

TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH

CÔNG NGHỆ THÔNG TIN

---  ---

BÀI TẬP MÔN HỌC

KHAI PHÁ DỮ LIỆU

# Đề Tài

So Sánh Phân Bố Các Trường Đại

Học Trong Khu Vực Và Thế Giới.

Dự Đoán Mô Hình Học Sinh Muốn

Đi Du Học

Sinh viên thực hiện: Văn Công Tuấn

Nguyễn Hoài Nam

Lê Thị Thu Uyên

Lớp: DH20IT01

Tháng 08 năm 2023

MỤC LỤC



Contents

[Đề Tài 1](#_Toc143462293)

[I. Mô Tả Dữ Liệu Bài Toán 3](#_Toc143462294)

[**II.** **Tiền Xử Lý Dữ Liệu** 6](#_Toc143462295)

[1. Thu Thập Dữ Liệu 6](#_Toc143462296)

[2. Sửa chữa dữ liệu 7](#_Toc143462297)

[3. Làm sạch dữ liệu 10](#_Toc143462298)

[4. Chuẩn hóa dữ liệu 12](#_Toc143462299)

[III. Phân cụm các trường đại học 12](#_Toc143462300)

[1. Phân cụm theo châu lục 12](#_Toc143462301)

[a) Phân cụm theo khu vực: 12](#_Toc143462302)

[b) Phân cụm theo tỷ lệ chấp nhận: 14](#_Toc143462303)

[2. Phân cụm theo khu vực 16](#_Toc143462304)

[2.1 Asian 17](#_Toc143462305)

[a) Phân cụm theo từng quốc gia 17](#_Toc143462306)

[**b) Phân cụm theo tỷ lệ chấp nhận** 18](#_Toc143462307)

[2.2 Europe 21](#_Toc143462308)

[3. Phân Theo Quốc Gia (Pakistan) 26](#_Toc143462309)

[a) Phân theo từng vùng của quốc gia 26](#_Toc143462310)

[b) Phân theo tỷ lệ chấp nhận đối với quốc gia 28](#_Toc143462311)

[4. So Sánh Mật Độ Phân Bố Và Nhận Xét 30](#_Toc143462312)

[IV. Cây Quyết Định Trường Đại Học 30](#_Toc143462313)

[V. Nhận Xét Tổng Quan 31](#_Toc143462316)

[**Tài Liệu Tham Khảo** 32](#_Toc143462317)

1. Mô Tả Dữ Liệu Bài Toán
2. Dataset Top 300 Universities of World

Tập dataset này chứa các thông tin về xếp hạng, loại hình quản lý, số lượng sinh viên, điểm số, hỗ trợ tài chính và các chỉ số khác của các trường đại học hàng đầu trên toàn thế giới. Tập dataset gồm 24 cột và 300 hàng dữ liệu. Mỗi hàng trong dữ liệu đại diện cho một trường đại học cụ thể. Các cột đó là:

* Name: Tên của trường đại học.
* World\_Ranking: Xếp hạng của trường đại học trên toàn cầu.
* Count\_World\_Ranking: Vị trí xếp hạng tương đối của trường đại học trên toàn cầu.
* Region\_Ranking: Xếp hạng của trường đại học trong khu vực (châu lục) nơi nó đặt.
* Count\_Region\_Ranking: Vị trí xếp hạng tương đối của trường đại học trong khu vực nơi nó đặt.
* Country\_Ranking: Xếp hạng của trường đại học trong quốc gia.
* Count\_Country\_Ranking: Vị trí xếp hạng tương đối của trường đại học trong quốc gia.
* Region: Khu vực (châu lục) nơi trường đại học đặt.
* Country: Quốc gia nơi trường đại học đặt.
* City/State: Thành phố hoặc tiểu bang nơi trường đại học đặt.
* Control: Loại hình quản lý của trường đại học (công lập, tư nhân, không vụ lợi).
* Acceptance\_Rate: Tỷ lệ chấp nhận của trường đại học (tỷ lệ sinh viên được nhận vào so với số lượng sinh viên đăng ký).
* Publication: Số lượng công bố mà trường đại học đã có.
* Language: Ngôn ngữ chính được sử dụng trong trường đại học.
* Average SAT: Điểm trung bình của kỳ thi SAT (Scholastic Assessment Test) đối với sinh viên nhập học.
* Average ACT: Điểm trung bình của kỳ thi ACT (American College Testing) đối với sinh viên nhập học.
* Aid receiving: Tỷ lệ sinh viên nhận được hỗ trợ tài chính.
* Undergraduates: Số lượng sinh viên đại học.
* Graduates: Số lượng sinh viên sau đại học.
* Highest Degree: Cấp độ học vị cao nhất mà trường đại học cung cấp.
* Citations: Số lượng trích dẫn của công trình nghiên cứu từ trường đại học.
* Average aid awarded: Số tiền trung bình hỗ trợ tài chính nhận được.
* Net average cost to attend: Chi phí trung bình net để tham gia trường đại học (sau khi đã nhận được hỗ trợ tài chính).
* Top\_World\_Ranking: Xếp hạng hàng đầu của trường đại học trên toàn cầu.

1. Dataset Euro2

Tập dataset này bao gồm thông tin về các trường đại học hàng đầu trên thế giới trong khu vực châu Âu. Tập dataset gồm 20 cột và 100 hàng dữ liệu. Mỗi trường đại học được mô tả bởi một hàng trong bảng dữ liệu. Các thuộc tính (cột) trong dữ liệu bao gồm:

* Name: Tên của trường đại học.
* World\_Ranking: Xếp hạng của trường đại học trên toàn thế giới.
* Count\_World\_Ranking (tỷ lệ): Vị trí xếp hạng tương đối của trường đại học trên toàn thế giới.
* Region\_Ranking: Xếp hạng của trường đại học trong khu vực nơi nó đặt.
* Count\_ Region\_Ranking (tỷ lệ): Vị trí xếp hạng tương đối của trường đại học trong khu vực nơi nó đặt.
* Coutry\_Ranking: Xếp hạng của trường đại học trong quốc gia nơi nó đặt.
* Count\_Coutry\_Ranking (tỷ lệ): Vị trí xếp hạng tương đối của trường đại học trong quốc gia nơi nó đặt.
* Coutry: Quốc gia nơi trường đại học đặt.
* City: Thành phố nơi trường đại học đặt.
* Type: Loại hình của trường đại học (Non-profit).
* Count\_Type (tỷ lệ): Tỷ lệ đại diện của loại hình trường đại học trong tập dữ liệu.
* Acceptance\_Rat: Tỷ lệ chấp nhận của trường đại học (tỷ lệ sinh viên được nhận vào so với số lượng sinh viên đăng ký).
* Publication: Số lượng công bố mà trường đại học đã có.
* High\_degree: Học vị cao nhất mà trường đại học đó cung cấp (Doctorate, Master, Bachelor).
* Count\_High\_degree (tỷ lệ): Tỷ lệ đại diện của học vị cao nhất mà trường đại học cung cấp trong tập dữ liệu.
* Web: Địa chỉ trang web của trường đại học.
* Language: Ngôn ngữ được sử dụng trên trang web của trường đại học.
* Phone: Số điện thoại liên hệ của trường đại học.
* Address: Địa chỉ vật lý của trường đại học.
* Top Rank: Xếp hạng hàng đầu của trường đại học trong khu vực Europe (thường được phân loại thành các nhóm hàng đầu, ví dụ: Top 50, Top 70).

1. Dataset Asia

Tập dataset này là một bộ thông tin về các trường đại học hàng đầu trong khu vực châu Á. Tập dataset gồm 19 cột và 100 hàng dữ liệu. Các cột đó là:

* Name: Tên của trường đại học.
* World\_Ranking: Xếp hạng của trường đại học trên toàn cầu.
* Count\_World\_Ranking: Vị trí xếp hạng tương đối của trường đại học trên toàn cầu.
* Asian\_Ranking: Xếp hạng của trường đại học trong khu vực châu Á.
* Count\_Asian\_Ranking: Vị trí xếp hạng tương đối của trường đại học trong khu vực châu Á.
* Coutry\_Ranking: Xếp hạng của trường đại học trong quốc gia.
* Coutry: Quốc gia nơi trường đại học đặt.
* City: Thành phố nơi trường đại học đặt.
* Type: Loại hình của trường đại học (ví dụ: công lập, tư nhân).
* Count\_Type: Số lượng trường đại học thuộc cùng loại hình.
* Acceptance\_Rat: Tỷ lệ chấp nhận của trường đại học (tỷ lệ sinh viên được nhận vào so với số lượng sinh viên đăng ký).
* Publication: Số lượng công bố mà trường đại học đã có.
* High\_degree: Cấp độ học vị cao nhất mà trường đại học cung cấp.
* Count\_High\_degree: Số lượng chương trình học vị cao nhất mà trường đại học cung cấp.
* Web: Địa chỉ trang web của trường đại học.
* Language: Ngôn ngữ được sử dụng trên trang web của trường đại học.
* Phone: Số điện thoại liên hệ của trường đại học.
* Address: Địa chỉ vật lý của trường đại học.
* TopRank: Xếp hạng hàng đầu của trường đại học trong quốc gia.

Bạn có thể sử dụng dữ liệu này để nghiên cứu và so sánh các trường đại học, phân tích các yếu tố như xếp hạng, tỷ lệ chấp nhận, công bố, loại hình và địa chỉ trụ sở. Ngoài ra, bạn cũng có thể tạo bảng xếp hạng và biểu đồ để trực quan hóa dữ liệu và hiểu rõ hơn về sự phân bố và các mối quan hệ giữa các trường đại học trong khu vực châu Á.

1. Dataset Pakistan

Tập dataset Pakistan cung cấp thông tin chi tiết về các trường đại học ở Pakistan và xếp hạng của chúng trên cấp độ toàn cầu, khu vực châu Á và trong nước Pakistan. Tập dataset gồm 19 cột và 100 hàng dữ liệu. Các cột đó là:

* Name: Tên của trường đại học.
* World\_Ranking: Xếp hạng của trường đại học trên toàn cầu.
* Count\_World\_Ranking: Vị trí xếp hạng tương đối của trường đại học trên toàn cầu.
* Asian\_Ranking: Xếp hạng của trường đại học trong khu vực châu Á.
* Count\_Asian\_Ranking: Vị trí xếp hạng tương đối của trường đại học trong khu vực châu Á.
* Pakistani\_ranking: Xếp hạng của trường đại học trong Pakistan.
* Count\_Pakistani\_ranking: Vị trí xếp hạng tương đối của trường đại học trong Pakistan.
* City: Thành phố nơi trường đại học đặt.
* Type: Loại hình của trường đại học (ví dụ: công lập, tư nhân).
* Count\_Type: Số lượng trường đại học thuộc loại hình tương tự.
* Acceptance\_Rat: Tỷ lệ chấp nhận của trường đại học (tỷ lệ sinh viên được nhận vào so với số lượng sinh viên đăng ký).
* Publication: Số lượng công bố mà trường đại học đã có.
* Web: Địa chỉ trang web của trường đại học.
* Language: Ngôn ngữ được sử dụng trên trang web của trường đại học.
* Phone: Số điện thoại liên hệ của trường đại học.
* Address: Địa chỉ vật lý của trường đại học.
* Top Rank Pakistan: Xếp hạng hàng đầu của trường đại học trong Pakistan.
* Lon (Longitude): Kinh độ của vị trí địa lý của trường đại học
* Lat (Latitude): Vĩ độ của vị trí địa lý của trường đại học

Dữ liệu này có thể được sử dụng để thực hiện các tác vụ như:

* So sánh xếp hạng của các trường đại học ở Pakistan trên cấp độ toàn cầu, khu vực châu Á và trong nước.
* Phân tích các yếu tố như số lượng công bố, tỷ lệ chấp nhận, loại hình trường đại học để đánh giá và so sánh các trường.
* Tìm hiểu vị trí địa lý của các trường đại học và khám phá các mối quan hệ giữa vị trí địa lý và xếp hạng.
* Tìm kiếm thông tin cụ thể về trường đại học như địa chỉ, trang web, số điện thoại liên hệ và ngôn ngữ được sử dụng trên trang web của trường đại học.
* Xây dựng các bảng xếp hạng và biểu đồ để trực quan hóa và phân tích dữ liệu.

1. Dataset Lon & Lat – Top 300 universities of World

Tập dataset này chứa các thông tin về tọa độ địa lý của từng trường. Tập dataset gồm 4 cột và 300 hàng dữ liệu. Các cột đó là:

* Name: Tên của trường đại học.
* Lat: Giá trị vĩ độ (latitude) của vị trí địa lý của trường đại học.
* Lon: Giá trị kinh độ (longitude) của vị trí địa lý của trường đại học.
* region: Vùng lục địa hoặc khu vực địa lý mà trường đại học thuộc về.

Dữ liệu này có thể được sử dụng để thực hiện các tác vụ như hiển thị các trường đại học trên bản đồ, phân loại các trường theo khu vực địa lý, phân tích địa lý và so sánh các trường đại học trên cơ sở vị trí địa lý của chúng.

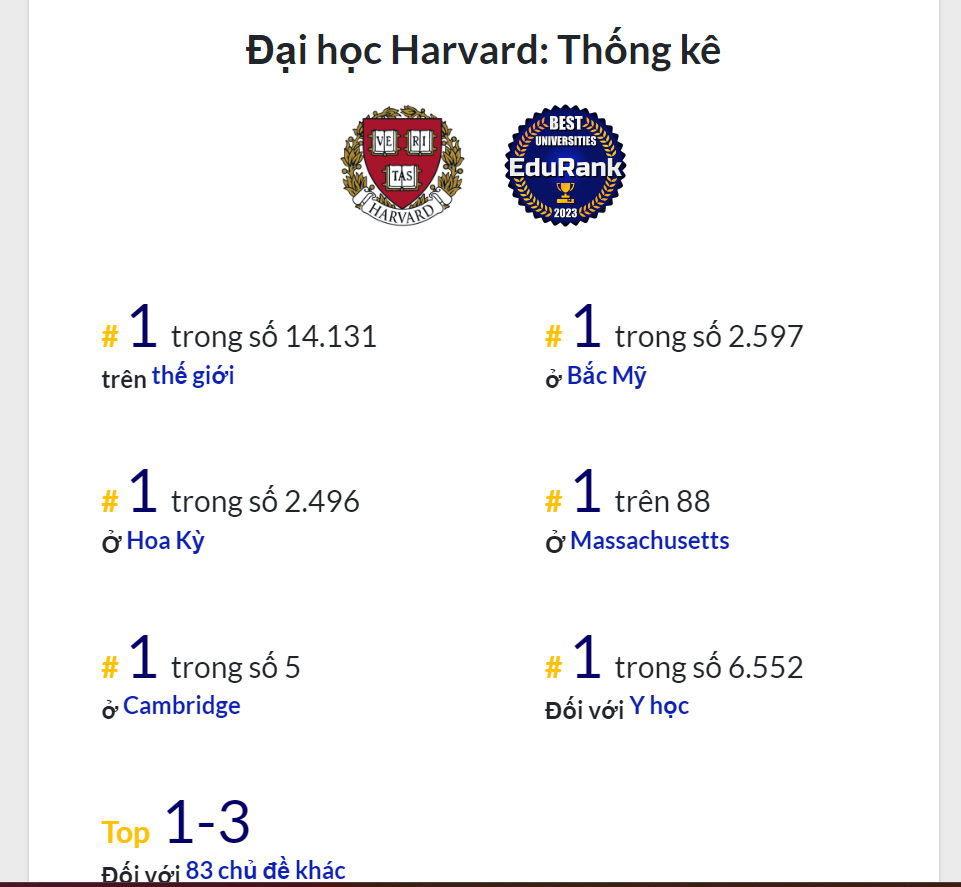
1. Dataset DSHocSinhTests

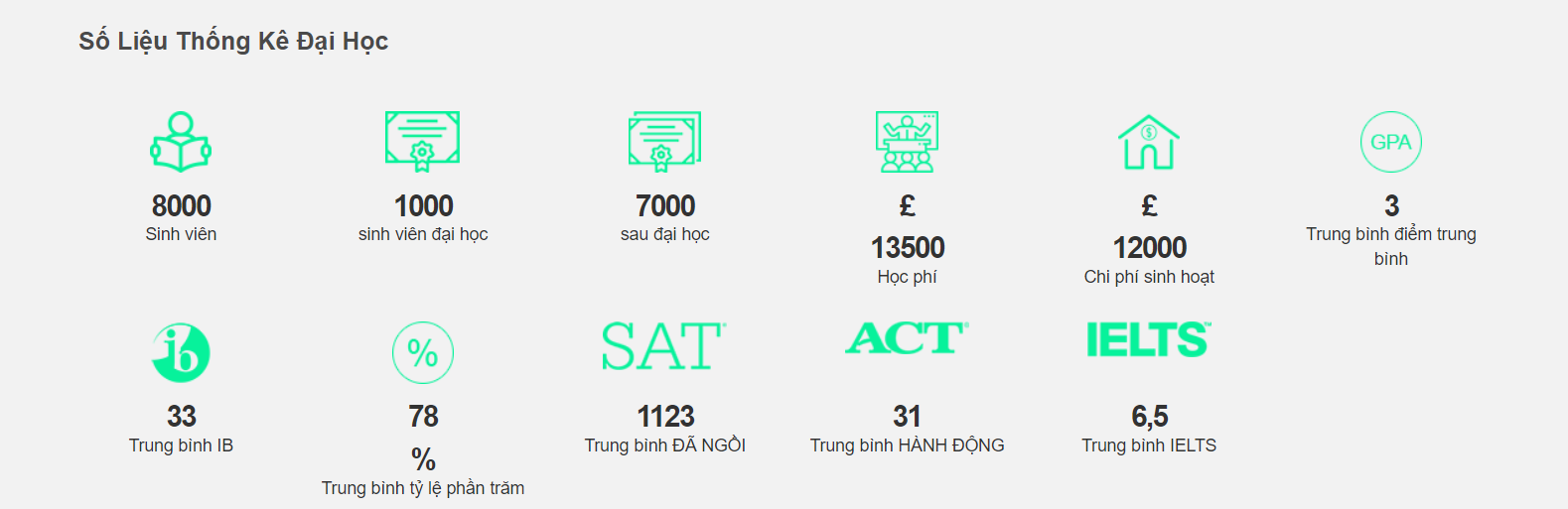
Tập dataset này được sử dụng để phân tích và xây dựng mô hình dự đoán xác suất đủ điều kiện để vào một trường đại học top đầu của một sinh viên dựa trên các thuộc tính đã cho. Dataset này gồm 7 cột và 21 dòng dữ liệu. Mỗi hàng dữ liệu là thuộc tính của một học sinh. Các thuộc tính đó là:

* Họ, Tên: Họ và tên của một sinh viên
* acceptance\_rate: Tỷ lệ chấp nhận vào trường đại học, được biểu diễn dưới dạng phần trăm.
* Average SAT: Điểm trung bình của bài kiểm tra SAT.
* Average ACT: Điểm trung bình của bài kiểm tra ACT.
* Average aid awarded: Số tiền hỗ trợ tài chính trung bình được trao cho sinh viên.
* Net average cost to attend: Chi phí trung bình để tham gia trường đại học sau khi được hỗ trợ tài chính.

1. **Tiền Xử Lý Dữ Liệu**
2. Thu Thập Dữ Liệu

- Thu thập dữ liệu trên trang <https://edurank.org/> và https://www.gotouniversity.com





1. Sửa chữa dữ liệu
2. Data set Asian có 14 cột trong đó có thiếu :

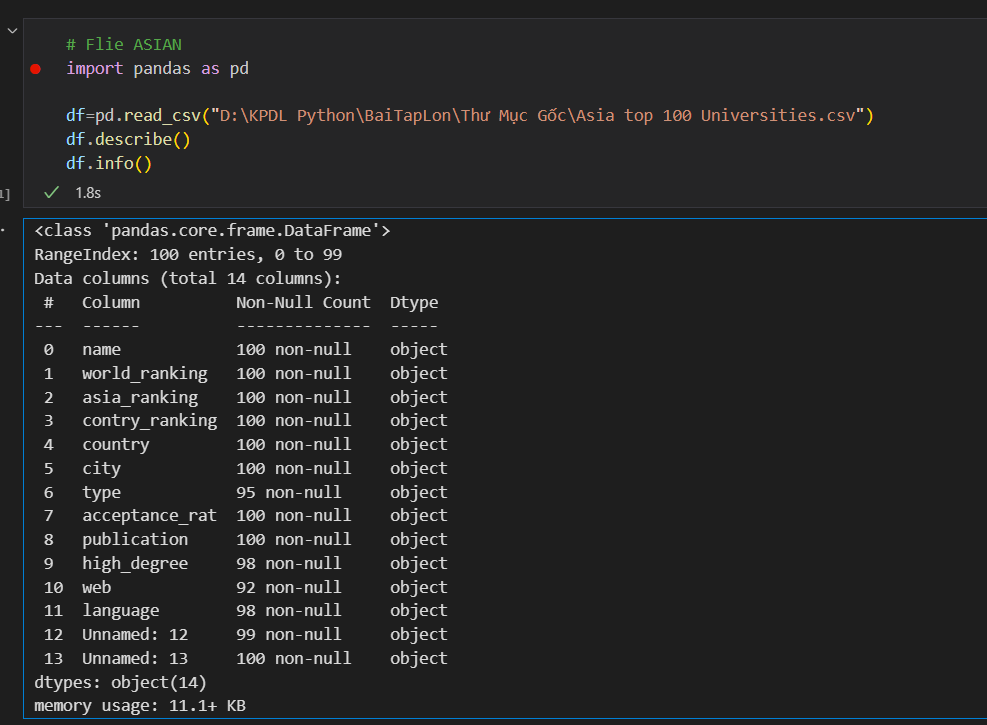
Cột type thiếu 5 dữ liệu

Cột high\_degree thiếu 2 dữ liệu

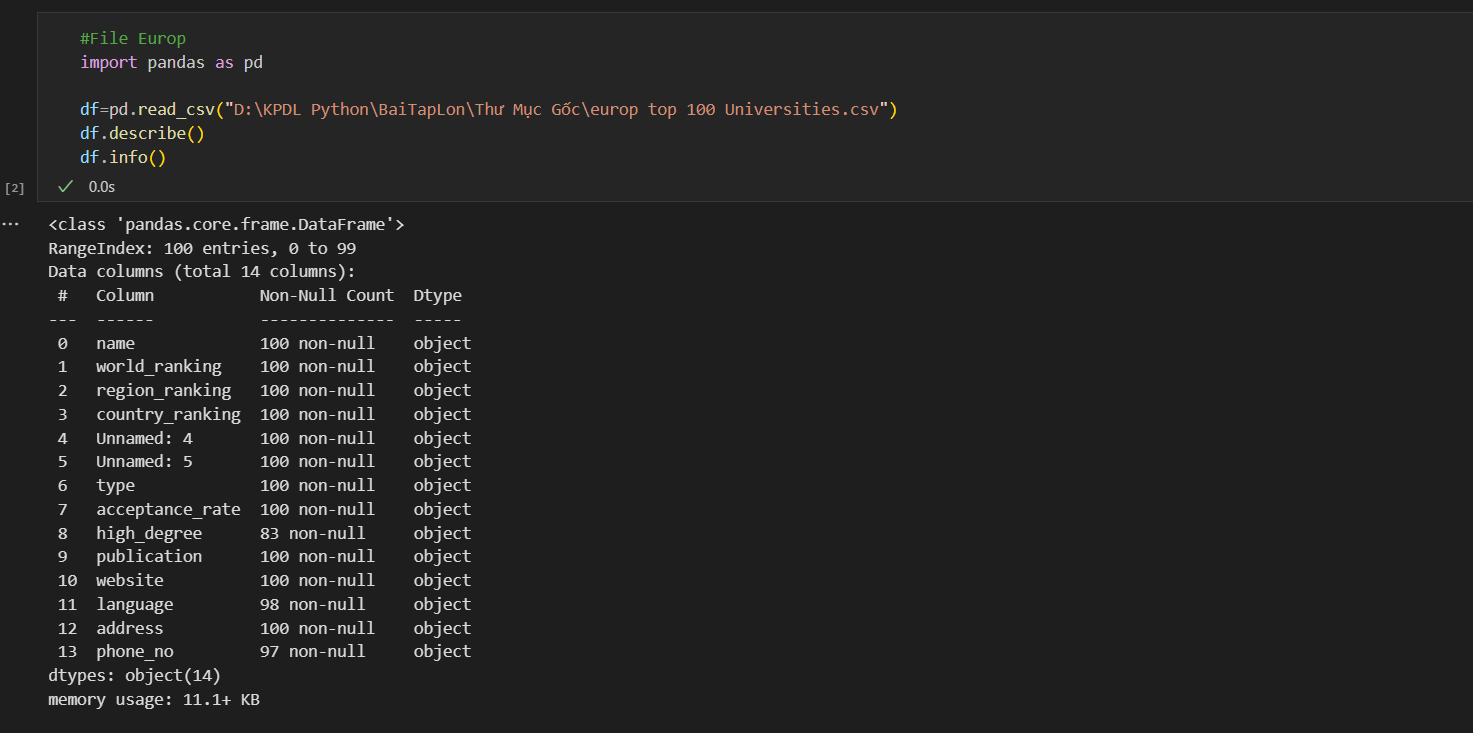
Cột web thiếu 2 dữ liệu

Cột language thiếu 2 dữ liệu

Thiếu tên 2 cột cuối



1. Data set Europe



Data set Asian có 14 cột trong đó có thiếu :

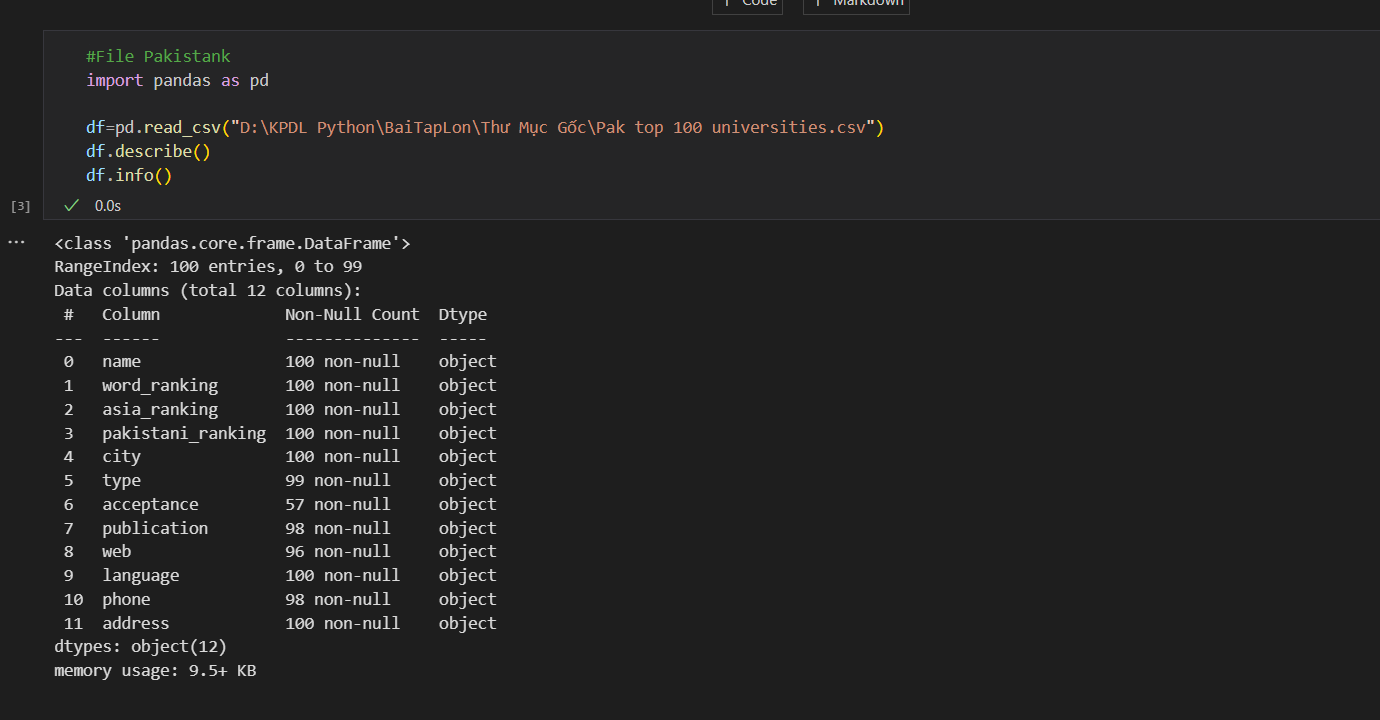
Cột high\_degree thiếu 17 dữ liệu

Cột phone thiếu 3 dữ liệu

Cột language thiếu 2 dữ liệu

Thiếu tên cột 5 và 6 cuối

1. Data set Pakistank



Data set Pakistank có 12 cột trong đó thiếu :

Cột type thiếu 1 dữ liệu

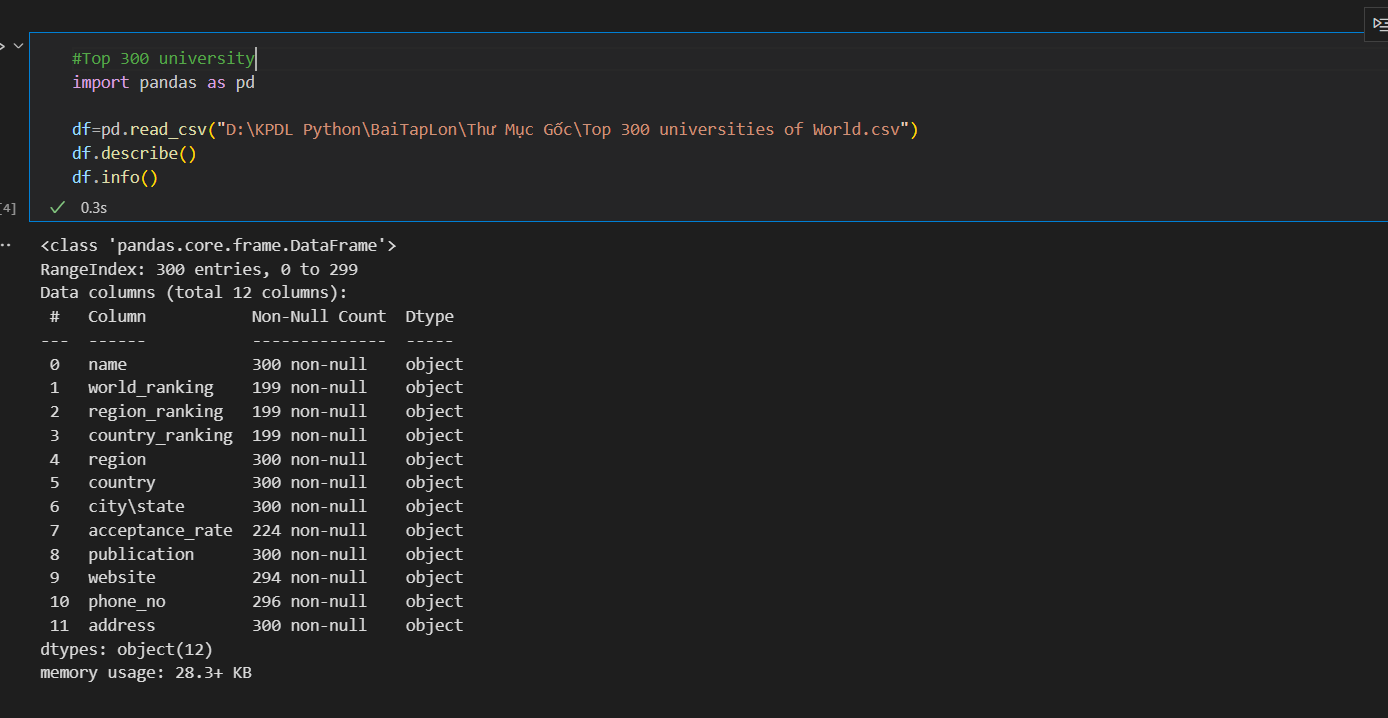
Cột acceptance thiếu 43 dữ liệu

Cột publication thiếu 2 dữ liệu

Cột web thiếu 4 dữ liệu

Cột phone thiếu 2 dữ liệu

1. Top 300 university



1. Data set Top 300 university có 12 cột trong đó thiếu :

Cột world\_raking thiếu 101 dữ liệu

Cột region\_ranking thiếu 101 dữ liệu

Cột country\_ranking thiếu 101 dữ liệu

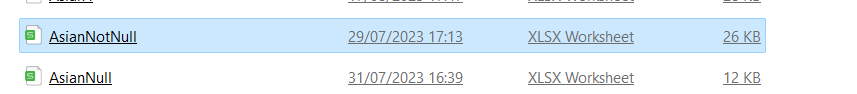
Cột web thiếu 6 dữ liệu

Cột phone thiếu 4 dữ liệu

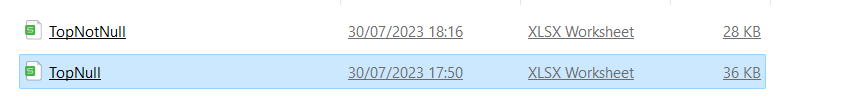
Cột acceptance\_rat thiếu 76 dữ liệu

1. Làm sạch dữ liệu

Đầu tiên chia 2 file của các data set trên thành 2 cái , 1 file không có dữ liệu null và not null của các hàng :





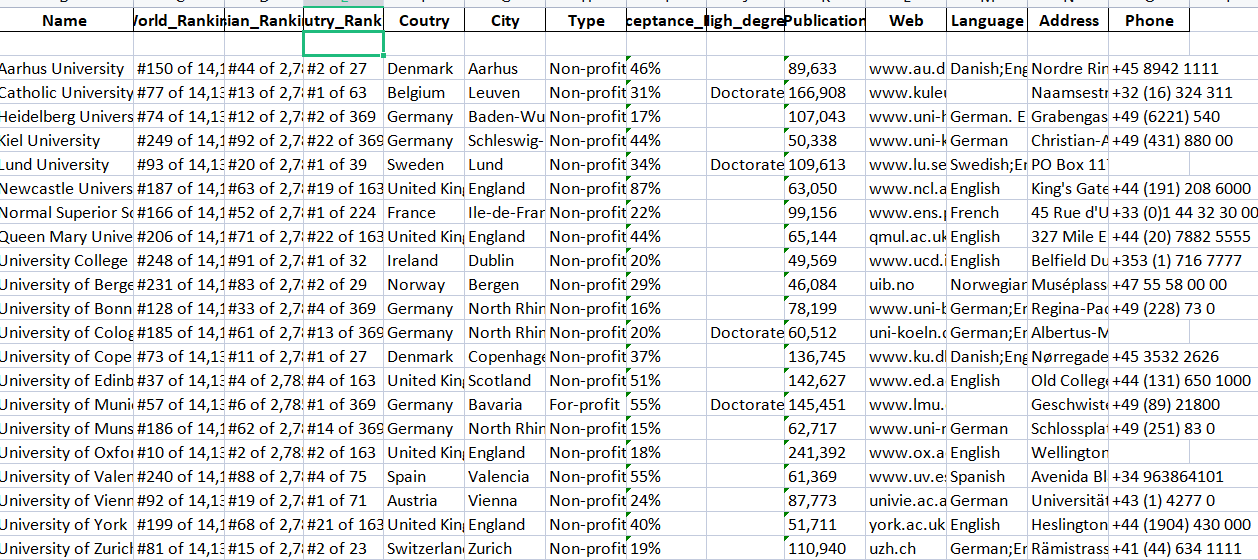


Điền vào những ô còn thiếu dữ liệu thông qua file có ô null:

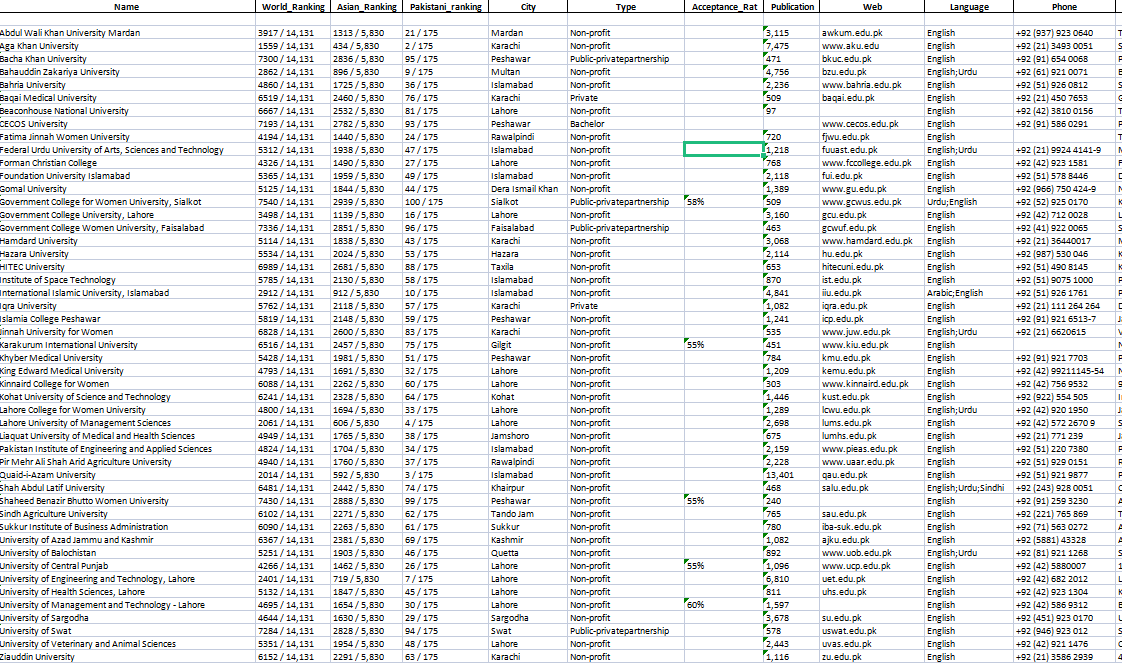
Asian



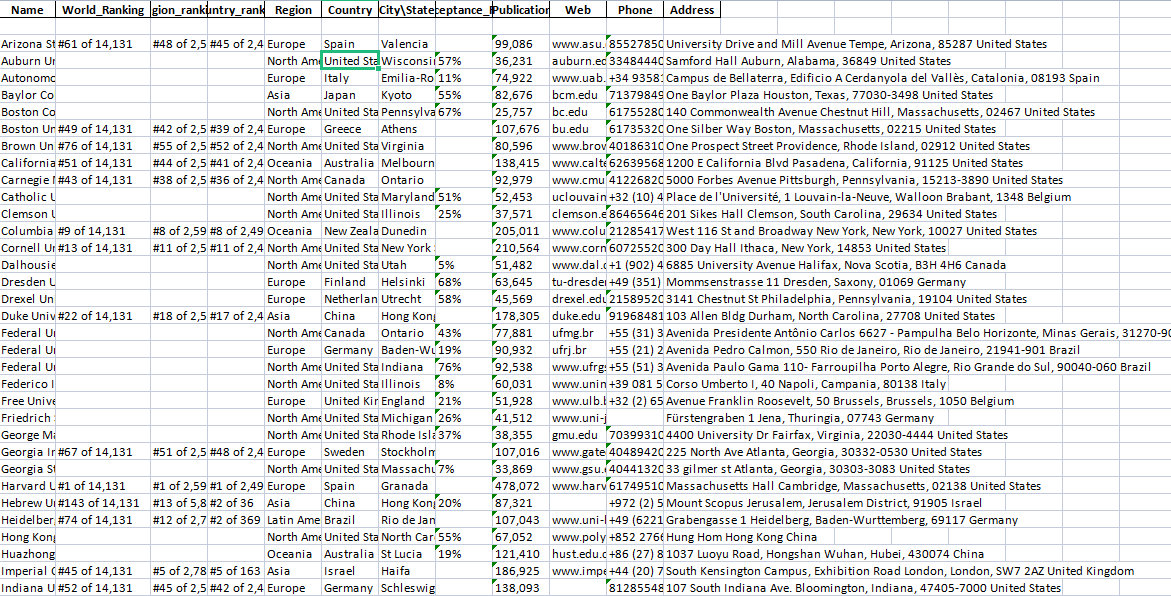
Europe



Pakistan



Top 300 university

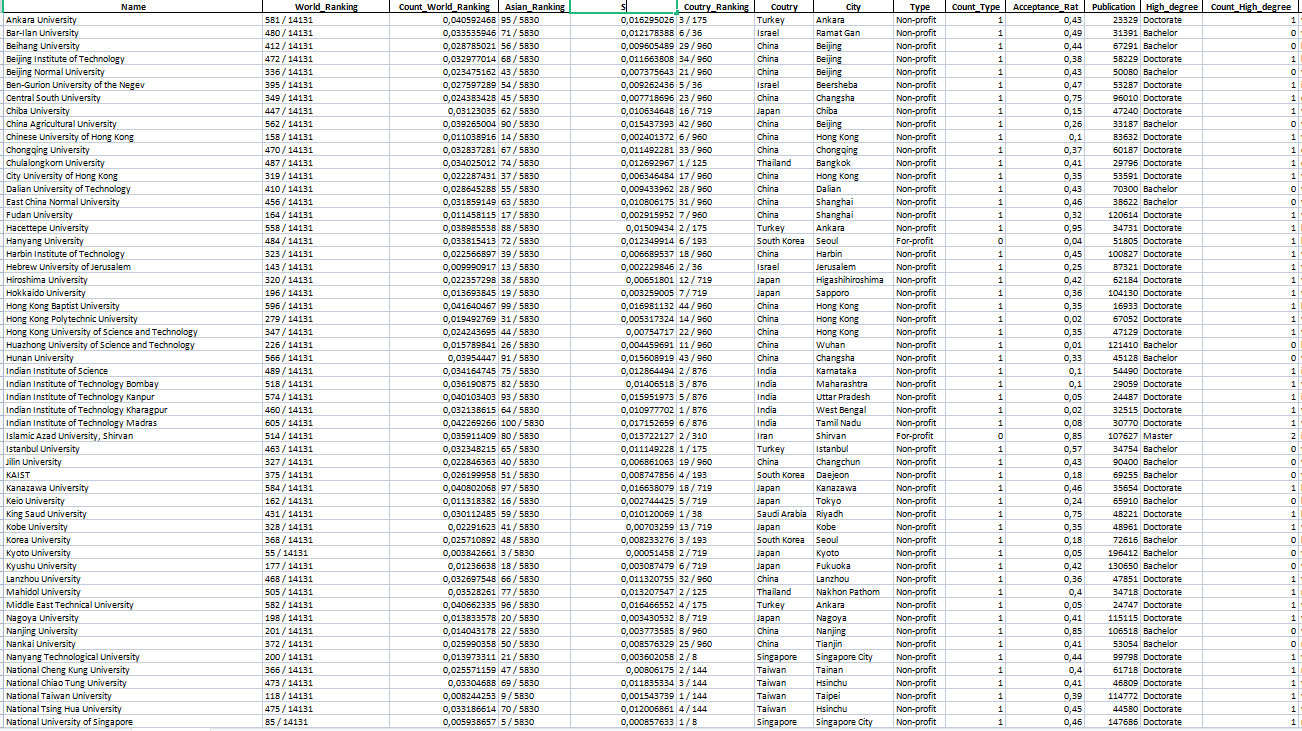


Xóa bỏ các dấu đưa về thành số nếu cột đó là cột số

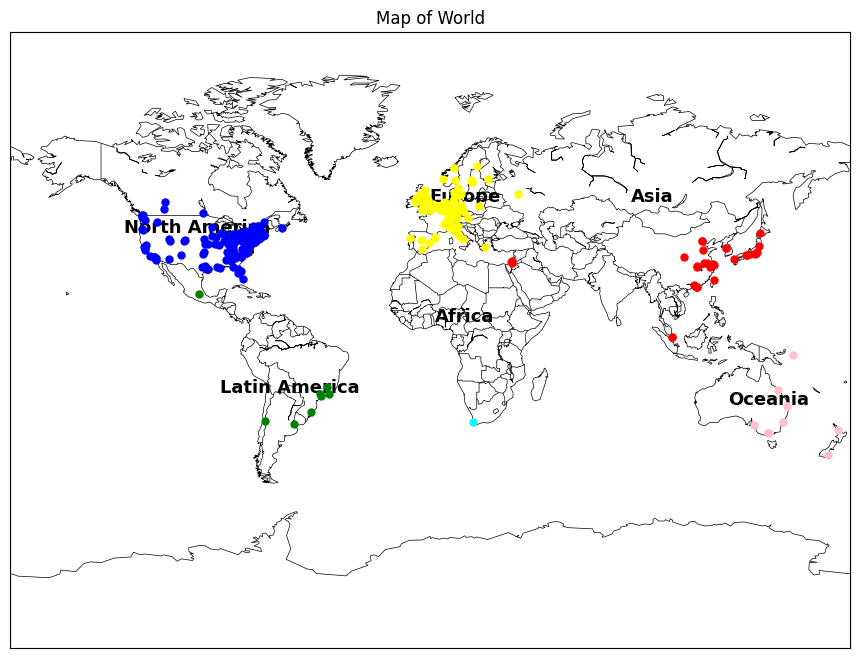
1. Chuẩn hóa dữ liệu

Chuẩn hóa dữ liệu các cột của từng file

Cột world\_ranking, Asian\_ranking, Acceptace\_rat , high\_degree,type



1. Phân cụm các trường đại học
2. Phân cụm theo châu lục
3. Phân cụm theo khu vực:



Trường đại học top đầu tập trung nhiều ở châu Âu (Europe) và Bắc Mỹ (North America) có một số lý do chính:

* Lịch sử và truyền thống: Châu Âu và Bắc Mỹ có lịch sử phát triển lâu đời trong lĩnh vực giáo dục đại học. Các trường đại học ở hai khu vực này đã tồn tại từ hàng thế kỷ và có truyền thống giáo dục chất lượng cao. Sự phát triển và danh tiếng của các trường đại học hàng đầu đã được xây dựng và phát triển qua nhiều thế hệ.
* Đầu tư và nguồn lực: Châu Âu và Bắc Mỹ đã đầu tư lớn vào giáo dục và nghiên cứu. Chính phủ và tổ chức tài trợ trong hai khu vực này cung cấp nguồn lực lớn để phát triển và duy trì các trường đại học hàng đầu. Điều này bao gồm cung cấp tài trợ cho nghiên cứu, cơ sở hạ tầng hiện đại và hỗ trợ tài chính cho sinh viên.
* Chất lượng giáo dục: Các trường đại học hàng đầu ở châu Âu và Bắc Mỹ thường đạt được tiêu chuẩn cao về chất lượng giảng dạy, nghiên cứu và cơ hội phát triển cá nhân cho sinh viên. Các trường này thường có các chương trình đào tạo đa dạng, đội ngũ giảng viên và nhà nghiên cứu hàng đầu, và liên kết mạng lưới với các tổ chức và công ty địa phương và quốc tế.
* Môi trường đa dạng và quốc tế: Châu Âu và Bắc Mỹ thu hút sinh viên từ khắp nơi trên thế giới, tạo ra một môi trường đa dạng với sự giao lưu văn hóa và trao đổi kiến thức. Sinh viên quốc tế thường muốn tới các trường đại học ở hai khu vực này để trải nghiệm cuộc sống và học tập trong một môi trường quốc tế.
* Nền kinh tế phát triển: Châu Âu và Bắc Mỹ có nền kinh tế phát triển mạnh mẽ, tạo điều kiện thuận lợi cho đầu tư vào giáo dục và nghiên cứu. Sự phát triển kinh tế này cũng mang lại nhiều cơ hội việc làm và sự tương tác với các công ty và tổ chức nghiên cứu hàng đầu, tạo điều kiện cho sinh viên tiếp cận các cơ hội nghề nghiệp và hợp tác nghiên cứu.

1. Phân cụm theo tỷ lệ chấp nhận:

* Đối với thế giới:
  + Phân thành 2 cụm

A graph of different points

Description automatically generated with medium confidence

* + Phân thành 3 cụm

A graph of different colored dots

Description automatically generated

* Đối với khu vực:
  + Phân thành 2 cụm

A chart with yellow and purple dots

Description automatically generated

* + Phân thành 3 cụm

A chart of dots and numbers

Description automatically generated with medium confidence

* Với k = 9 thì  khi train có error cao nhưng khi áp dụng vào test lại có error thấp => Chọn k = 9 bắt đầu vì khi k = 9 error bão hòa.

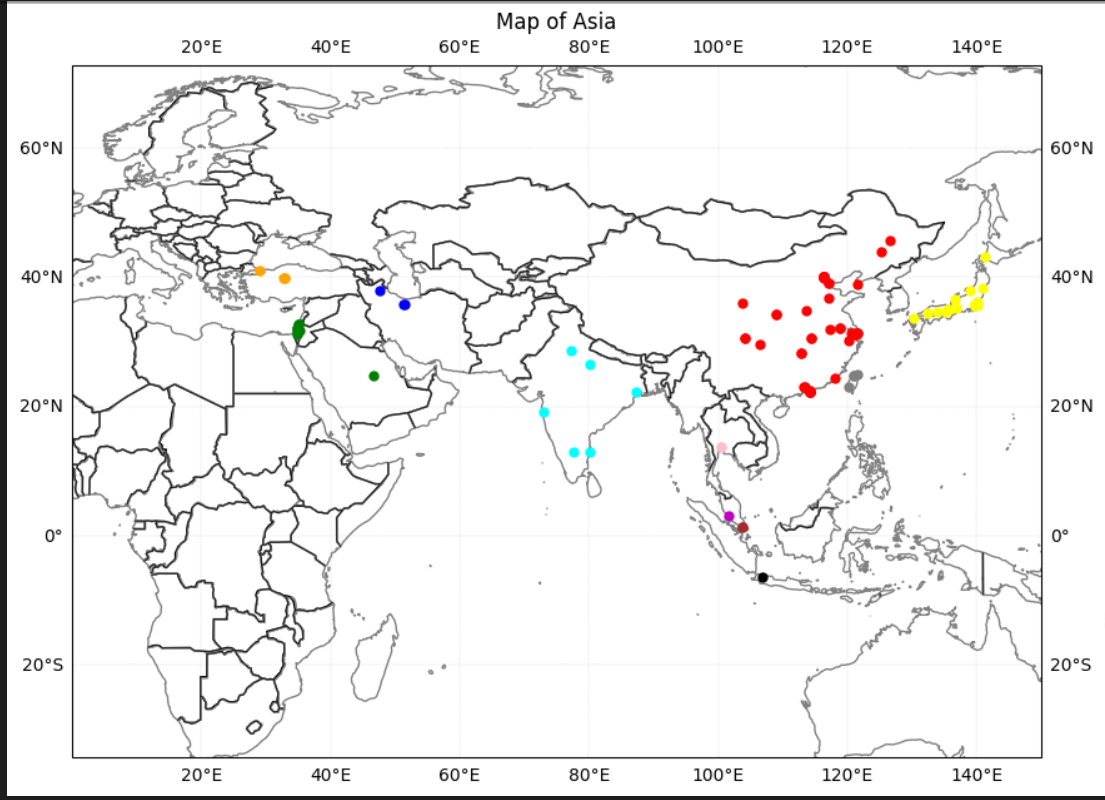
A graph with numbers and lines

Description automatically generated

1. Phân cụm theo khu vực

2.1 Asian

1. Phân cụm theo từng quốc gia

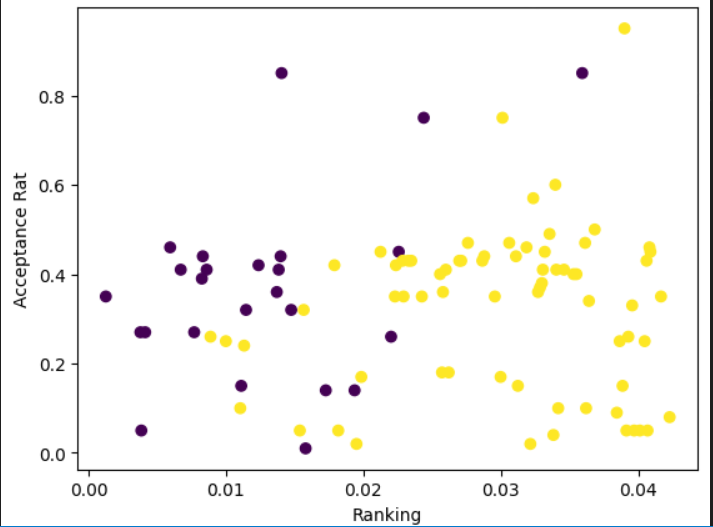


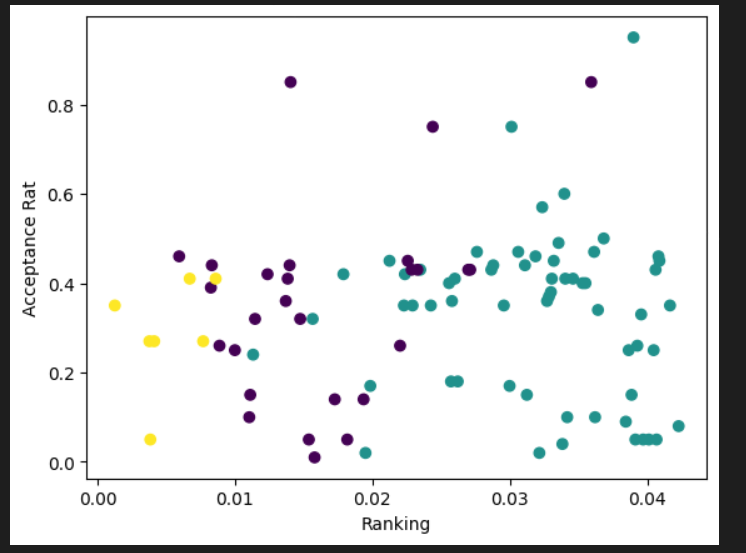
Ta thấy các trường top đầu thường tập trung nhiều ở 2 nước là Trung Quốc và Nhật Bản . Sự tập trung của các trường đầu hàng đầu ở Trung Quốc và Nhật Bản có thể được giải thích bởi một số yếu tố sau đây:

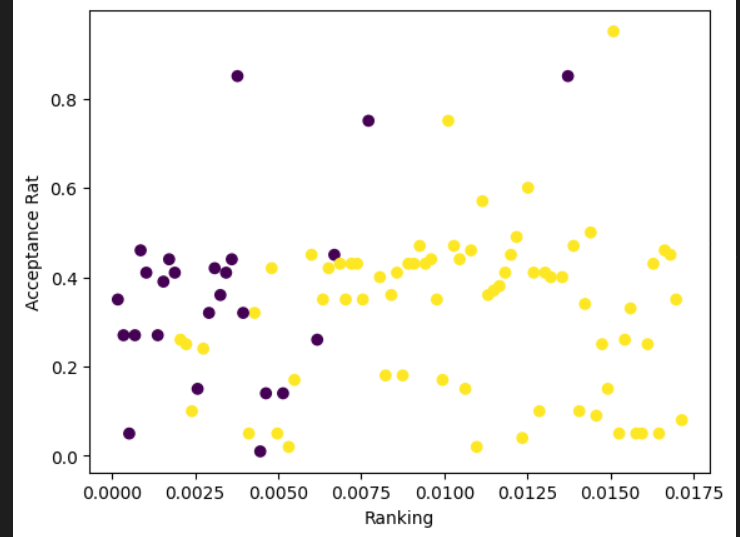
* Kinh tế và phát triển công nghệ: Cả Trung Quốc và Nhật Bản đều là hai quốc gia có kinh tế phát triển và sự đầu tư mạnh mẽ vào lĩnh vực công nghệ và giáo dục. Trung Quốc là một trong những quốc gia có tốc độ tăng trưởng kinh tế nhanh nhất thế giới và đã đầu tư mạnh vào nghiên cứu và phát triển công nghệ. Nhật Bản cũng nổi tiếng với sự tiên phong trong lĩnh vực công nghệ và có nền kinh tế mạnh mẽ.
* Đầu tư vào giáo dục: Cả Trung Quốc và Nhật Bản gắn sự quan trọng đặc biệt vào giáo dục và đã đầu tư lớn vào hệ thống giáo dục của mình. Họ có các trường đại học và viện nghiên cứu hàng đầu, cung cấp môi trường học tập và nghiên cứu tiên tiến cho sinh viên và nhà nghiên cứu.
* Nghiên cứu và sáng tạo: Cả Trung Quốc và Nhật Bản đều có một lịch sử lâu đời và giàu truyền thống trong lĩnh vực nghiên cứu và sáng tạo. Các trường đại học và viện nghiên cứu của hai quốc gia này thường có sự tập trung cao các nhà khoa học, nhà nghiên cứu và những người có đóng góp quan trọng trong các lĩnh vực khác nhau.
* Hợp tác quốc tế: Cả Trung Quốc và Nhật Bản đều đã thiết lập mối quan hệ hợp tác quốc tế mạnh mẽ trong lĩnh vực giáo dục và nghiên cứu. Điều này thu hút nhiều sinh viên và nhà nghiên cứu từ các quốc gia khác đến Trung Quốc và Nhật Bản để học tập và tham gia vào hoạt động nghiên cứu.

1. **Phân cụm theo tỷ lệ chấp nhận**

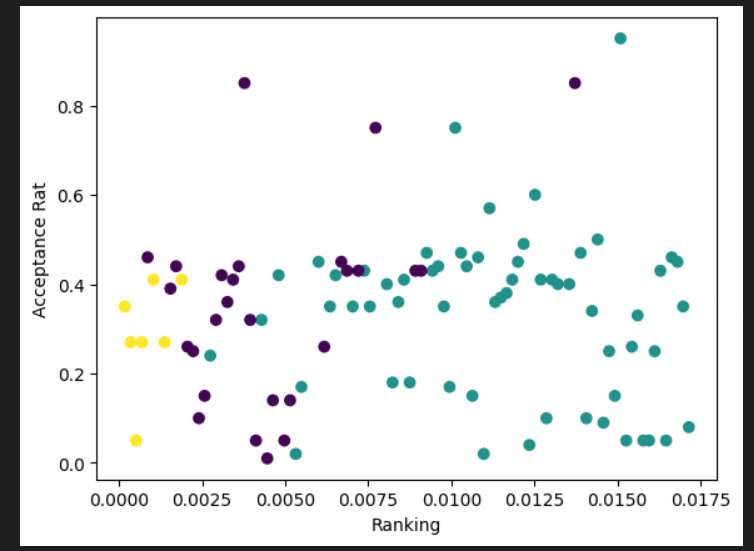
* Đối với thế giới
  + Phân thành 2 cụm



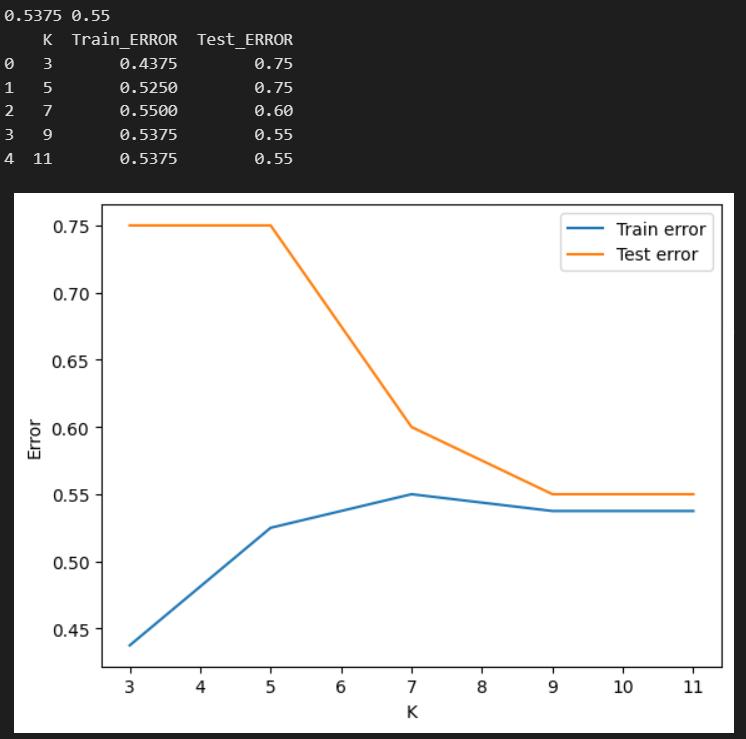
* + Phân thành 3 cụm:
* 
* Đối với Asian
  + Phân thành 2 cụm



* + Phân thành 3 cụm

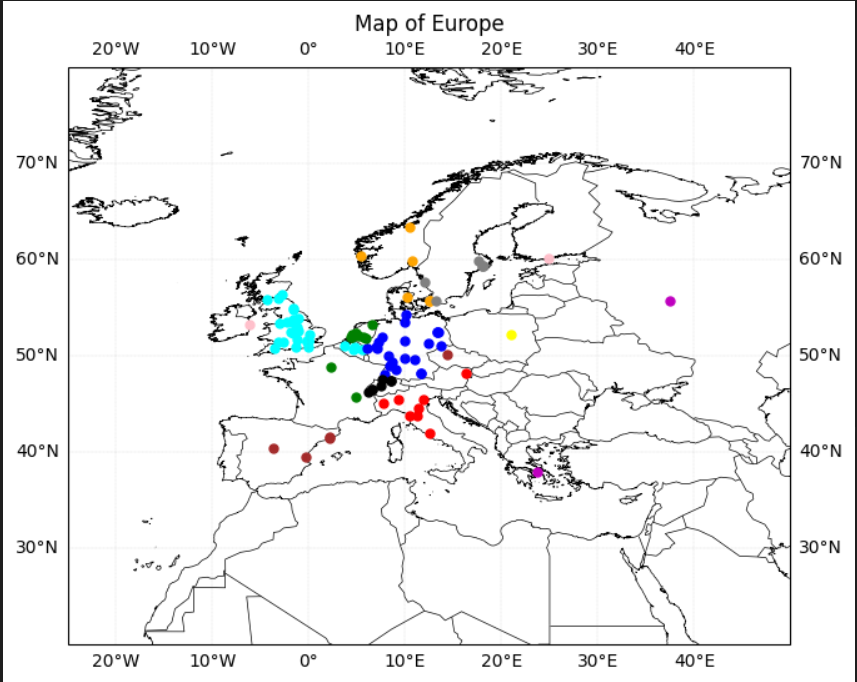


* Với k = 9 thì  khi train có error cao nhưng khi áp dụng vào test lại có error thấp => Chọn k = 9 bắt đầu vì khi k = 9 error bão hòa.



2.2 Europe

a) Phân cụm theo quốc gia



Các trường đại học ở châu Âu tập trung vào các nước phát triển mạnh về nền giáo dục .

Các trường đại học hàng đầu ở Châu Âu thường nằm tại các quốc gia có lịch sử và truyền thống giáo dục mạnh mẽ, đầu tư lớn vào nghiên cứu và giáo dục, và có sự phát triển kinh tế ổn định. Dưới đây là một số quốc gia và thành phố nổi tiếng với các trường đại học hàng đầu của mình:

+Anh: Anh có một số trường đại học hàng đầu thế giới như Đại học Oxford, Đại học Cambridge, Imperial College London và University College London. Các trường này có lịch sử lâu đời, nguồn lực mạnh mẽ và đội ngũ giảng viên và sinh viên xuất sắc.

+Đức: Đức có hệ thống giáo dục cao cấp rất đa dạng và được công nhận trên toàn cầu. Các trường đại học hàng đầu của Đức bao gồm Đại học Heidelberg, Đại học Munich, Đại học Technical Munich và Đại học Freiburg. Thành phố Berlin cũng thu hút sự chú ý với nhiều trường đại học danh tiếng như Đại học Humboldt và Đại học Teknik Berlin.

+Pháp: Pháp có một số trường đại học hàng đầu như Đại học Paris-Saclay, Đại học Sorbonne, Đại học Pierre và Marie Curie (Paris VI) và Đại học Paris-Dauphine. Thành phố Paris, với lịch sử văn hóa và giáo dục lâu đời, là trung tâm của nhiều trường đại học hàng đầu.

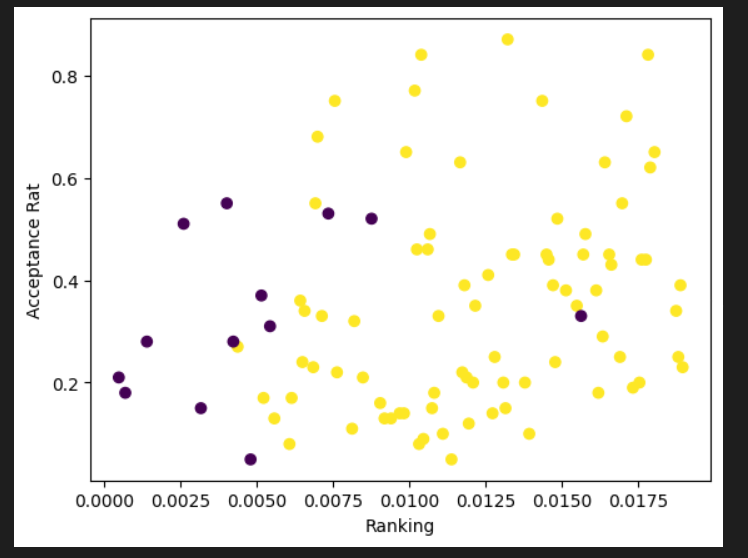
+Hà Lan: Hà Lan có một số trường đại học tốt được xếp hạng cao như Đại học Amsterdam, Đại học Leiden, Đại học Utrecht và Đại học Wageningen. Các trường đại học này thường nằm ở các thành phố như Amsterdam, Leiden và Utrecht.

+Thụy Điển: Thụy Điển có một hệ thống giáo dục cao cấp với chất lượng cao và hướng tới nghiên cứu. Các trường đại học hàng đầu của Thụy Điển bao gồm Đại học Stockholm, Đại học Uppsala và Đại học Lund.

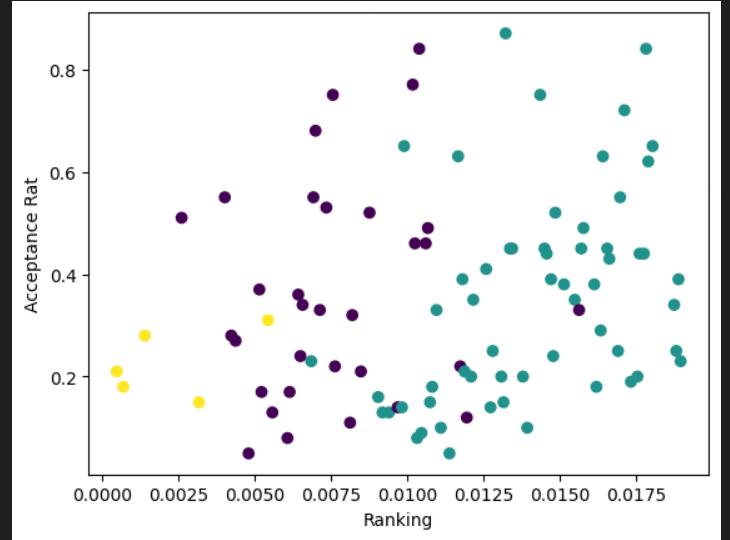
Lý do các trường đại học hàng đầu thường nằm tại các quốc gia và thành phố này có thể là do lịch sử và truyền thống của nền giáo dục, đầu tư mạnh mẽ vào nghiên cứu và học tập, sự tập trung tài chính và nguồn lực, cùng với khả năng thu hút và giữ chân các tài năng xuất sắc từ khắp nơi trên thế giới.

**b)Phân cụm theo tỷ lệ chấp nhận**

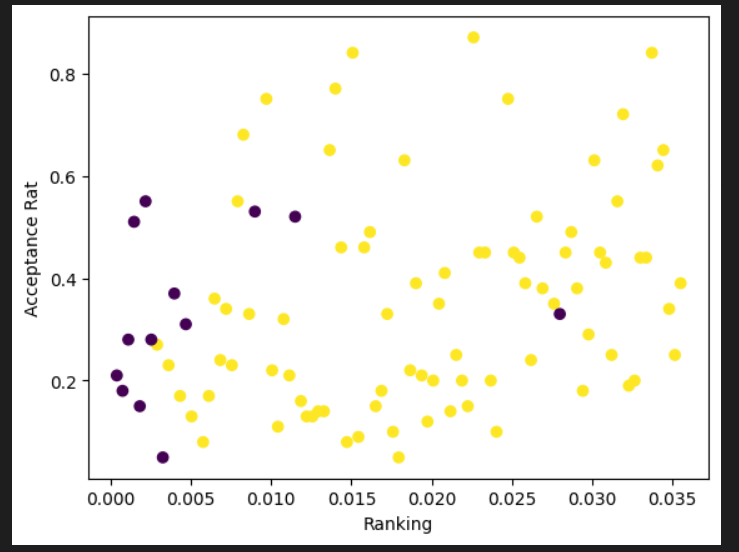
* Đối với thế giới
  + Phân thành 2 cụm :



* + Phân thành 3 cụm:



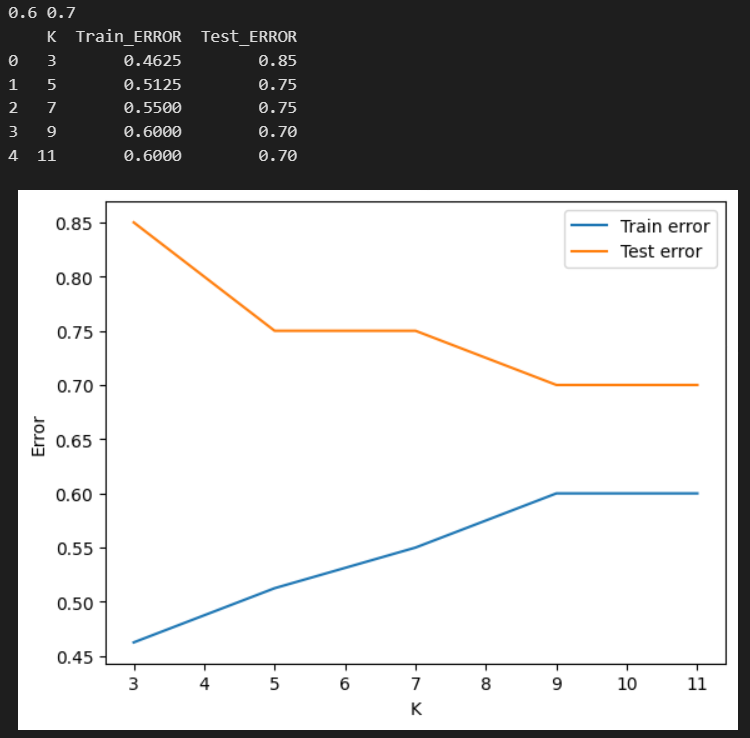
* Đối với Europe
  + Phân thành 2 cụm



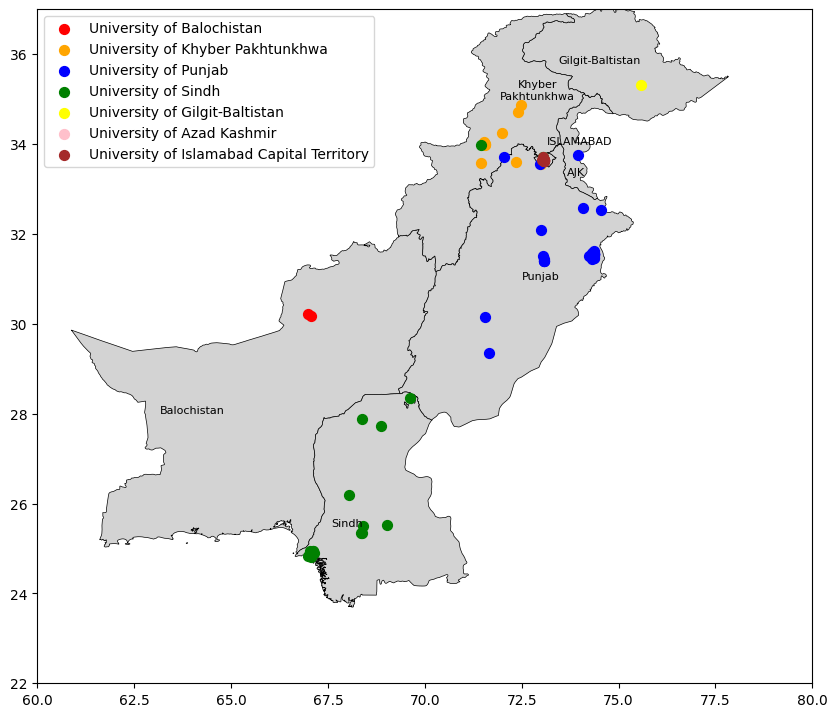
* + Phân thành 3 cụm



* Với k = 9 thì  khi train có error cao nhưng khi áp dụng vào test lại có error thấp => Chọn k = 9 bắt đầu vì khi k = 9 error bão hòa.



1. Phân Theo Quốc Gia (Pakistan)
2. Phân theo từng vùng của quốc gia

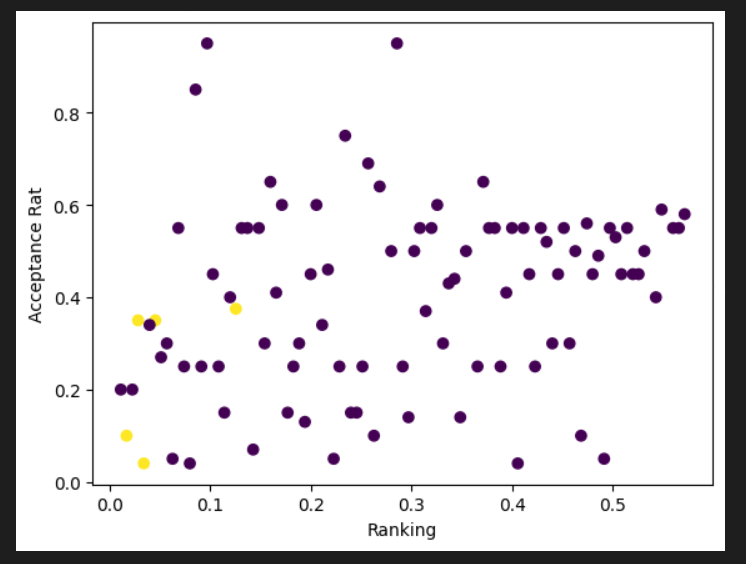


Trường đại học top đầu tập trung đông tại tỉnh Punjab và Sindh của Pakistan vì:

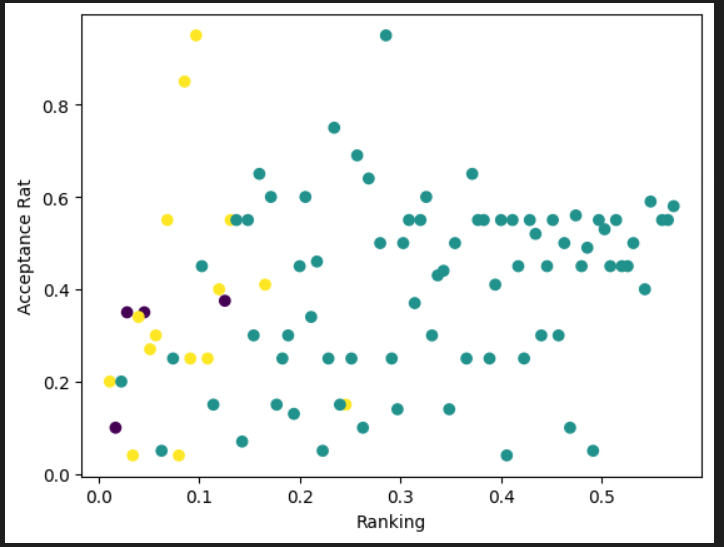
* Truyền thống giáo dục và lịch sử: Punjab và Sindh có lịch sử lâu đời và truyền thống phát triển giáo dục. Các trường đại học hàng đầu như Punjab University, Lahore University of Management Sciences (LUMS) ở Punjab và University of Karachi, Aga Khan University ở Sindh đã có một lịch sử phát triển dài và được công nhận trong giới học thuật. Sự uy tín này thu hút sinh viên và giảng viên tài năng từ khắp nơi tới học và giảng dạy.
* Dân số đông: Punjab và Sindh là hai tỉnh có dân số đông nhất tại Pakistan. Dân số đông tạo ra một nguồn cung cấp lớn của sinh viên tiềm năng, và các trường đại học hàng đầu tận dụng lợi thế này để thu hút và tuyển chọn những ứng viên xuất sắc. Đồng thời, sự cạnh tranh giữa các trường đại học trong khu vực này cũng thúc đẩy sự phát triển và nâng cao chất lượng giáo dục.
* Tài nguyên và cơ sở hạ tầng: Punjab và Sindh đã đầu tư đáng kể vào tài nguyên và cơ sở hạ tầng giáo dục. Các trường đại học hàng đầu được trang bị cơ sở vật chất hiện đại, thư viện, phòng thí nghiệm và các tiện ích học thuật khác để hỗ trợ quá trình giảng dạy, nghiên cứu và phát triển học thuật. Sự đầu tư này thu hút giảng viên và nghiên cứu viên tài năng và cung cấp môi trường học tập tốt cho sinh viên.
* Kinh tế và doanh nghiệp: Punjab và Sindh là hai tỉnh phát triển kinh tế của Pakistan, với nhiều cơ hội việc làm và mối quan hệ với các doanh nghiệp và công ty. Các trường đại học hàng đầu tại Punjab và Sindh tận dụng mối quan hệ này bằng cách thiết lập liên kết với các doanh nghiệp và tổ chức để cung cấp cơ hội thực tập, đào tạo và tuyển dụng cho sinh viên. Điều này giúp sinh viên có môi trường học tập thực tế và cơ hội tiếp cận với thị trường lao động sau khi tốt nghiệp.
* Mối quan hệ quốc tế: Các trường đại học hàng đầu tại Punjab và Sindh đã xây dựng mối quan hệ quốc tế mạnh mẽ với các trường đại học và tổ chức giáo dục hàng đầu trên toàn thế giới. Điều này mang lại lợi thế hợp tác nghiên cứu, trao đổi sinh viên và giảng viên, và thu hút sự quan tâm của sinh viên quốc tế đến các trường đại học này.

1. Phân theo tỷ lệ chấp nhận đối với quốc gia

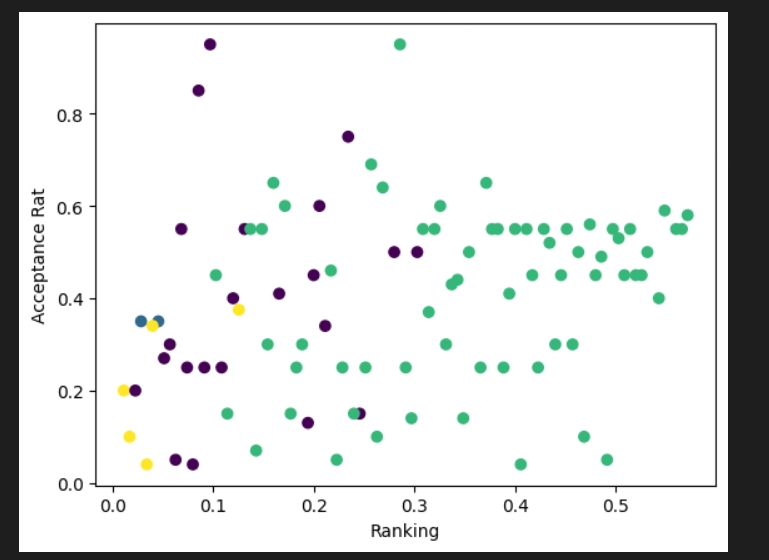
* Phân 2 cụm



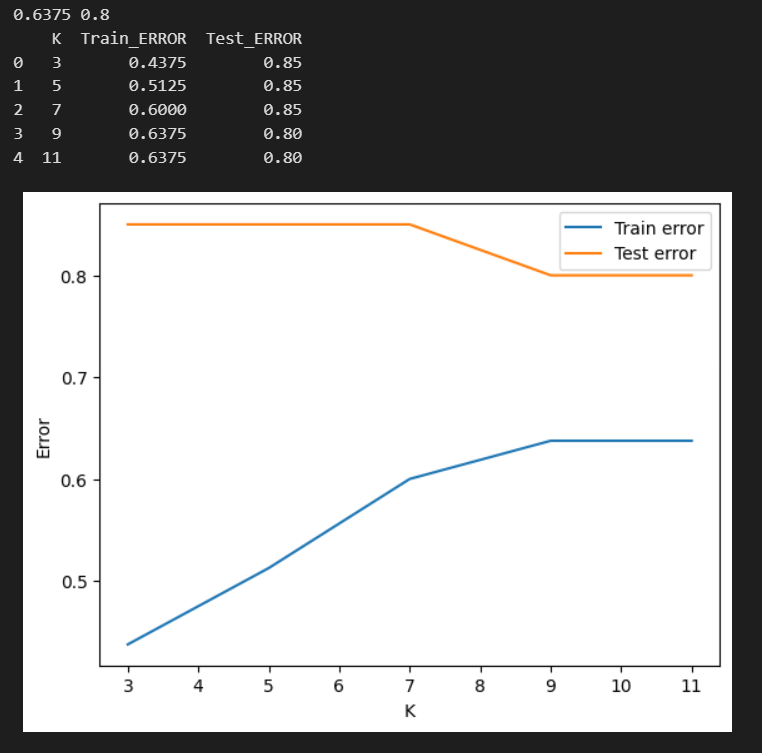
* Phân 3 cụm



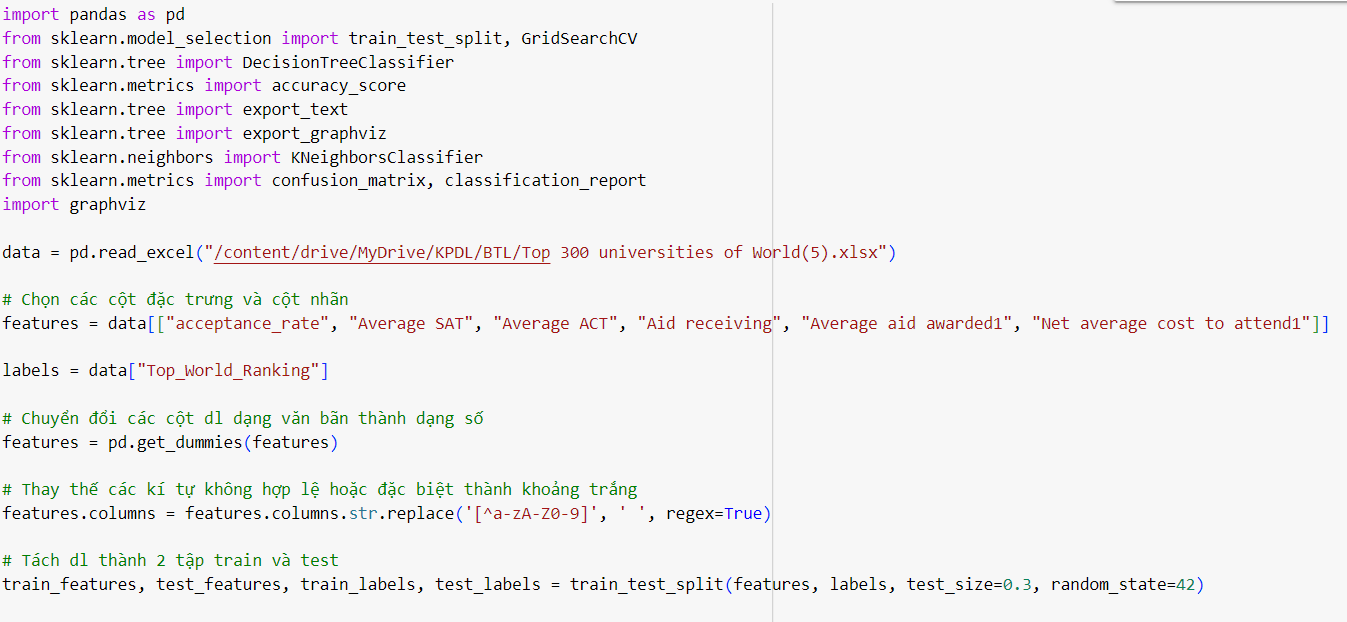
* Phân 4 cụm



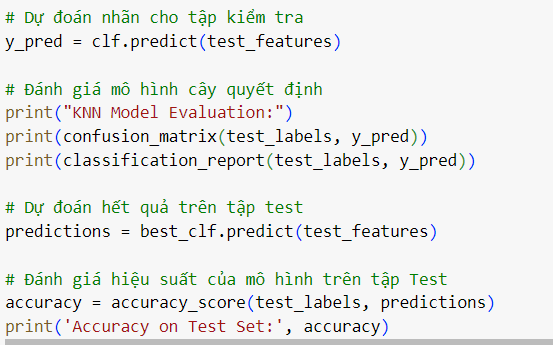
* Với k = 9 thì  khi train có error cao nhưng khi áp dụng vào test lại có error thấp => Chọn k = 9 bắt đầu vì khi k = 9 error bão hòa.



1. So Sánh Mật Độ Phân Bố Và Nhận Xét
2. Cây Quyết Định Trường Đại Học





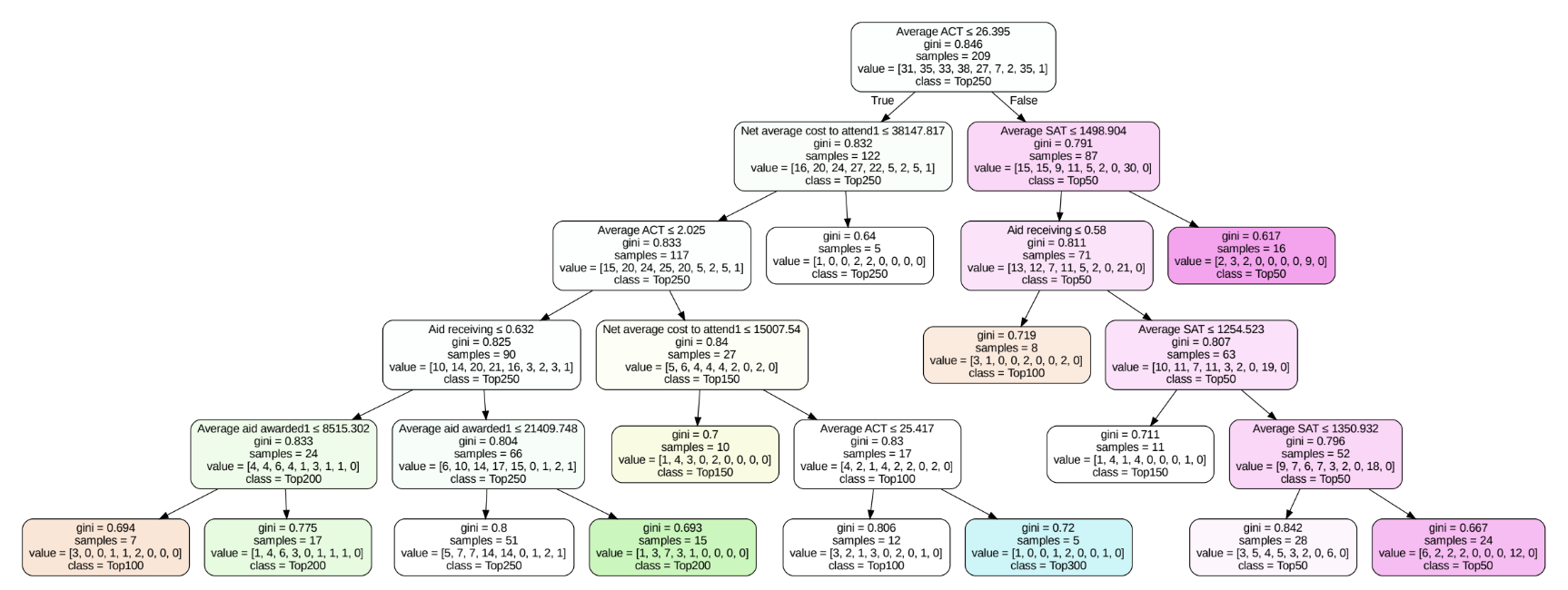


Ở đây ta sử dụng thư viện graphviz để in ra đồ thị cây quyết định với độ sâu là 5 cho dễ nhìn và tránh những trường hợp cây có quá nhiều nhánh con gây khó hiểu cho mô hình cây quyết định.

Kết quả độ chính xác sau khi chạy Python:



Model cây quyết định:



*Model cây quyết định*

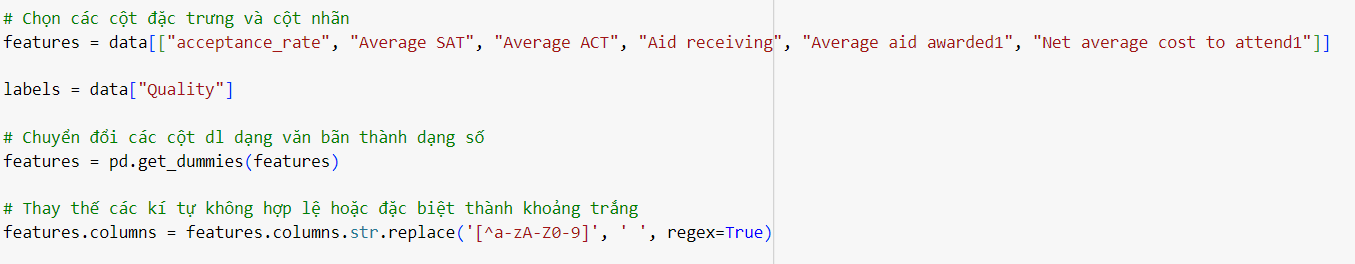
Ở hình trên, dòng đầu tiên là điều kiện của Node, nếu không có điều kiện thì là Node kết quả cho thấy. Kết quả ở đây là dòng cuối cùng

Ghi class, ví dụ như: class = Top100.

Dù đã phân tích và giới hạn lại độ phức tạp của cây trên nhưng mà độ chính xác khi mà áp dụng vào tập kiểm thử thì khá là thất, chỉ cỡ 32.54%. Một chỉ số khá thấp nên ta có thể nói rằng mô hình cây quyết định này đang bị Overfitting. Overfitting sẽ xảy ra khi mô hình cây từ tập huấn luyện có độ chính xác thấp khi áp dụng lên tập kiểm thử.

Do đó, để mà tăng độ chính xác tránh trường hợp bị Overfitting thì ta có thể biến đổi dữ liệu Top\_World\_Ranking thành 2 giá trị đơn giản là Goog và Bag được thêm vào cột mới với tên Quality. Với Good bao gồm Top50, Top100, Top150 và Bag thì gồm Top200, Top250, Top300.

Bây giờ ta sẽ kiểm chứng lại trên python bằng cách thay đổi giá trị class bằng cột Quality.

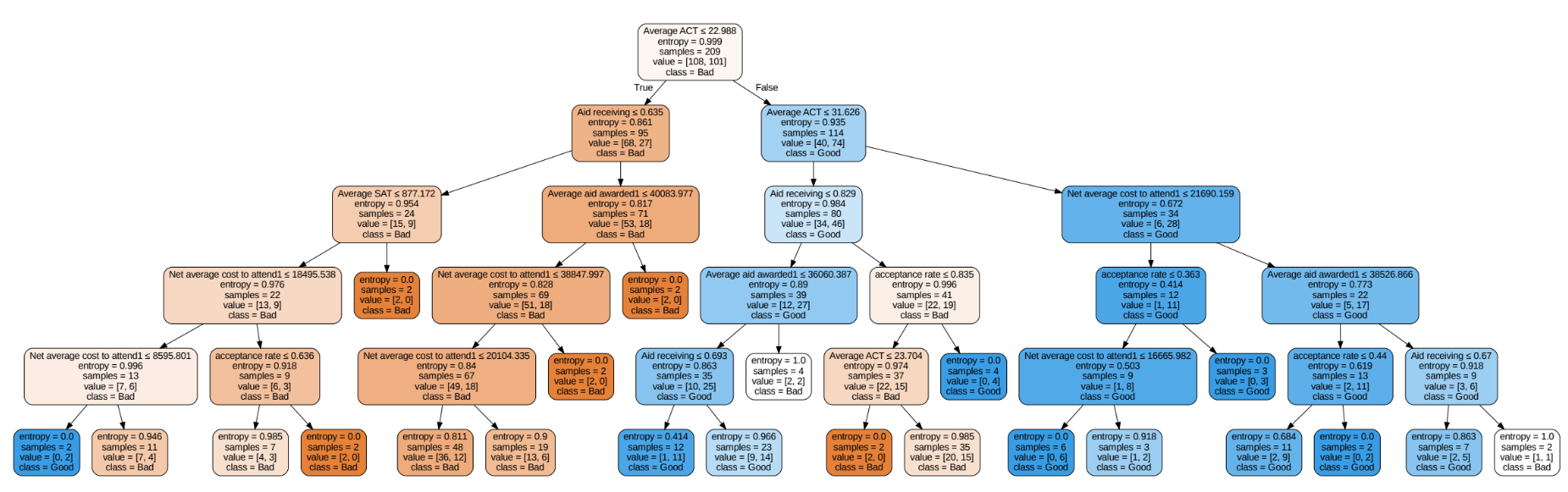


Chạy chương trình lần nữa kết quả là:



Độ chính xác vào bây giờ là 73.17%, cao hơn 40% nhiều so với cây quyết định ban đầu.

Sau đây là cây quyết định mới tạo thành:



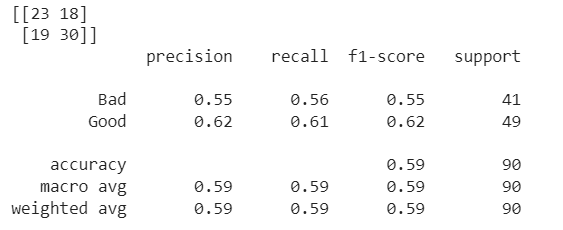
*Model cây quyết định mới*

Dựa theo cây quyết định mới này, chúng ta có thể khẳng định Average ACT là nhân tố quyết định đánh giá chất lượng của trường học. Khi mà Average ACT <= 22.988 thì chất lượng Bag với value = [108,101] ứng với 108 trường học Bag và 101 trường học Good.

Kiểm tra kiểm thử để nhìn rõ bản chất đúng sai của mô hình khi cho vào tập dữ liệu test:





Kết quả phân tích độ chính xác của cây:

Nhìn thấy, độ chính xác khi quyết định chất lượng trường học là Good là 62%, còn với Bag thì là 55%. Ma trận 2x2 cho thấy rõ điều này hơn: mô hình cây đánh giá chình xác được 23 Bag và 30 Good đúng như giá trị quality của tập test. Ngược lại, nó cũng đánh giá sai về 18 Bag thành Good và 19 Good thành Bag.

Ta có thể kết luận rằng mô hình cây quyết định này có độ chính xác khá tốt khi ta huấn luyện 70% tập dữ liệu và dùng 30% tập dữ liệu ban đầu để kiểm thử chất lượng của trường học ra 2 loại Good và Bag. Khi ta biến đổi giá trị cột chất lượng Top\_World\_Ranking đơn giản lại hơn so với ban đầu thì cây quyết định có được dễ nhìn và có một độ chính xác cao hơn.

1. Nhận Xét Tổng Quan

# **Tài Liệu Tham Khảo**

1. Tài liệu môn học Data Mining của TS. Nguyễn Tiến Đạt**.**