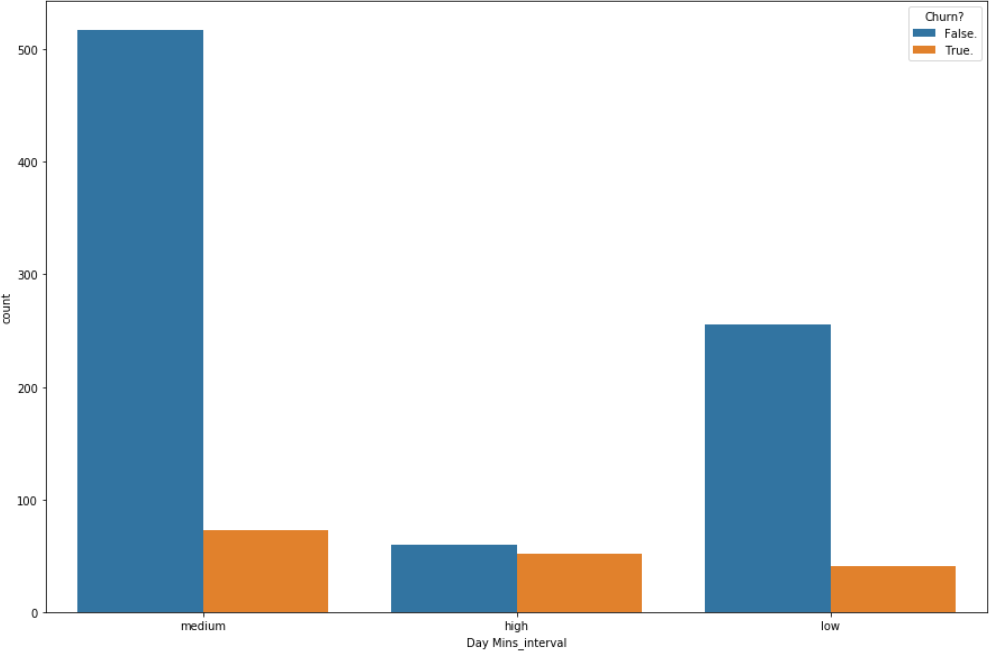
* + 1. Phân lớp các thuộc tính liên tục:

Dựa vào mục *Mô tả dữ liệu: phần* ***I. Data***

* Chia các thuộc tính Day Mins, Day Calls, Eve Mins, Eve Calls, Night Mins, Night Calls thành 3 phần là low, median và high.

VD: với cách phân bố của thuộc tính Day Mins như sau, cột y là tần số xuất hiện, cột x là giá trị số phút gọi. Ta chia làm 3 khoảng, low=[0, 150), median=[150, 250), high=[250, max].

*Hình 7: Tương quan giữa Day Mins với phân phối chuẩn.*

* Thuộc tính Customer service calls (như *hình 7* ) thành 2 khoảng giá trị low, high. Low=[0,3], high=[4,max].
* Thuộc tính Day Mins:

Quan hệ giữa Day Mins với Churn?

🢥 Từ biểu đồ ta thấy, với số phút gọi trong ngày cao thì tỉ lệ cao khách hàng này sẽ rời bỏ dịch vụ.

🡺 Ta chỉ xét trường hợp số phút gọi trong ngày cao.

# **Code:**

- File p**process.py** nhận đầu vào là file churn.txt, đầu ra là file: **ChurnProcessed.csv** chỉ chứa các thuộc tính và mẫu cần thiết để đưa vào Weka tìm ra các tập luật sinh.

* Chạy bằng tham số dòng lệnh: p**process.py churn.txt**

# **Experiments**

* 1. **Mục đích phân tích dữ liệu:**

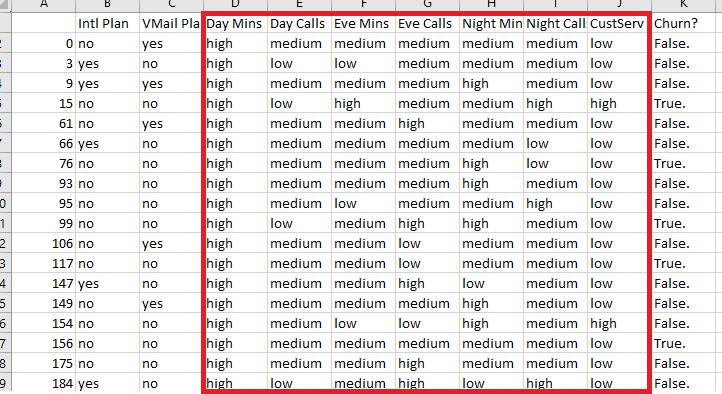
Từ những phân tích dữ liệu trên, những luật kết hợp có ý nghĩa mà chúng ta muốn sinh ra ( một số kết quả có tính chất bắc cầu dẫn đến người đó có khả năng là churner ):

* Những luật dẫn đến người đó thực hiện cuộc gọi đến dịch vụ khách hàng.
* Những luật dẫn đến người đó sử dụng dịch vụ International plan.
* Những luật dẫn đến người đó sử dụng dịch vụ VoiceMail Plan ( sử dụng ở độ tin cậy thấp )
* Những luật dẫn đến người đó có số phút gọi trong ngày cao.
* Những luật dẫn trực tiếp đến 1 người có khả năng là churn. *( rất khó trong tập dataset bị mất cân bằng )*
  1. **Thực nghiệm**

## Các phương pháp tiền xử lí

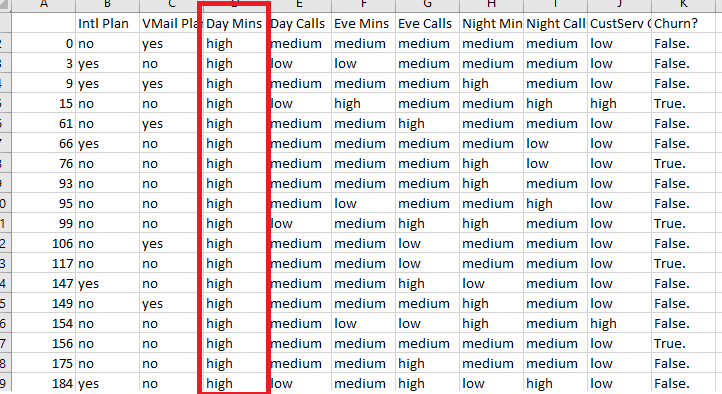
* Xóa các thuộc tính không cần thiết: State, Accound Length, VMail Message, Day Charge, Eve Charge, Night Charge, Intl Charge. Các chứng minh đã làm ở phần [phân tích dữ liệu](#_bookmark3).
* Phân lớp cho các thuộc tính: Day Mins, Day Calls, Eve Mins, Eve Calls, Nignt Mins, Nignt Calls. Đã trình bày ở [phần gom nhóm dữ liệu trong các thuộc tính thành các nhóm](#_bookmark14).

🡺 Được thực hiện trong file **pprocess.py**



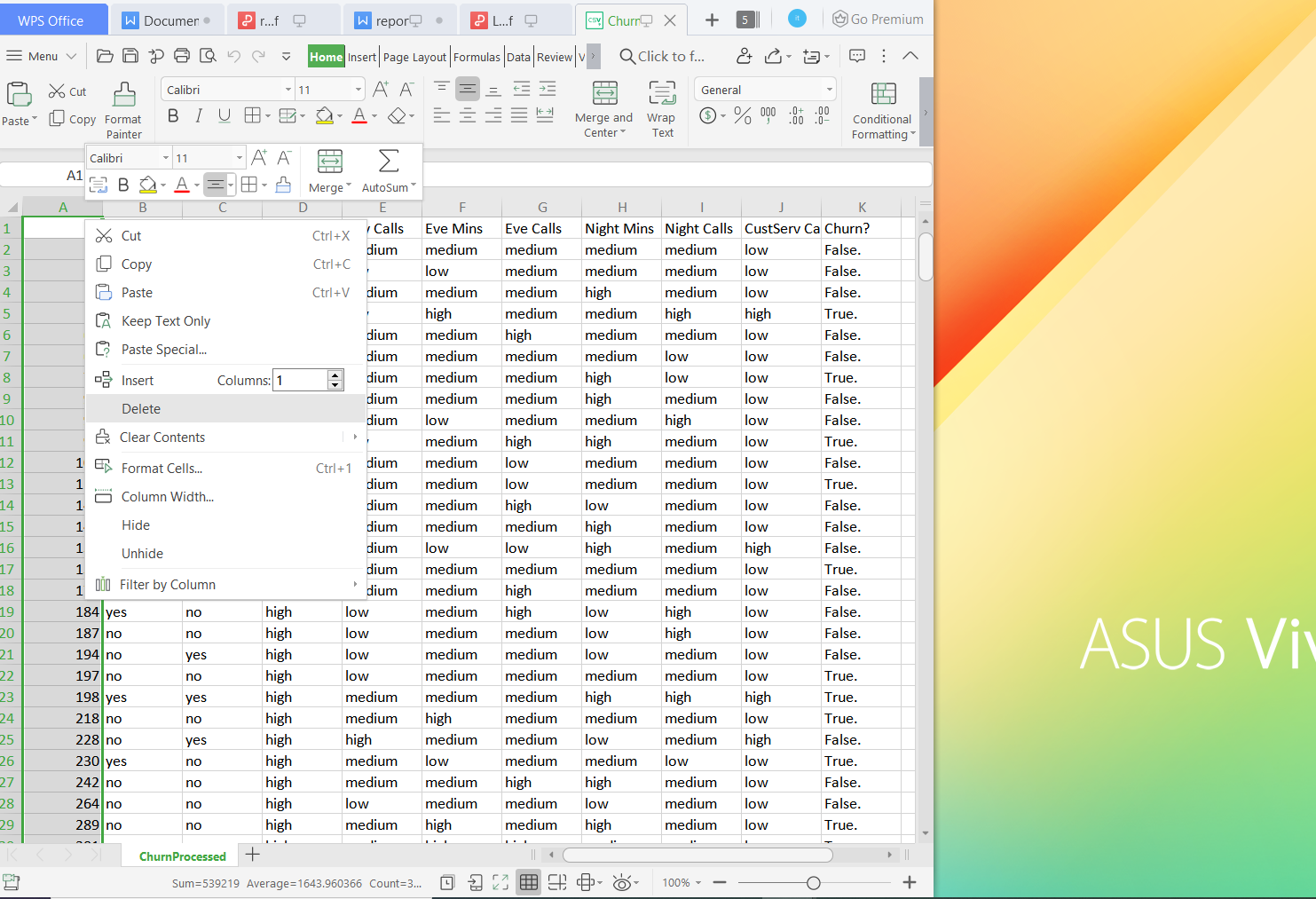
Các thuộc tính **Day Mins, Day Calls, Eve Mins, Eve Calls, Night Mins và Night Calls** đã được phân lớp.

* Lấy những dữ liệu có thuộc tính Day Mins =high.



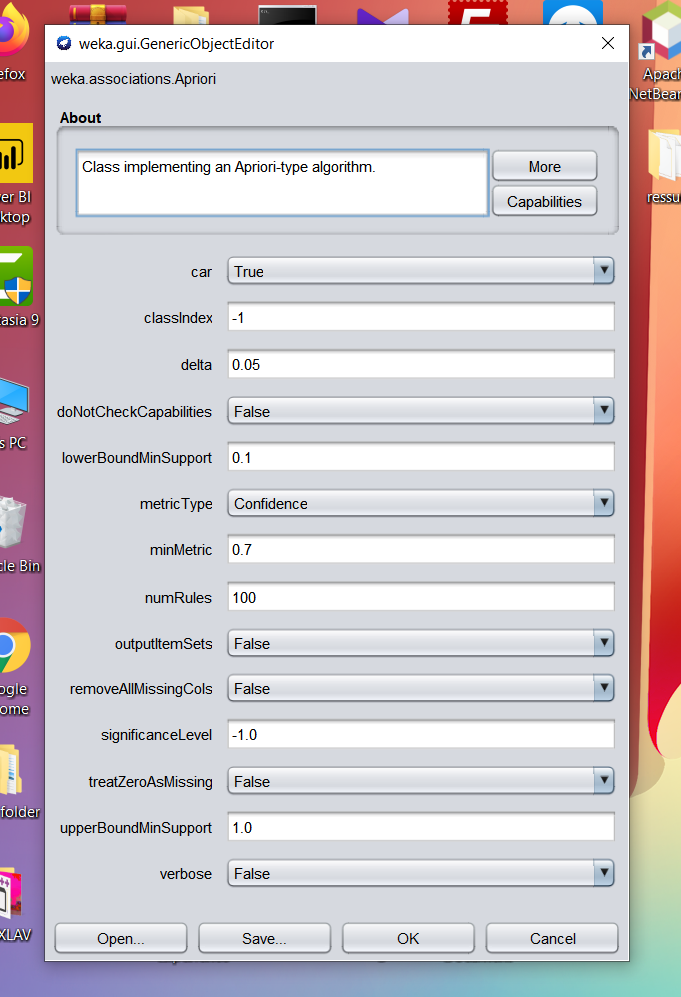
Thuộc tính Day Mins chỉ lấy các mẫu có **Day Mins= high**

🢥 Do sử dụng python để xuất ra file .csv nên ta cần xóa thuộc tính số thứ tự trong file ***ChurnProcessed.csv***



*Xóa cột đầu tiên trong file* ***ChurnProcessed.csv***

## \* Điều chỉnh các tham số:



**- Car**: Chọn True để sinh ra các luật hướng về thuộc tính lớp (churn)

**- numRules**: số luật kỳ vọng sinh ra (trên thực tế nhều lúc numRules sinh ra thâp hơn kỳ vọng *(có thể là rất nhiều)*).

**metricType**: Loại độ đo được sử dụng: Confidence, lift, leverage, convinction. Ở đây mình dùng độ đo Confidence.

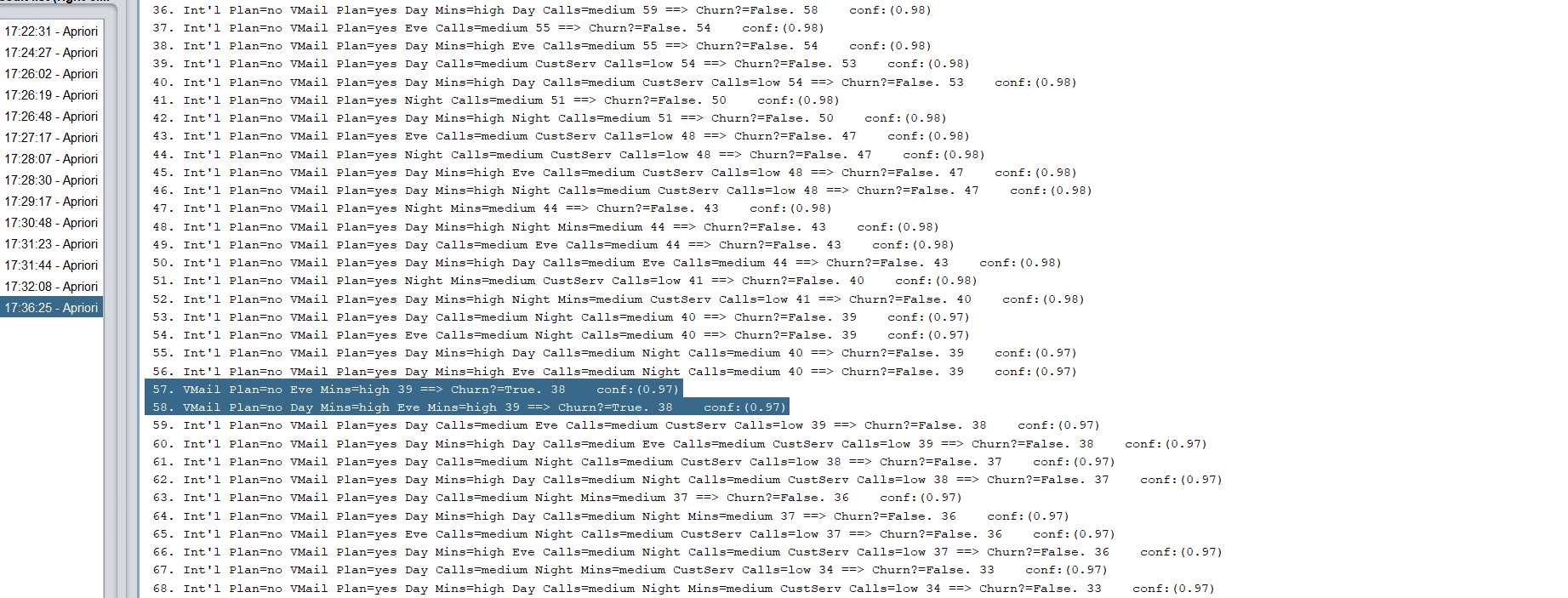
**- lowBoundMinSupport**: độ hỗ trợ nhỏ nhất.

## \*Kết quả phân tích:

* Nhận xét quan trọng:

So với khi chưa xử lý với file process.py thì số luật thu được thấp hơn rất nhiều so với kỳ vọng.

🡺 Sau khi trải qua bước xử lý thì số luật thu được đạt số lượng RẤT TỐT so với kỳ vọng:



Kết quả phân tích khi chạy với thuật toán Apriori

Dưới đây là các tập luật sinh ra Churn?=True cho Độ tin cậy(conf) cao nhất. (conf với Churn?=True giảm dần trong tập kết quả)

**57.** VMail Plan=no Eve Mins=high 39 ==> Churn?=True. 38  **conf:(0.97)**

**58.** VMail Plan=no Day Mins=high Eve Mins=high 39 ==> Churn?=True. 38 **conf:(0.97**)

**135**. VMail Plan=no Night Mins=high CustServ Calls=low 42 ==> Churn?=True. 36 **conf:(0.86)**

**136.** VMail Plan=no Day Mins=high Night Mins=high CustServ Calls=low 42 ==> Churn?=True. 36 **conf:(0.86)**

## \*Ý nghĩa

Dựa vào kết quả trên đều cho thấy rằng. Nếu Số phút gọi ban đêm (Nigh Mins) và số phút gọi vào ban ngày (Day Mins) cao (high) thì khả năng rất cao là khách hàng này sẽ rời khỏi dịch vụ của công ty.

🡺 ***Cần có những ưu đãi với những người có số phút gọi cao để giữ khách hàng ở lại.***

1. **Tóm tắt kết quả:**
   1. **Cách đánh giá kết quả:**

Đánh giá dựa trên độ tin cậy của luật. Độ tin cậy được tính bằng cách: Giả sử ta có luật X-> Y thì độ tin cậy sẽ là:



Nếu độ tin cậy của luật phải lớn hơn độ tin cậy nhỏ nhất đã chỉ định từ trước thì nghĩa là kết quả có thể tin tưởng được:  
***confident ≥ minconfident***

* 1. **(Những) tập luật tốt nhất thu được:**

**57.** VMail Plan=no Eve Mins=high 39 ==> Churn?=True. 38 conf:(0.97)

**58.** VMail Plan=no Day Mins=high Eve Mins=high 39 ==> Churn?=True. 38 conf:(0.97)

**135.** VMail Plan=no Night Mins=high CustServ Calls=low 42 ==> Churn?=True. 36 conf:(0.86)

**136.** VMail Plan=no Day Mins=high Night Mins=high CustServ Calls=low 42 ==> Churn?=True. 36 conf:(0.86)

## \*Ý nghĩa:

* Luật **57**: Nếu không sử dụng dịch vụ Vmail và số phút gọi vào ban đêm cao thì người này rất có khả năng rời khỏi dịch vụ.
* Luật **58**: Nếu khách hàng này không dùng dịch vụ Vmail và có số phút gọi vào ban ngày và ban đêm cao thì khách hàng này khả năng cao sẽ từ bỏ sử dụng dịch vụ.
* Luật **135**: Nếu khách hàng này không dùng dịch vụ Vmail và có số phút gọi vào ban đêm và gọi đến dịch vụ CSKH cao thì khách hàng này khả năng cao sẽ từ bỏ sử dụng dịch vụ.
* Luật **136**: Nếu khách hàng này không dùng dịch vụ Vmail và có số phút gọi vào ban ngày&đêm và gọi đến dịch vụ CSKH cao thì khách hàng này khả năng cao sẽ từ bỏ sử dụng dịch vụ.
  1. **Ưu và nhược điểm trong bài tập:**

1. Ưu điểm

* Có khả năng dùng python để mô phỏng các quan hệ của dữ liệu, biến đổi và chỉnh sửa dữ liệu.
* Sử dụng được thuật toán Apriori để phát sinh luật.
* Phát hiện ra thuộc tính lớp bị mất cân bằng và điều chỉnh bằng cách lấy một phần dữ liệu cân bằng để sinh luật.
* Phát hiệu và loại bỏ các thuộc tính “nhiễu”.

1. Nhược điểm:

- ‘Chưa thực sự’ hiểu rõ bản chất toán học của các thuật toán sinh luật.

- Thời gian để hiểu rõ nội dung dữ liệu thực tế còn lâu.

- Đọc và ‘hiểu’ tài liệu tiếng anh chỉ mới dừng lại ở mức cơ bản.

- Chưa sinh ra đươc luật với các thuật toán quan trọng khác (nhưng vẫn cho ra kết qua cần hướng đến với Apriori).

# **Nguồn ham khảo:**

*<http://bis.net.vn/forums/p/384/675.aspx>*

*<https://github.com/IBM/telco-customer-churn-on-icp4d>*

*<https://medium.com/learning-intelligence/understanding-and-implementation-of-apriori-algorithm-with-python-part-2-ff037ceab254>*

Chuỗi bài giảng Khai phá dữ liệu: *<http://ndhcuong.blogspot.com/p/khai-pha-du-lieu.html>*

*và nhiều nguồn khác*

**\_ \_ HẾT\_ \_**