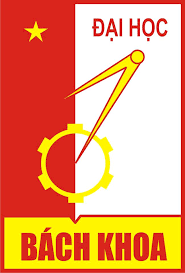
**※ TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI ※**

**VIỆN CÔNG NGHỆ THÔNG TIN VÀ**

**TRUYỀN THÔNG**

**-------------------------------------------------------------------------**

**TIỂU LUẬN BỘ MÔN XÁC SUẤT THỐNG KÊ**



**-------------------------------------------------------------------------**

**Sinh viên: Đỗ Thành Nam**

**MSSV: 20176829**

## I. Thông tin sinh viên và dữ liệu xử lí

Họ tên: Đỗ Thành Nam

MSSV: 20176829

Lớp: Việt Nhật A – K62

Đường link dữ liệu: <https://archive.ics.uci.edu/ml/datasets/Chess+%28King-Rook+vs.+King%29>

Phần mềm sử dụng: Microsoft Excel 2013

## II. Tổng kết, đánh giá tiểu luận

**BẢNG TỔNG KẾT ĐÁNH GIÁ TIỂU LUẬN**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Kết quả Ghi chú (ghi rõ nếu đánh “v” vào kết quả có) | | | | | | |
| Tiêu chí | | |  | Có | Không |  |
| I | Về trình bày | | |  |  |  |
| 1 |  | Căn lề | |  |  |  |
| 2 |  | Header & footer | |  |  |  |
| 3 |  | Mục lục | |  |  |  |
| 4 |  | Bảng dữ liệu nhìn được tiêu đề | |  |  |  |
| 5 |  | Bảng tổng kết | |  |  |  |
| 6 |  | Kết luận | |  |  |  |
| 7 |  | Tài liệu tham khảo | |  |  |  |
| 8 |  | Không sai chính tả | |  |  |  |
| 9 |  | Cỡ chữ <14 | |  |  |  |
| II | Về nội dung | | |  |  |  |
| 1 |  | Đủ nội dung theo yêu cầu đề bài | |  |  |  |
|  |  |  | Số thuộc tính phân tích |  |  | Phân tích … thuộc tính |
| 2 |  | Xử lí thô dữ liệu | |  |  |  |
|  |  |  | Bỏ dữ liệu |  |  |  |
|  |  |  | Thay thế dữ liệu |  |  |  |
|  |  |  | Số hóa dữ liệu dạng chữ |  |  |  |
|  |  |  | Chia khoảng |  |  |  |
|  |  |  | Khác (đề nghị ghi rõ ở cột ghi chú) |  |  |  |
| 3 |  | Vẽ bảng phân phối Pareto | |  |  |  |
| 4 |  | Tính đặc trưng | |  |  |  |
|  |  |  | Kỳ vọng |  |  |  |
|  |  |  | Mod |  |  |  |
|  |  |  | Median |  |  |  |
|  |  |  | Phương sai mẫu |  |  |  |
|  |  |  | Độ lệch chuẩn mẫu |  |  |  |
| 5 |  | Phần cộng điểm | |  |  |  |
|  |  |  | Bảng phân tích tương quan |  |  |  |
|  |  |  | Bảng phân tích ANOVA |  |  |  |
|  |  |  | Dự báo hồi quy |  |  |  |
|  |  |  | Kiểm định phân phối |  |  |  |
|  |  |  | Kiểm định giả thiết cho kỳ vọng |  |  |  |
|  |  |  | Kiểm định giả thiết cho phương sai |  |  |  |
|  |  |  | Phân lớp |  |  |  |

**MỤC LỤC**

[I. Thông tin sinh viên và dữ liệu xử lí 1](#_Toc514291991)

[II. Tổng kết, đánh giá tiểu luận 2](#_Toc514291992)

[III. Đề bài thi xác suất thống kê K62 Việt Nhật 4](#_Toc514291993)

[IV. Giới thiệu về dữ liệu 6](#_Toc514291994)

[V. Phân tích tương quan giữa các thuộc tính 9](#_Toc514291995)

[VI. Phân lớp dữ liệu 10](#_Toc514291996)

[VII. Kết luận 12](#_Toc514291997)

[1. Các việc đã làm 12](#_Toc514291998)

[2. Kiến thức thu được 12](#_Toc514291999)

[VIII. Tài liệu tham khảo 13](#_Toc514292000)

## III. Đề bài thi xác suất thống kê K62 Việt Nhật

Yêu cầu 1. Lấy dữ liệu theo yêu cầu của giáo viên

Nguồn tổng

<http://archive.ics.uci.edu/ml/datasets.html>

hoặc <http://mlr.cs.umacs.edu/ml/datasets.html>

Yêu cầu 2. Sử dụng phần mềm tính được các thông số (tham khảo chi tiết ở yêu cầu 3)

Yêu cầu 3. In bài làm ra giấy A4 (tiểu luận nhỏ), trong đó ghi rõ:

Yêu cầu 4. Nộp vở bài tập ở nhà

Yêu cầu 5. Gặp giảng viên theo lịch hẹn để được thi vấn đáp

CHI TIẾT CHO YÊU CẦU 3

Yêu cầu chung cho tiểu luận

Có bìa tiểu luận, mục lục, kết luận, tài liệu tham khảo

SAI chính tả (<1 lỗi/trang)

Có bình luận các kết quả làm được

Trình bày đẹp

Trang 1

Họ và tên

Mã số sinh viên

Lớp

Đường link dữ liệu được phân

Phần mềm sử dụng để tính toán

Trang 2

Tổng kết đánh giá tiểu luận

Trang 3

Mục lục

Trang 4

Đề bài của mình

Trang 5 (có thể nhiều hơn 1 trang)

Mô tả cấu trúc của dữ liệu (các thuộc tính, số lượng các bản ghi dữ liệu, các thuộc tính dữ liệu, nguồn dữ liệu, sử dụng để….)

Chi tiết bảng dữ liệu

Trang 6-n (có thể nhiều hơn 1 trang)

Phân tích từng thuộc tính (mỗi thuộc tính 1 trang)

Tóm tắt thông tin thuộc tính: loại dữ liệu (liên tục, rời rạc, catalog, nếu liên tục thì min, max, nếu catalog thì nó là những loại gì,…)

Bảng phân phối thực nghiệm + đồ thị

Ước lượng cho kì vọng, mod, med (ước lượng điểm, ước lượng khoảng)

Ước lượng cho phương sai (ước lượng điểm, ước lượng khoảng)

Làm thêm (sẽ được cộng điểm)

Phân tích tương quan 02 thuộc tính

Phân lớp được dữ liệu (với bài toán phân lớp. Gợi ý: sử dụng Bayets)

Dự báo được dữ liệu (với bài toán chuỗi. Gợi ý: sử dụng hồi quy)

Kiểm định được các thuộc tính

Phân tích ANOVA

Trang n+1

Kết luận: Tóm tắt những gì mình làm, nêu những kiến thức thu được

Trang n+2

Tài liệu tham khảo

**Đề bài: Chess (King-Rook vs. King)**

Link dữ liệu: <https://archive.ics.uci.edu/ml/datasets/Chess+%28King-Rook+vs.+King%29>

## IV. Giới thiệu về dữ liệu

**Mô tả dữ liệu:**

Tập dữ liệu thống kê: Chess Endgame Database (vua trắng và xe trắng vs vua đen)

**Cấu trúc dữ liệu**:

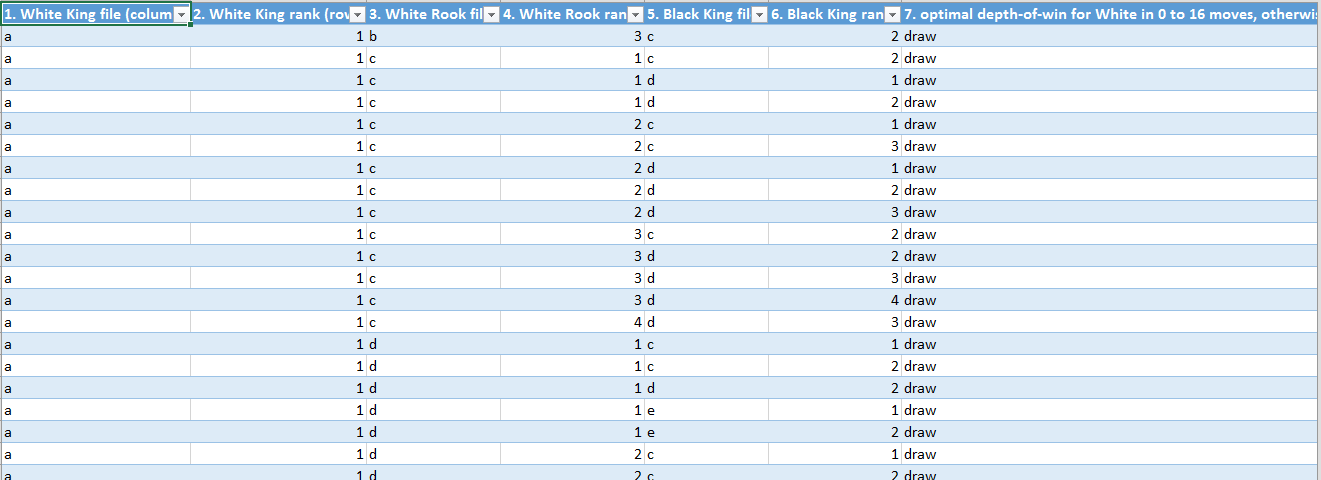
1. Số lượng bản ghi dữ liệu: 28056
2. Số thuộc tính: 6
3. Đặc điểm thuộc tính: Categorical, integer.
4. Nguồn dữ liệu: được tạo bởi Michael Bain và Arthur van Hoff của học viện Turing, Glasgow, UK.
5. Sử dụng để:

Tính toán khả năng chiến thắng tối ưu của cờ trắng từ 0 đến 16 nước đi, hoặc là hòa so với cờ đen (bên trắng: vua và 1 xe, bên đen: vua).

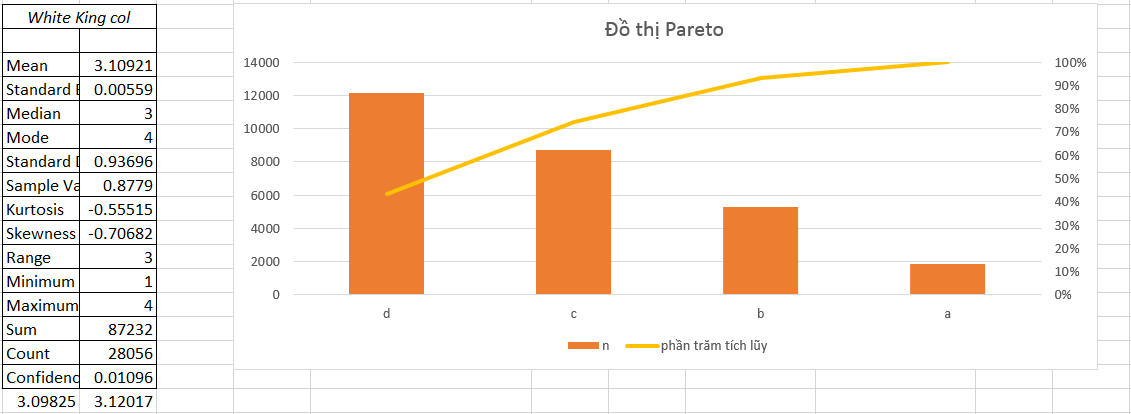
**Chi tiết bảng dữ liệu:**

1. Các thuộc tính trong bảng dữ liệu:

* Từ cột 1 đến cột 6: lần lượt là tọa độ của vua trắng, xe trắng, vua đen trên bàn cờ.

1. Chi tiết bảng dữ liệu:
2. Các thuộc tính phân tích:

* **Phân tích thuộc tính 1 White King column**
* Loại dữ liệu: Catalog (a, b, c, d)
* Bảng phân phối và biểu đồ:



* Điểm kỳ vọng: 3,10921
* Ước lượng cho kì vọng:

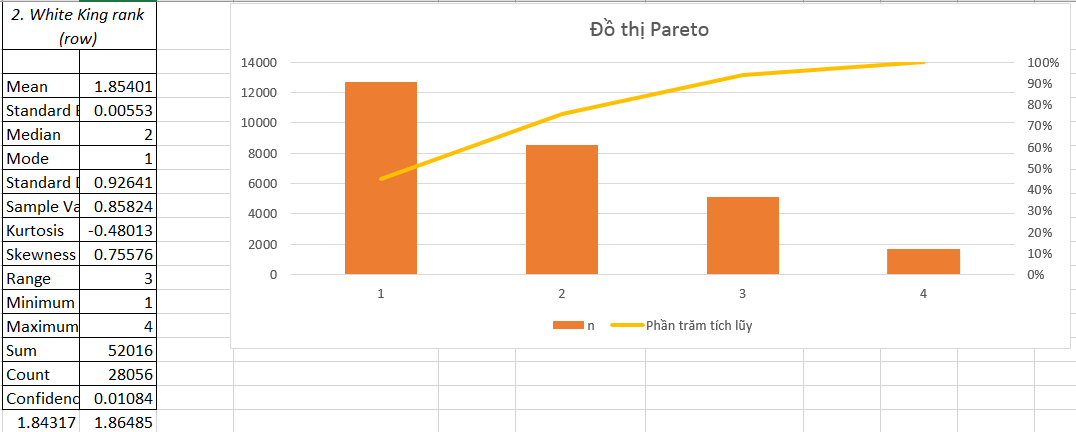
+ Min: 3,09825

+ Max: 3,12017

* Mod: 4
* Med: 3
* Ước lượng cho phương sai:

Ước lượng điểm: 0,8779

* **Phân tích thuộc tính 2: White King row**
* Loại dữ liệu: rời rạc
* Bảng phân phối và biểu đồ:



* Điểm kỳ vọng: 1,85401
* Ước lượng cho kì vọng:

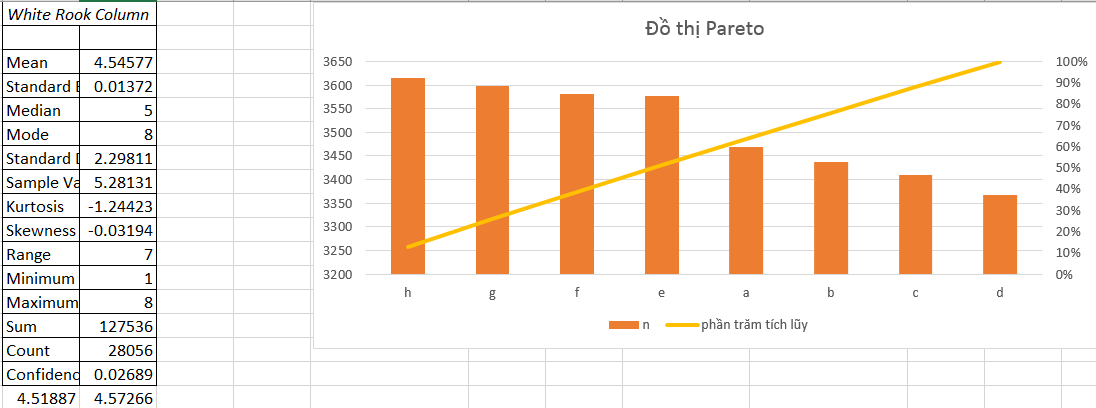
+ Min: 1,84317

+ Max: 1,86485

* Mod: 1
* Med: 2
* Ước lượng cho phương sai:

Ước lượng điểm: 0,85824

* **Phân tích thuộc tính 3 White Rook column**
* Loại dữ liệu: Catalog
* Bảng phân phối và biểu đồ:



* Điểm kỳ vọng: 4,54577
* Ước lượng cho kì vọng:

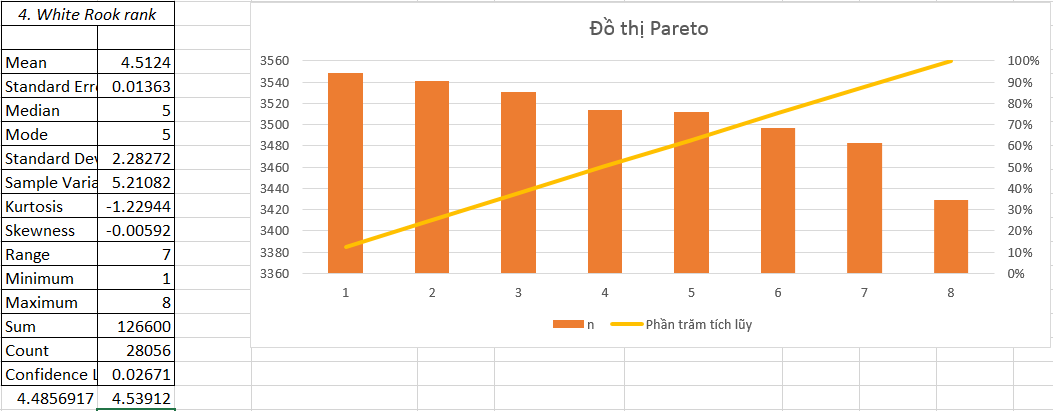
+ Min: 4,51887

+ Max: 4,57266

* Mod: 8
* Med: 5
* Ước lượng cho phương sai:

Ước lượng điểm: 5,28131

* **Phân tích thuộc tính 4 White Rook row**
* Loại dữ liệu: rời rạc
* Bảng phân phối và biểu đồ:



* Điểm kỳ vọng: 4,5124
* Ước lượng cho kì vọng:

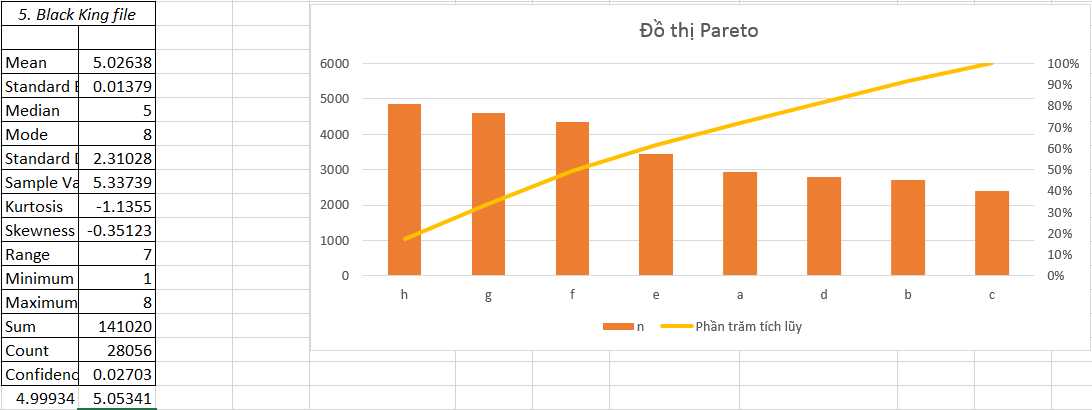
+ Min: 4,4856917

+ Max: 4,53912

* Mod: 5
* Med: 5
* Ước lượng cho phương sai:

Ước lượng điểm: 5,21082

* **Phân tích thuộc tính 5 Black King column**
* Loại dữ liệu: Catalog
* Bảng phân phối và biểu đồ:



* Điểm kỳ vọng: 5,02638
* Ước lượng cho kì vọng:

+ Min: 4,99934

+ Max: 5,05341

* Mod: 8
* Med: 5
* Ước lượng cho phương sai:

Ước lượng điểm: 5,33739

* **Phân tích thuộc tính 6: Black King row**
* Loại dữ liệu: rời rạc
* Bảng phân phối và biểu đồ:



* Điểm kỳ vọng: 4,45181
* Ước lượng cho kì vọng:

+ Min: 4,4255

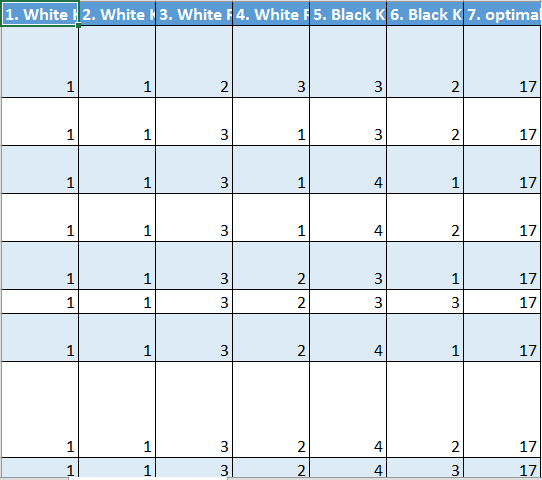
+ Max: 4,47812

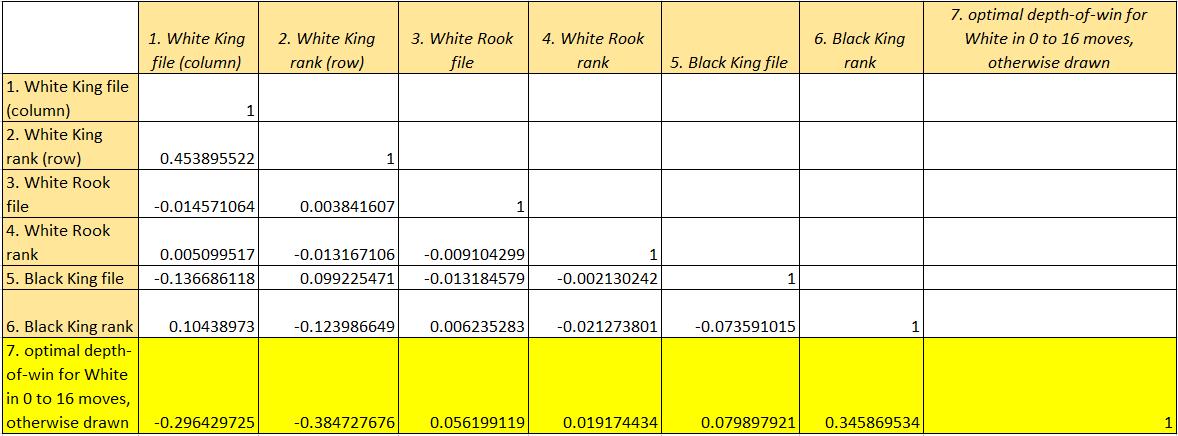
* Mod: 5
* Med: 4
* Ước lượng cho phương sai:

Ước lượng điểm: 5,05524

## V. Phân tích tương quan giữa các thuộc tính

Bảng dữ liệu các thuộc tính được xét:



Bảng tương quan giữa các thuộc tính:

* Dựa vào bảng tương quan ta thấy:

+ R­white king col  = -0,3 < 0 và R­white king rank = -0,38 < 0. Thuộc tính White King column và White King rank có tương quan nghịch với Optimal depth-of-win và ảnh hưởng đáng kể đến Optimal depth-of-win. Nếu tăng giá trị của 2 thuộc tính này thì sẽ mất nhiều nước cờ hơn để bên trắng thắng.

+ R­Black King rank = 0,346: Thuộc tính Black King rank có tương quan thuận với Optimal depth-of-win và ảnh hưởng đáng kể đến Optimal depth-of-win.

* Các thuộc tính White King column, White King rank, Black King rank có thể coi như độc lập với nhau.

## VI. Phân lớp dữ liệu

* **Phân lớp dữ liệu bằng Bayes:**
* Đối với mỗi bản ghi z mới vừa thu được, tính giá trị của biểu thức sau với mỗi phân lớp ci C = {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17}(0->16 tương ứng với số nước, 17 là hòa):

***P(ci).(xj|ci)***

* xj tương ứng là các thuộc tính trình bày trong bảng trên.
* P(c0) = 0.000749 , P(c­1)= 0.002513, P(c2)= 0.008662, P(c­3)= 0.003048, P(c­4)= 0.007218, P(c­5)= 0.016628, P(c6) = 0.021547 , P(c­7)= 0.024435, P(c8)= 0.051436, P(c­9)= 0.05983, P(c­10)= 0.069775, P(c­11)= 0.102443, P(c12) = 0.129658, P(c­13)= 0.149334, P(c14)= 0.161418, P(c­15)= 0.077421, P(c­16)= 0.014062, P(c­17)= 0.099824.
* Xác định phân lớp của z là lớp có thể nhất c\*:

***c\* = argmax P(ci).(xj|ci)***

VD: Bản ghi dữ liệu vừa thu thập được:



+ *Bước 1: Tính toán dựa theo công thức*

* Phân lớp 1:

P(c1).(xj| c1) = 0

* Phân lớp 2:

P(c2).(xj| c2) = 0

……………………………………………………………………………..

* Phân lớp 17:

P(c17).(xj| c17) = 1.55E-06

+ *Bước 2: xác định c\* lớn nhất:*

c\* = argmax P(ci).(xj|ci) = 6.41E-06

*+ Bước 3: kết luận, nghiệm thu kết quả*

Dựa vào kết quả thu được ở bước 2 ta phân bản ghi vừa thu được vào phân lớp 12.

## VII. Kết luận

### Các việc đã làm

Xử lý trên Excel:

* Nhập và xử lý dữ liệu thô.
* Cách tạo bảng phân phối thực nghiệm và vẽ đồ thị Pareto cho các thuộc tính.
* Tính sự tương quan giữa các thuộc tính.

Báo cáo:

* Mô tả cấu trúc dữ liệu.
* Phân tích các thuộc tính đã xử lý trên Excel.
* Phân tích tương quan giữa các thuộc tính.
* Phân lớp dữ liệu.

### Kiến thức thu được

* Biết sử dụng, thao tác cơ bản đối với chương trình tính Excel.
* Xử lý số liệu, ý nghĩa các giá trị trong bảng phân phối.
* Các kiến thức về phân lớp, sử dụng Bayes trong phân lớp.

## VIII. Tài liệu tham khảo

1. Sử dụng Bayes trong Machine Learning – Thân Quang Khoát.

2. Giáo trình xác suất thống kê – Lê Đức Vĩnh.