# TRƯỜNG ĐẠI HỌC THỦ DẦU MỘT

# Giới thiệu - Đại học Thủ Dầu MộtKHOA KỸ THUẬT CÔNG NGHỆ

# BÀI TẬP LỚN MÔN PHÂN TÍCH THỐNG KÊ

# Đề tài

# PHÂN TÍCH ĐIỂM CƠ SỞ NGÀNH CÔNG NGHỆ THÔNG TIN

# SVTH: HOÀNG KIM TUYẾN

# Mã SV: 1824801040043

# Lớp: D18HT01

# Tháng 07/2020

# LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của TS. Bùi Thanh Hùng. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây.

Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong báo cáo còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung báo cáo của mình.** Trường Đại học Thủ Dầu Một không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*Bình Dương, ngày 23 tháng 07 năm 2020*

*Người thực hiện  
 (ký tên và ghi rõ họ tên)*

**MỤC LỤC**

[LỜI CAM ĐOAN 1](#_Toc46406420)

[PHẦN 1 1](#_Toc46406421)

[1.1 Bài 1 1](#_Toc46406422)

[1.2 Bài 2 1](#_Toc46406423)

[1.3 Bài 3 2](#_Toc46406424)

[PHẦN 2 4](#_Toc46406425)

[2.1. Mục đích đề tài 4](#_Toc46406426)

[2.2. Câu hỏi nghiên cứu 4](#_Toc46406427)

[2.3. Thu thập dữ liệu 4](#_Toc46406428)

[2.4. Phân tích dữ liệu 4](#_Toc46406429)

[2.5. Dự đoán 8](#_Toc46406430)

[PHẦN 3: TỰ CHẤM 9](#_Toc46406431)

[DANH MỤC TÀI LIỆU THAM KHẢO 10](#_Toc46406432)

# PHẦN 1

* + 1. Bài 1

*Một cửa hàng thực phẩm nhận thấy thời gian vừa qua trung bình một khách hàng mua 99 nghìn đồng thực phẩm trong ngày. Nay cửa hàng chọn ngẫu nhiên 15 khách hàng thấy trung bình một khách hàng mua* ***xy*** *nghìn đồng trong ngày và phương sai mẫu điều chỉnh là s2 = (2 nghìn đồng)2 . Với mức ý nghĩa là y% , thử xem có phải sức mua của khách hàng hiện nay thực sự giảm sút.*

*Với: xy: là 2 số cuối mã sinh viên của em ( xy = 43)*

*y: là số cuối mã sinh viên của em (y =3)*

Giải

Giả sử: Ho: a = 99

a là sức mua của khách hàng hiện.

ao = 99 là sức mua của khách hàng trước đây.

n = 15, , = 43, s =2, α = 3%

Do α = 3% => γ = 97% => = 1.076

= = = 108.44 >

Vậy ta không chấp nhận Ho

Kết luận: Với mức ý nghĩa là 0%, sức mua của khách hàng hiện nay có giảm sút.

* + 1. Bài 2

*Nhà máy sản xuất 100.000 sản phẩm trong đó có 30.000 sản phẩm loại 2, còn lại là sản phẩm loại 1. KCS đến kiểm tra và lấy ra 150 sản phẩm để thử. Trong 2 trường hợp chọn lặp và chọn không lặp. Hãy tính xác suất để số sản phẩm loại 2 mà KCS phát hiện ra:*

1. *Từ 043 đến 043 + 3\*10*
2. *Ít hơn 043 + 10*

*Với xyz: là 3 số cuối mã sinh viên của em*

Giải

Gọi X là số sản phẩm loại 2 có trong 150 sản phẩm đem kiểm tra

Ta có: X : B(150;0,3)

Do n = 150 lớn, p = 0,3 không quá gần 0 và 1

Nên ta xấp xỉ X N(µ = 45, 2 = 31.5)

1. P(43 ≤ X ≤ 73) = = = 0.6406
2. P(0 ≤ X ≤ 53) = = = 0.4236 + 0,5 = 0.9236
   * 1. Bài 3

*Một gia đình có x+y người con. Giả sử xác suất sinh con trai, con gái như nhau. Tính xác suất:*

*a) Không có con trai.*

*b) Có x con trai và y con gái.*

*c) Số trai từ x đến y (nếu y>x) (hoặc từ y đến x (nếu x>y)).*

*Với x, y là 2 số cuối mã sinh viên của em*

Giải

Xác xuất sinh con trai và con gái bằng nhau nên:

P(sinh con trai) = P(sinh con gái) = 0.5

MSSV của em: 18248010400 43 

* x + y = 7

a, Xác xuất không có con trai:

P = 

b, Xác xuất *có x con trai và y con gái ( x =4, y = 3)*

P = 

c, Số trai từ y -> x ( từ 3 đến 4 )

P = 

# PHẦN 2

## 2.1. Mục đích đề tài

Đối với các sinh viên, khi bắt đầu theo học đại học, tất cả sinh viên đều phải học qua các môn cơ sở ngành của mình. Vì thế, trong đề tài này, em quyết định thu thập điểm các môn cơ sở ngành của các sinh viên CNTT – Trường Đại Học Thủ Dầu Một.  
Đề tài này trả lời các câu hỏi sau đây: các sinh viên ngành CNTT có học tốt các môn cơ sở ngành không, có nên có thái độ nghiêm túc trong việc học các môn cơ sở ngành hay không, , mức độ quan trọng của các môn cơ sở ngành đối với sinh viên CNTT sau khi học các môn chuyên ngành, tỉ lệ sinh viên CNTT học tốt các môn cơ sở ngành

## 2.2. Câu hỏi nghiên cứu

* Các sinh viên ngành CNTT có học tốt các môn cơ sở ngành không?
* Có nên có thái độ nghiêm túc trong việc học các môn cơ sở ngành hay không?
* Có nên có thái độ nghiêm túc trong việc học các môn cơ sở ngành hay không?
* Mức độ quan trọng của các môn cơ sở ngành đối với sinh viên CNTT sau khi học các môn chuyên ngành?
* Tỉ lệ sinh viên CNTT học tốt các môn cơ sở ngành?

## 2.3. Thu thập dữ liệu

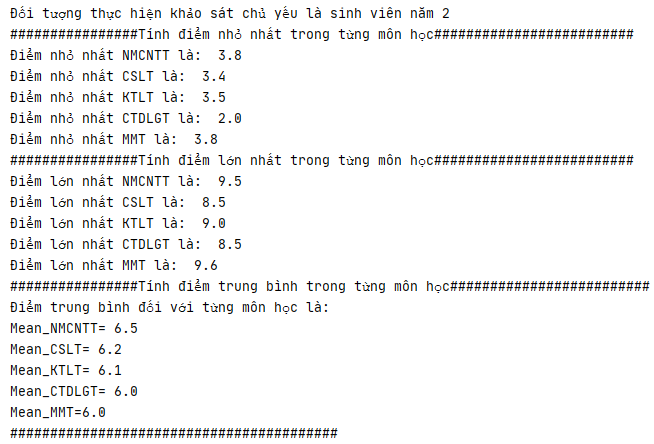
Trong thời gian nhận được yêu cầu đề tài, em đã thu thập dữ liệu bằng cách tạo google Form, sau đó gửi cho các bạn sinh viên trên Facebook, nhờ các bạn trong lớp tiến hành khảo sát. Trong vòng 2 tuần em thu thập được hơn 150 dòng dữ liệu

Link khảo sát: *https://forms.gle/QrLXnjWhZAH9PEht7*

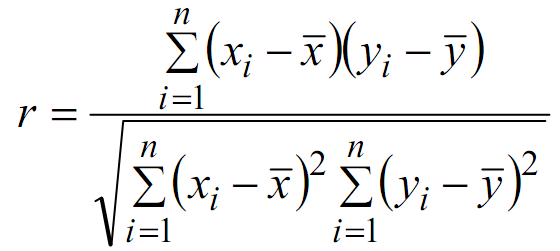
Bộ dữ liệu bao gồm 6 cột, cột thứ nhất chứa sinh viên năm thứ mấy, 5 cột còn lại chứa điểm của sinh viên.

## 2.4. Phân tích dữ liệu

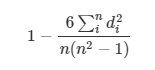
Phân tích giá trị Lớn nhất, Nhỏ nhất, Trung bình, Phần tử xuất hiện nhiều nhất trong mỗi cột



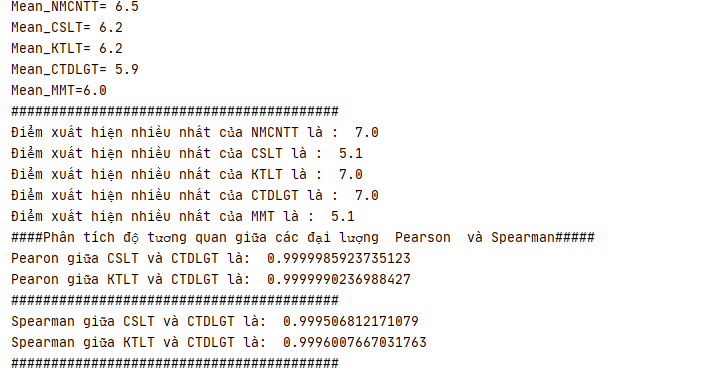
Hệ số tương quan (r) là một chỉ số thống kê đo lường mối liên hệ tương quan giữa hai biến số, như giữa MỨC ĐỘ HÀI LÒNG (y) và TIỀN LƯƠNG (x). Hệ số tương quan có giá trị từ -1 đến 1. Hệ số tương quan bằng 0 (hay gần 0) có nghĩa là hai biến số không có liên hệ gì với nhau; ngược lại nếu hệ số bằng -1 hay 1 có nghĩa là hai biến số có một mối liên hệ tuyệt đối. Nếu giá trị của hệ số tương quan là âm (r <0) có nghĩa là khi x tăng cao thì y giảm (và ngược lại, khi x giảm thì y tăng); nếu giá trị hệ số tương quan là dương (r > 0) có nghĩa là khi x tăng cao thì y cũng tăng, và khi x tăng cao thì y cũng tăng theo.

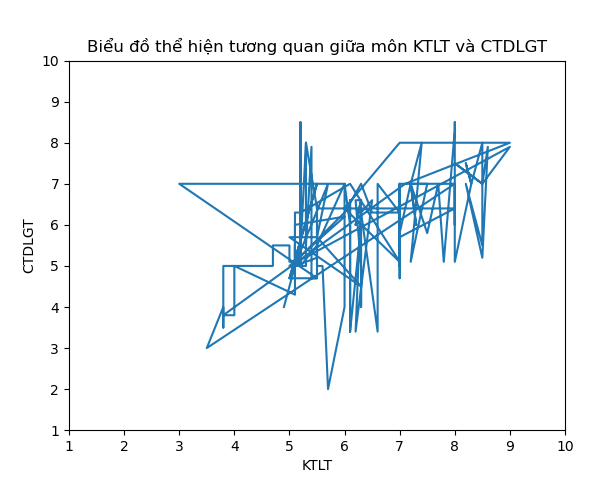
Có nhiều hệ số tương quan, hệ số tương quan thông dụng nhất: hệ số tương quan Pearson r, được định nghĩa như sau: Cho hai biến số x và y từ n mẫu, hệ số tương quan Pearson được ước tính bằng công thức sau đây:

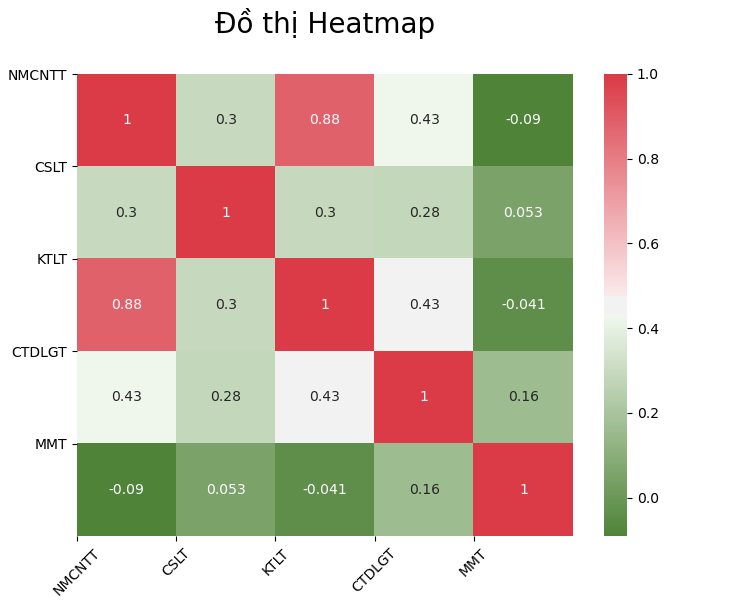
Tương quan hạng (spearman):

* Đánh giá mức độ tương quan của 2 hạng của 2 biến (rank-ordered variables), sử dụng khi phân phối của tổng thể được giả sử không phải là phân phối chuẩn hoặc trong trường hợp có các giá trị quan sát bất thường (lớn quá hoặc nhỏ quá)

di: hiệu hạng của 2 biến







## http://3qeqpr26caki16dnhd19sv6by6v.wpengine.netdna-cdn.com/wp-content/uploads/2014/12/naive-bayes-classifier.jpg2.5. Dự đoán

https://viblo.asia/uploads/a468626e-0831-4efb-b4be-537f5329f050.png

Lý thuyết Bayes thì có lẽ không còn quá xa lạ với chúng ta nữa rồi. Nó chính là sự liên hệ giữa các xác suất có điều kiện. Điều đó gợi ý cho chúng ta rằng chúng ta có thể tính toán một xác suất chưa biết dựa vào các xác suất có điều kiện khác. Thuật toán ***Naive Bayes*** cũng dựa trên việc tính toán các xác suất có điều kiện đó. Nghe tên thuật toán là đã thấy gì đó ngây ngô rồi. Tại sao lại là **Naive** nhỉ. Không phải ngẫu nhiên mà người ta đặt tên thuật toán này như thế. Tên gọi này dựa trên một giả thuyết rằng các chiều của dữ liệu X=(x\_1, x\_2, …., x\_n)X=(x1​,x2​,....,xn​) là độc lập về mặt xác suất với nhau.  Chúng ta có thể thấy rằng giả thuyết này có vẻ khá ***ngây thơ*** vì trên thực tế điều này có thể nói là không thể xảy ra tức là chúng ta rất ít khi tìm được một tập dữ liệu mà các thành phần của nó không liên quan gì đến nhau. Tuy nhiên, giả thiết ngây ngô này lại mang lại những kết quả tốt bất ngờ. Giả thiết về sự độc lập của các chiều dữ liệu này được gọi là Naive Bayes (xin phép không dịch). Cách xác định class của dữ liệu dựa trên giả thiết này có tên là **Naive Bayes Classifier (NBC)**. Tuy nhiên dựa vào giả thuyết này mà bước training và testing trở nên vô cùng nhanh chóng và đơn giản. Chúng ta có thể sử dụng nó cho các bài toán large-scale. Trên thực tế, **NBC** hoạt động khá hiệu quả trong nhiều bài toán thực tế, đặc biệt là trong các bài toán phân loại văn bản, ví dụ như lọc tin nhắn rác hay lọc email spam

Dự đoán bằng phương pháp Bayes tự viết và đánh giá bằng 3 độ đo (Accuracy, Confusion Matrix score và F1 score)

# PHẦN 3: TỰ CHẤM

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nội dung | Ý | Thang điểm | Điểm chấm | Ghi chú |
| Phần 1 | Bài 1 | 1 điểm | 1 |  |
| Bài 2 | 1 điểm | 1 |  |
| Bài 3 | 1 điểm | 1 |  |
| Phần 2 | Dữ liệu | 1 điểm | 1 |  |
|  | Phân tích | 2 điểm | 2 |  |
|  | Độ tương quan | 2 điểm | 1 |  |
|  | Dự báo | 2 điểm | 0 |  |
| **Tổng** | | 10 | 7 |  |

# DANH MỤC TÀI LIỆU THAM KHẢO

1. <https://viblo.asia/p/gioi-thieu-ve-matplotlib-mot-thu-vien-rat-huu-ich-cua-python-dung-de-ve-do-thi-yMnKMN6gZ7P>
2. <https://viettuts.vn/python/mo-dun-statistics-trong-python>
3. <https://viblo.asia/p/gioi-thieu-ve-numpy-mot-thu-vien-chu-yeu-phuc-vu-cho-khoa-hoc-may-tinh-cua-python-maGK7kz9Kj2>