ỦY BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**



BÁO CÁO ĐỒ ÁN CUỐI KỲ

HỌC PHẦN: PHÂN TÍCH DỮ LIỆU

**ĐỀ TÀI: Phân tích ảnh hưởng của một số yếu tố đến kết quả học tập của sinh viên hệ CLC khoá 21 tại Đại học Sài Gòn**

THÀNH VIÊN:

Lê Minh Phúc – 3119411051-DCT119C1

Nguyễn Công Thọ – 3119411077-DCT119C1

Lê Công Minh - 3119411037-DCT119C2

GVHD: PGS.TS Nguyễn Tuấn Đăng

NĂM HỌC: 2023-2024/HK2

TP. HỒ CHÍ MINH, THÁNG 4 NĂM 2024

MỤC LỤC

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH ẢNH, SƠ ĐỒ

DANH MỤC CHỮ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| CLC | : | Chất lượng cao |
| GPA | : | Grade Point Average |
| csv | : | Comma separated value |

MỞ ĐẦU

1. Lý do chọn đề tài

Giáo dục đại học là một trong những vấn đề cấp thiết trong việc đào tạo nguồn nhân lục chất lượng cao nhằm phục vụ cho việc công nghiệp hoá hiện đại hoá đất nước trong thời đại Cách mạng công nghiệp 4.0 hiện nay. Trong đó, chương trình đào tạo hệ Chất lượng cao (CLC) là một trong những vấn đề cần được quan tâm vì các sinh viên hệ CLC nhận được những chế độ chính sách đặc biệt về phương pháp học tập, môi trường đào tạo và mục tiêu đào tạo. Vì thế, nhóm chúng em chọn đề tài này với lý do nhằm hiểu rõ những yếu tố ảnh hưởng đến kết quả học tập của sinh viên hệ CLC. Qua đó, có thể đưa ra các biện pháp hỗ trợ kịp thời và phù hợp giúp cho sinh viên đạt được kết quả tốt nhất.

1. Khái quát về đề tài

Đề tài nghiên cứu tập trung vào kết quả học tập vào năm nhất và năm hai của một số sinh viên thuộc khoá 21 ngành Công nghệ thông tin hệ đào tạo Chất lượng cao tại trường Đại học Sài Gòn cùng một số yếu tố ảnh hưởng như sau:

* Quãng đường từ nhà đến trường
* Thu nhập do việc làm thêm
* Số lượng bạn bè
* Số lượng anh chị em trong nhà

1. Mục đích nghiên cứu

Mục đích nghiên cứu của đề tài gồm:

* Xác định các yếu tố ảnh hưởng đến kết quả học tập của sinh viên hệ CLC khóa 21 tại Đại học Sài Gòn.
* Đánh giá mức độ ảnh hưởng của từng yếu tố.
* Đề xuất các biện pháp hỗ trợ phù hợp để giúp sinh viên hệ CLC nâng cao kết quả học tập.

1. Phạm vi nghiên cứu

Đề tài nghiên cứu được thực hiện trên đối tượng là các sinh viên khoá 21 ngành Công nghệ thông tin thuộc hệ đào tạo CLC đang theo học tại Đại học Sài Gòn và một số yếu tố ảnh hưởng đến kết quả học tập được đề cập tại mục 2

1. Phương pháp nghiên cứu

Một số phương pháp nghiên cứu được áp dụng trong đề tài:

* Phương pháp khảo sát và phỏng vấn được áp dụng trong bước thu thập và xây dựng bộ dữ liệu
* Phương pháp thống kê được áp dụng trong bước phân tích dữ liệu nhằm đưa ra kết quả phân tích cùng nhận xét đánh giá mức độ ảnh hưởng đến kết quả học tập của các yếu tố được nêu.

1. Cấu trúc của đề tài

Báo cáo của đề tài gồm 5 phần:

* Phần mở đầu
* Chương 1: Giới thiệu bộ dữ liệu
* Chương 2: Xử lý dữ liệu
* Chương 3: Nhận xét đánh giá
* Phần kết luận

CHƯƠNG 1: GIỚI THIỆU BỘ DỮ LIỆU

1. GPA – tiêu chí đánh giá kết quả học tập

* GPA (Grade Point Average – điểm trung bình các môn học) là chỉ số được các trường đại học sử dụng một cách phổ biến trong việc đánh giá kết quả học tập của sinh viên trong suốt quá trình học tập tại trường. Chỉ số GPA có thể được đánh giá theo toàn khoá học, một học kỳ hay một năm học.
* Chỉ số GPA được xác định bằng cách tính trung bình cộng tích số điểm của các môn đã học nhân với hệ số tín chỉ của từng môn học, sau đó chia cho tổng số tín chỉ của các môn đã học (theo công thức (1))

(1)

Trong đó:

p: số điểm của môn học

t: số tín chỉ của môn học

* Để có thể hiểu rõ hơn về chỉ số GPA thì ta có ví dụ từ bảng số liệu từ Bảng 1:

Bảng 1. Bảng số liệu ví dụ điểm số của một số môn học dùng để tính GPA của một sinh viên

|  |  |  |
| --- | --- | --- |
| Môn học | Số điểm | Số tín chỉ |
| A | 9 | 3 |
| B | 8 | 2 |
| C | 7 | 3 |
| D | 10 | 4 |
| E | 9 | 3 |

Với số liệu từ Bảng 1, ta có thể tính chỉ số GPA như sau:

* Mức điểm GPA thường được sử dụng ở thang điểm 4 (0.0 đến 4.0) nhằm để đánh giá thành tích học tập cụ thể:
  + 0.0: Mức kém (F)
  + 1.0: Mức yếu (D)
  + 2.0: Mức trung bình (C)
  + 3.0: Mức tốt (B)
  + 4.0: Mức xuất sắc (A)
* Theo quy định của Đại học Sài Gòn thì kết quả học tập được tính theo 2 loại GPA là GPA thang điểm 10 và GPA thang điểm 4 (theo thang quy đổi bảng 2)

Bảng 2. Bảng số liệu quy đổi điểm số GPA thang điểm 10 sang thang điểm 4 theo quy định của Đại học Sài Gòn

|  |  |  |
| --- | --- | --- |
| Thang điểm 10 | Thang điểm 4 | Mức điểm |
| < 4.0 | 0.0 | F |
| >=4,0 và <5.5 | 1.0 | D |
| >=5.5 và <7.0 | 2.0 | C |
| >=7.0 và <8.5 | 3.0 | B |
| >=8.5 | 4.0 | A |

1. Thu thập dữ liệu

* Trong đề tài này, dữ liệu được thu thập trong phạm vi khoá 21 ngành Công nghệ thông tin hệ đào tạo chất lượng cao (CLC) đang học tập tại trường Đại học Sài Gòn. Có thể thu hẹp phạm vi nghiên cứu bằng cách thu thập dữ liệu của các sinh viên trong một lớp học nằm trong phạm vi nghiên cứu cụ thể là lớp DCT121C3
* Nhóm chúng em đã thu thập một số thông tin liên quan đến kết quả học tập trong khoảng thời gian năm nhất và năm hai của các sinh viên thuộc lớp DCT121C3 cùng với một số thông tin về các yếu tố ảnh hưởng đến kết quả học tập của họ, cụ thể:
  + Các thông tin cá nhân của sinh viên (gồm mã sinh viên, tên và lớp): Các thông tin này là các thông tin cá nhân nhạy cảm có thể từ chối cung cấp
  + Kết quả học tập trong năm nhất và năm hai của sinh viên: được tính theo GPA (thang điểm 10) tích luỹ tương ứng năm nhất và năm 2
  + Một số thông tin liên quan đến các yếu tố ảnh hưởng đến kết quả học tập của sinh viên
    - Khoảng cách từ nhà đến trường (tính bằng km)
    - Thu nhập do làm thêm của sinh viên (tính theo triệu đồng)
    - Số anh chị em của sinh viên
    - Số bạn bè thân thiết của sinh viên
* Hình thức thu thập dữ liệu: Dữ liệu được thu thập dưới dạng bảng câu hỏi và thu thập bằng công cụ Google Form. Đồng thời kết quả thu thập được lưu trong một bảng tính Google Spreadsheet và có thể tải về dưới dạng file csv (comma separated value)

CHƯƠNG 2: XỬ LÝ DỮ LIỆU

1. Công cụ và môi trường thực nghiệm
2. Công cụ thực nghiệm

* Công cụ lưu trữ dữ liệu: Bộ dữ liệu được lưu trong một tập tin csv (các giá trị trong tập tin được phân tách bằng dấu phẩy) vì tập tin csv có cấu trúc đơn giản dễ sử dụng được hỗ trợ bởi các công cụ phổ biến như pandas hay pyspark. Tuy nhiên, đối với dữ liệu phức tạp thì tập tin csv không được dùng để lưu trữ vì tập tin csv không có định dạng cụ thể nào. Vì vậy, với quy mô của một đồ án môn học, dữ liệu thường có cấu trúc tương đối đơn giản và có ít dữ liệu do đó tập tin csv là công cụ phù hợp để lưu trữ bộ dữ liệu này.
* Công cụ phân tích dữ liệu: pyspark là một interface của Apache Spark là một thư viện xử lý dữ liệu lớn mã nguồn mở mạnh mẽ. Nó cho phép bạn sử dụng ngôn ngữ lập trình Python để viết các chương trình Spark để xử lý và phân tích dữ liệu lớn trên cụm máy tính. Trong đề tài này, nhóm chúng em tận dụng một số tiện ích của pyspark như Spark DataFrame (trong việc xử lý dữ liệu dạng bảng) và Spark SQL (trong việc xử lý dữ liệu có cấu trúc) cùng với tính linh hoạt của nó trong việc hỗ trợ nhiều loại tập tin trong đó có dạng csv.
* Công cụ trực quan hoá dữ liệu: Thư viện Matplotlib của python là một công cụ trực quan hoá mã nguồn mổ tiêu chuẩn và phổ biến vì được sự hỗ trợ mạnh mẽ từ cộng đồng. Bên cạnh đó, Matplotlib còn có khả năng tuỳ chỉnh các thuộc tính của biểu đô một cách linh hoạt.

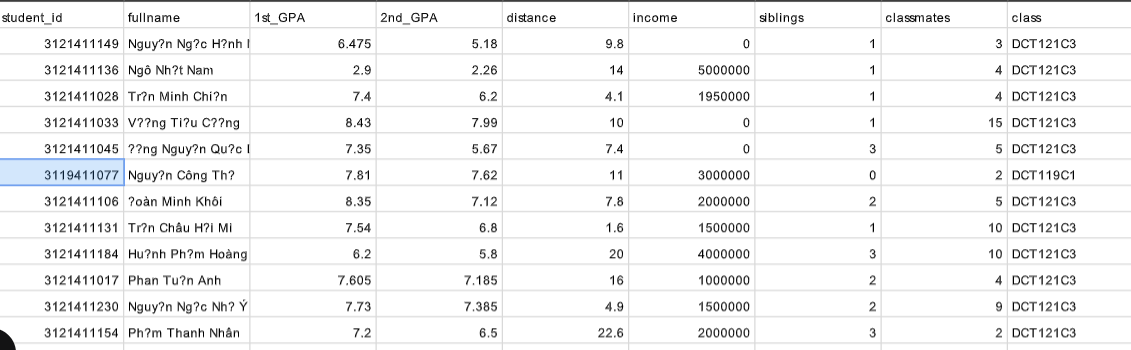
1. Môi trường thực nghiệm

* Nhóm chúng em sử dụng Google Colab để tiến hành thực hiện đề tài này cùng với việc lưu trữ bộ mã nguồn và dữ liệu trên Google Drive vì có thể chia sẻ cho đội nhóm và dễ dàng quản lý tiến độ của các thành viên thông qua lịch sử hoạt động của các thành viên được ghi nhận trên thư mục của dự án
* Một tiện ích khác khi sử dụng Google Colab là việc môi trường lập trình được hệ thống máy chủ cũa Google cài đặt sẵn hầu hết mọi thư viện hỗ trợ cơ bản vả có giao diện như một Jupyter notebook giúp cho việc lập trình và quan sát kết quả dễ dàng hơn trên một ô (coding cell) cũng như khả năng chú thích các tài liệu quan trọng trước khi lập trỉnh thông qua các markdown.
* Tuy nhiên, không phải thư viện nào của python cũng được Google cài đặt sẵn trong đó có thư viện pyspark vì vậy trước khi thực hiện chạy các ô code phía sau thì phải cài đặt pyspark bằng cú pháp

|  |
| --- |
| %pip install pyspark |

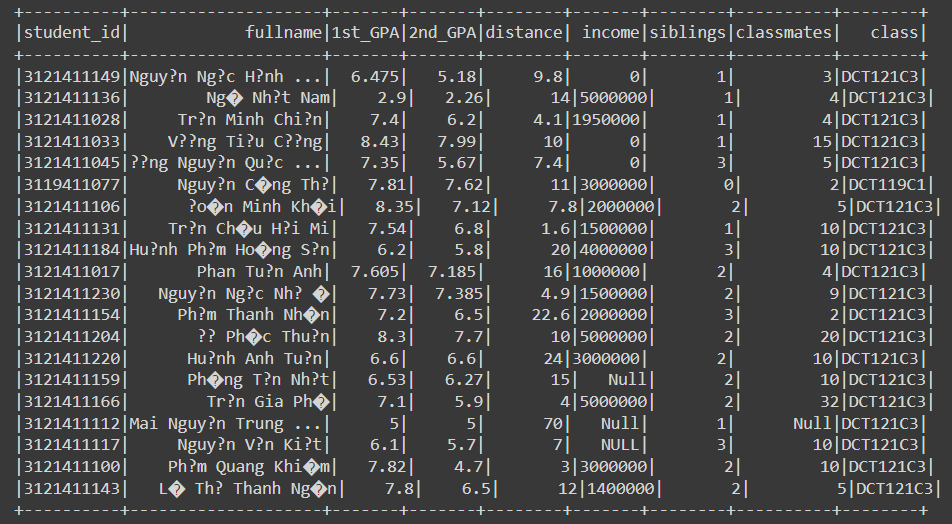
1. Tiền xử lý dữ liệu
2. Chỉnh sửa kiểu dữ liệu của bộ dữ liệu

* Như đã giới thiệu trong chương 1, bộ dữ liệu đã được thu thập là một tập tin csv gồm 9 trường dữ liệu (data fields) tượng trưng cho 9 cột và 31 thể hiện (data instance) tượng trưng cho 31 dòng dữ liệu



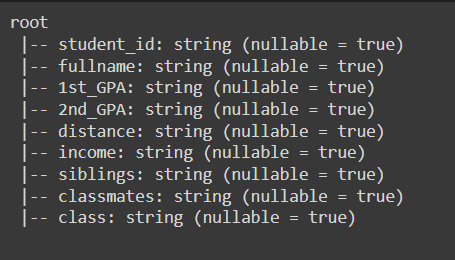
Hình 1. Hình ảnh ví dụ của một mẫu dữ liệu của bộ dữ liệu

* Theo hình ảnh minh hoạ từ Hình 1, bộ dữ liệu gồm các trường dữ liệu sau:
  + Student\_id: mã số sinh viên của sinh viên
  + Fullname: tên của sinh viên
  + 1st\_GPA: chỉ số GPA của sinh viên vào năm thứ nhất (tính theo thang điểm 10)
  + 2nd\_GPA: chỉ số GPA của sinh viên vào năm thứ hai (tính theo thang điểm 10)
  + Distance: khoảng cách từ nhà đến trường của mỗi sinh viên (tính theo km)
  + Income: thu nhập của sinh viên đến từ công việc làm thêm (tính theo VNĐ)
  + Siblings: Số anh chị em của sinh viên
  + Classmates: Số lượng bạn thân cùng lớp
  + Class: là lớp của sinh viên tham gia khảo sát



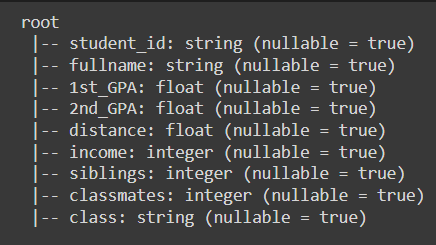
Hình 2. Dữ liệu được đọc bởi pyspark

* Một vấn đề gặp phải trong bộ dữ liệu (được thể hiện quan Hình 1 và Hình 2) là việc hiển thị tiếng Việt bị lỗi font chữ do lỗi mã hoá Unicode (UTF-8). Vì thế trường fullname có thể bị lược bỏ và các sinh viên có thể được quản lý nhờ vào mã số sinh viên và mã lớp của sinh viên.
* Tiếp theo là kiểu dữ liệu của các cột dữ liệu (schema) được ghi nhận nhờ phương thức printSchema() của pyspark



Hình 3. Kiểu dữ liệu ban đầu của các Schema trong bộ dữ liệu

* Theo hình 3, các schema được mặc định kiểu dữ liệu là kiểu chuỗi (string) mặc dù trong bảng dữ liệu (dataframe) còn có các cột mang kiểu dữ liệu là số nguyên hoặc số thực. Vì thế việc xử lý đổi kiểu dữ liệu là việc cần thiết nhằm phục vụ chp việc phân tích dữ liệu. Việc xử lý đổi kiểu dữ liệu được thể hiện qua Hình 4.



Hình 4. Hình ảnh các schema được chỉnh sửa kiểu dữ liệu

1. Xử lý dữ liệu khuyết thiếu (Handling missing data)

* Dữ liệu khuyết thiếu là một vấn đề phổ biến trong việc phân tích dữ liệu. Dữ liệu có thể bị khuyết thiếu vì một số nguyên nhân:
  + Lỗi thu thập dữ liệu
  + Lỗi nhập liệu
  + Người khảo sát từ chối cung cấp dữ liệu trong quá trình thu thập dữ liệu (có thể do dữ liệu cá nhân nhạy cảm nên từ chối cung cấp)
* Việc xử lý dữ liệu khuyết thiếu là rất quan trọng vì nó có thể ảnh hưởng đến độ chính xác và hiệu quả của các phân tích
* Có nhiều phương pháp khác nhau để xử lý dữ liệu khuyết thiếu. Phương pháp phù hợp nhất sẽ phụ thuộc vào loại dữ liệu khuyết thiếu, nguyên nhân gây ra dữ liệu khuyết thiếu và mục đích sử dụng dữ liệu. Một số phương pháp phổ biến bao gồm:
  + Xoá bỏ dòng dữ liệu khuyết thiếu
  + Gán giá trị
  + Dự đoán giá trị bằng máy học

1. Mã hoá dữ liệu cá nhân nhạy cảm
2. Tính điểm trung bình GPA và đánh giá xếp loại sinh viên

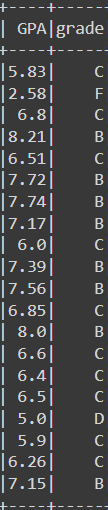
* Về việc đánh giá GPA của các sinh viên qua GPA của năm nhất và GPA của năm thứ hai thì nhóm chúng em sử dụng phương pháp tính giá trị trung bình của GPA trong hai năm học thông qua công thức (2)

Trong đó:

* GPA là giá trị GPA trung bình trong 2 năm học của sinh viên
* A là điểm GPA của sinh viên trong năm nhất
* B là GPA của sinh viên trong năm thứ hai

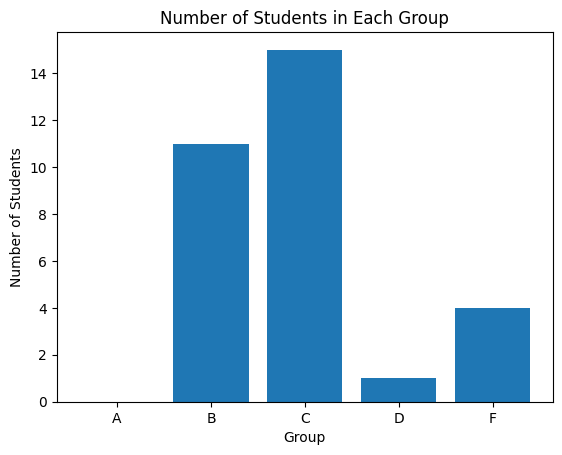
Lưu ý: Kết quả của việc tính điểm GPA được làm tròn 2 chữ số thập phân

* Khi đã tính được GPA của từng sinh viên, sau đó nhóm chúng em đã tiến hành đánh giá dựa trên mức thành tích (dựa theo thông tin của Bảng 2) thì nhóm chúng em có được thành tích của một số sinh viên tương ứng (theo hình 5)



Hình 5. Điểm số GPA và xếp loại theo mức thành tích của một số sinh viên

* Từ một phần bảng số liệu trích ra từ bộ dữ liệu (như hình 5), nhóm chúng em đã thống kê bao nhiêu sinh viên đạt hạng A, B, C, D và F tương ứng và đã trực quan hoá dưới dạng biểu đồ cột (như hình 6)



Hình 6. Hình ảnh minh hoạ số lượng sinh viên duo95c xếp theo các mức thành tích

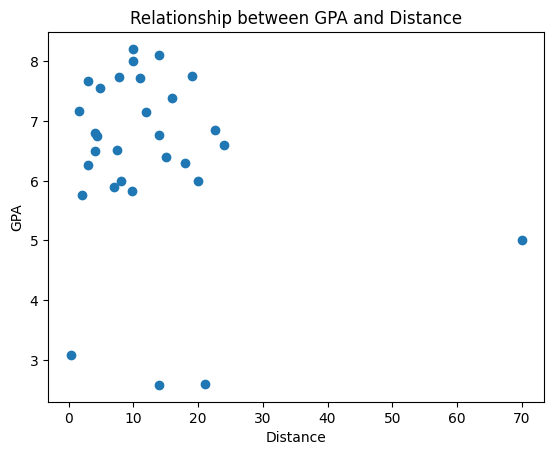
* Với biểu đồ cột như hình 6, nhóm chúng em đã thống kê được các nhóm sinh viên được chia theo thành tích học tập như bảng 3:

Bảng 3. Bảng số liệu thành tích học tập của sinh viên

|  |  |
| --- | --- |
| Thành tích học tập | Số lượng sinh viên |
| A | 0 |
| B | 11 |
| C | 15 |
| D | 1 |
| F | 4 |
| Tổng cộng | 31 |

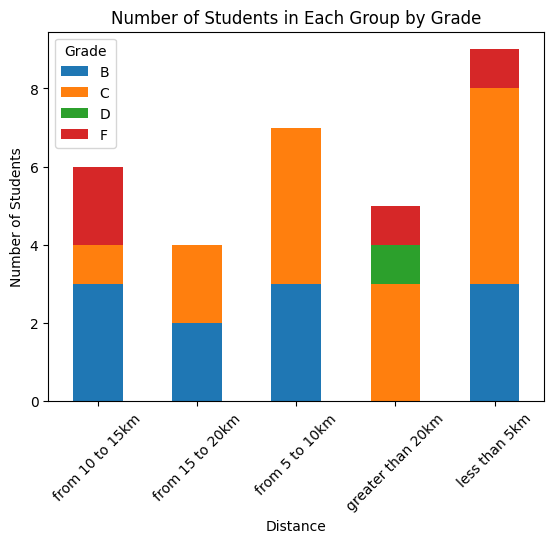
1. Phân tích một số yếu tố ảnh hưởng đến kết quả học tập của sinh viên
2. Phân tích sự ảnh hưởng của khoảng cách từ nhà đến trường đến thành tích học tập

* Mối liên hệ giữa khoảng cách từ nhà đến trường và thành tích học tập của sinh viên là một mối quan hệ phức tạp với nhiều yếu tố cần suy xét. Trong đó có hai giả thuyết liên quan:
  + Giả thuyết 1: Khoảng cách từ nhà đến trường ảnh hưởng tiêu cực đến thành tích học tập của sinh viên. Quãng đường đi học của sinh viên càng xa thì thể lực và tinh thần của sinh viên sẽ suy giảm ảnh hưởng đến sự tập trung của chính sinh viên trong việc học tập và tham gia các hoạt động ngoại khoá
  + Giả thuyết 2: Khoảng cách từ nhà đến trường tuy có tác động tiêu cực đến thành tích học tập của sinh viên nhưng không phải là tất cả. Một số sinh viên ở xa trường có thể rèn luyện cho mình một sự kỷ luật mạnh mẽ cùng với kỹ năng quản lý thời gian một cách chặt chẽ và khả năng tự lập của sinh viên. Đồng thời, việc ở xa trường sẽ cho sinh viên có động lực để thành công hơn vì bản thân họ đã nhận thức được rằng họ cần sự nỗ lực nhiều hơn để dat được mục tiêu của mình.
* Trong bộ dữ liệu đã thu thập thì mỗi sinh viên sẽ có một thành tích riêng và khoảng cách từ nhà đến trường của họ cũng khác nhau (được mô tả theo biểu đồ hình 7)



Hình 7. Biểu đồ mô tả sự phân bố sinh viên dựa vào thành tích và khoảng cách từ nhà đến trường

* Theo hình 7 thì ta có xu hướng chủ yếu của sinh viên được khảo sát có khoảng cách từ nhà đến trường là khoảng 20km. Tuy nhiên vẫn có ngoại lệ là một sinh viên ở cách trường đến 70km
* Xét về mặt thành tích học tập của các sinh viên được khảo sát theo thông tin mà hình 7 cung cấp cho thấy thành tích của sinh viên thường tập trung ở mức 5,5 đến hơn mức 8 một chút. Đồng thời cũng phát hiện 4 sinh viên có thành tích học tập dưới mức 5
* Khi nhận xét kết hợp giữa thành tích học tập và khoảng cách từ nhà đến trường thì ta có thể nhận thấy các sinh viên ở nhà tương dối gần (mức dưới 15km) thì thành tích học tập tương đối cao thậm chí có sinh viên có mức điểm trên 8 và các sinh viên có khoảng cách nhà xa hơn thì có khoảng cách kém hơn (có thể ví dụ một sinh viên phải đi 70km để đi học thì có thành tích ở mức dưới 5). Tuy nhiên, chúng ta vẫn có một ngoại lệ khác là ba bạn sinh viên có thành tích dưới mức 5 nhưng khoảng cách từ nhà đền trường chỉ khoảng 20km.
* Để có thể hiểu rõ hơn mối quan hệ giữa thành tích học tập thì nhóm chúng em đã thống kê chia nhóm lại lại cho các sinh viên theo cách như sau:
  + Đầu tiên, nhóm chúng em sẽ chia khoảng cách từ nhà đến trường thành nhiều nhóm có quãng đường theo cột mốc cụ thể:
    - Nhóm nhà gần (dưới 5km)
    - Nhóm nhà tương đối gàn (từ 5 đến 10km)
    - Nhóm nhà trung bình (từ 10 đến 15km)
    - Nhóm nhà tương đối xa (từ 15 đến 20km)
    - Nhóm nhà xa (nhiều hơn 20km)
  + Đồng thời, về thành tích học tập vẫn chia thành các mốc A, B, C, D và F dựa theo thông tin từ bảng 2 và bảng 3
  + Khi đã có dữ liệu thống kê thì nhóm chúng em sử dụng biểu đồ cột chồng (stacked bar plot) để vẽ biểu đồ thống kê có bao nhiêu sinh viên theo từng nhóm khoảng cách từ nhà đến trường và đếm cụ thể có bao nhiêu sinh viên A, B, C, D và F trong từng nhóm (được mô tả trong hình 8)



Hình 8. Biểu đồ cột chồng mô tả việc thống kê số lượng sinh viên theo từng nhóm khoảng cách từ nhà đến trường

* Theo hình 8, biểu đồ hiển thị số lượng sinh viên được xếp theo các nhóm khoảng cách từ nhà đến trường. Đồng thời cũng đếm số sinh viên trong mỗi nhóm theo thành tích học tập (dữ liệu đã được thống kê theo bảng 4)

Bảng 4. Dữ liệu thống kê số sinh viên theo nhóm khoảng cách từ nhà đến trường và thành tích học tập

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Khoảng cách  Số sinh viên | Dưới 5km | Từ 5 đến 10km | Từ 10 đến 15km | Từ 15 đến 20km | Xa hơn 20km | Tổng cộng |
| Mức điểm A | 0 | 0 | 0 | 0 | 0 | 0 |
| Mức điểm B | 3 | 3 | 3 | 2 | 0 | 11 |
| Mức điểm C | 5 | 4 | 1 | 2 | 3 | 15 |
| Mức điểm D | 0 | 0 | 0 | 0 | 1 | 1 |
| Mức điểm F | 1 | 0 | 2 | 0 | 1 | 4 |
| Tổng cộng | 8 | 7 | 6 | 4 | 5 | 31 |

* Theo dữ liệu từ bảng 4, ta có 8 sinh viên đi học với khoảng cách dưới 5km, 7 sinh viên với quãng dường từ 5 đến 10km, 6 sinh viên có khoảng cách từ 15 đến 20km, 4 sinh viên có quãng đường đi học từ 15 đến 20km và 5 sinh viên có khoảng cách từ nhà dến trường xa hơn 20km.
* Về mặt thành tích học tập thì mức điểm cao nhất mà các sinh viên có thể đạt được là B (không có một sinh viên nào dạt mức điểm A). Đồng thời, ta có thể nhìn thấy được mức điểm B đều có mặt ở 4 nhóm quãng đường đi học gần hơn mà không có mặt sinh viên nào ở nhóm nhà xa trường hơn 20km. Mặt khác, ta thấy ở mức điểm D thì không có sinh viên nào thuộc 4 nhóm có khoảng cách nhà gần trong khoảng 20km và có 1 thuộc nhóm nhà xa hơn 20km. Điều này có thể nói lên được rằng giả thuyết nhà càng ở xa thì thể lực và khả năng tập trung cho việc học sẽ kém hơn (có thể chứng minh thêm bằng một dẫn chứng khác là các nhóm nhà có khoảng cách trong khoảng 20km thì thành tích cao nhất là B còn nhóm nhà xa hơn 20km thì thành tích cao nhất là C)
* Khi xét đến nhóm mức điểm F thì có 4 sinh viên có mặt tại 3 nhóm:
  + Nhóm gần hơn 5km
  + Nhóm từ 10 đến 15km
  + Nhóm xa hơn 20km
* Như vậy thì việc sinh viên học kém trong nhóm mức điểm F này có thể không liên quan đến việc nhà xa hay gần mà phụ thuộc vào các yếu tố khác ảnh hưởng đến chất lượng học tập của sinh viên trong đó có vấn đề ý thức kỷ luật của chính bản thân sinh viên

1. Phân tích sự ảnh hưởng của thu nhập do làm thêm đến thành tích học tập
2. Phân tích sự ảnh hưởng của số lượng anh/chị/em trong nhà đến thành tích học tập
3. Phân tích sự ảnh hưởng của số lượng bạn bè đến thành tích học tập

CHƯƠNG 3

KẾT LUẬN

PHỤ LỤC

Bộ mã nguồn chương trình

* Kết nối Google Drive

|  |
| --- |
| from google.colab import drive  drive.mount('/content/drive')  base\_path = '/content/drive/MyDrive/PTDL/data\_and\_code/final\_assignment' |

* Cài đặt pyspark

|  |
| --- |
| %pip install pyspark |

* Tạo Spark session

|  |
| --- |
| from pyspark.sql import SparkSession  def spark\_session(app\_name):  spark = SparkSession.builder \  .appName(app\_name)\  .getOrCreate()  # To make sure that we are working with the correct version of Spark  print("Running PySpark version: ", spark.version)  return spark  spark = spark\_session('assignment1') |

* Đọc dữ liệu từ file csv

|  |
| --- |
| dataset\_path='/content/drive/MyDrive/PTDL/data\_and\_code/final\_assignment /final\_assignment\_dataset.csv' #dataset path string  data = spark.read.csv(dataset\_path, header=True) |

* Đếm số dòng dữ liệu

|  |
| --- |
| # Count the number of rows in the DataFrame  row\_count = data.count()  # Print the number of rows  print("Number of rows in the DataFrame:", row\_count) |

* In lược đồ cấu trúc (schema)

|  |
| --- |
| data.printSchema() |

* Chỉnh sửa kiểu dữ liệu của các schema

|  |
| --- |
| from pyspark.sql.functions import col  data = data.withColumn("1st\_GPA", col("1st\_GPA").cast("float"))  data = data.withColumn("2nd\_GPA", col("2nd\_GPA").cast("float"))  data = data.withColumn("distance", col("distance").cast("float"))  data = data.withColumn("income", col("income").cast("integer"))  data = data.withColumn("siblings", col("siblings").cast("integer"))  data = data.withColumn("classmates", col("classmates").cast("integer")) |

* Gán các giá trị thuộc ô dữ liệu khuyết thiếu thành giá trị bằng 0

|  |
| --- |
| from pyspark.sql.functions import when  # Chuyển đổi giá trị null trong cột "income" thành 0  data = data.withColumn("income", when(data["income"].isNull(), 0).otherwise(data["income"]))  data = data.withColumn("classmates", when(data["classmates"].isNull(), 0).otherwise(data["classmates"]))  data.show() |

* Ứng dụng mã hoá AES mã hoá các thông tin cá nhân của sinh viên

|  |
| --- |
| %pip install pycryptodome  from Crypto.Cipher import AES  from Crypto.Util.Padding import pad, unpad  from Crypto.Random import get\_random\_bytes  import base64  from pyspark.sql.functions import udf  from pyspark.sql.types import StringType  # Khởi tạo khóa bí mật  key = b'n\x1ak\x06{\x12\xeb\xe3\x19=\x91d\xad\xa7\xaeY' # 128-bit (16 bytes) key  # Hàm để mã hóa dữ liệu  def encrypt\_data(data, key):  cipher = AES.new(key, AES.MODE\_CBC)  ct\_bytes = cipher.encrypt(pad(data.encode(), AES.block\_size))  iv = base64.b64encode(cipher.iv).decode('utf-8')  ct = base64.b64encode(ct\_bytes).decode('utf-8')  return iv + "|" + ct  # Hàm UDF để áp dụng mã hóa cho mỗi giá trị trong cột  encrypt\_udf = udf(lambda x: encrypt\_data(x, key), StringType())  # Áp dụng mã hóa cho các cột và thêm vào DataFrame  # Mã hóa các cột trên DataFrame hiện tại  df\_encrypted = data.withColumn("student\_id", encrypt\_udf(col("student\_id"))) \  .withColumn("fullname", encrypt\_udf(col("fullname"))) \  .withColumn("class", encrypt\_udf(col("class")))  # Hiển thị DataFrame sau khi mã hóa  df\_encrypted.show()  # Hàm để giải mã dữ liệu  def decrypt\_data(data, key):  iv, ct = data.split("|")  iv = base64.b64decode(iv)  ct = base64.b64decode(ct)  cipher = AES.new(key, AES.MODE\_CBC, iv)  pt = unpad(cipher.decrypt(ct), AES.block\_size)  return pt.decode('utf-8')  # Hàm UDF để áp dụng giải mã cho mỗi giá trị trong cột  decrypt\_udf = udf(lambda x: decrypt\_data(x, key), StringType())  # Áp dụng giải mã cho các cột đã mã hóa và thêm vào DataFrame  df\_decrypted = df\_encrypted.withColumn("student\_id", decrypt\_udf(df\_encrypted["student\_id"])) \  .withColumn("fullname", decrypt\_udf(df\_encrypted["fullname"])) \  .withColumn("class", decrypt\_udf(df\_encrypted["class"]))  # Hiển thị DataFrame sau khi giải mã  df\_decrypted.show()  # Lưu DataFrame đã được thay đổi thành tập tin CSV  output\_path = '/content/drive/MyDrive/PTDL/data\_and\_code/final\_assignment/final\_assignment\_dataset\_modified/'  df\_encrypted.write.csv(output\_path, header=True, mode="overwrite") |

* In thử mã số sinh viên của một sinh viên sau khi mã hoá

|  |
| --- |
| # In ra một dòng dữ liệu đã được mã hóa  df\_encrypted.select("student\_id").show(1, truncate=False)  # Giải mã dòng dữ liệu đó  print(decrypt\_data('HS9OhhjRbrRa2yaCbfrJVA==|DS0ki98W/Yap51SfBybtRA==',key)) |

* Tính điểm GPA của các sinh viên bằng cách áp dụng công thức (2) và làm tròn 2 số thập phân

|  |
| --- |
| from pyspark.sql.functions import round  GPA\_df = df\_encrypted.withColumn("GPA", round((col("1st\_GPA")+col("2nd\_GPA"))/2,2)) |

* Xếp loại sinh viên theo điều kiện của bảng 2

|  |
| --- |
| from pyspark.sql.functions import expr  GPA\_df = df\_encrypted.withColumn("GPA", round((col("1st\_GPA")+col("2nd\_GPA"))/2,2))  grade\_df = GPA\_df.withColumn("grade", expr("CASE WHEN GPA >= 8.5 THEN 'A' WHEN GPA < 8.5 AND GPA >=7 THEN 'B' WHEN GPA >= 5.5 AND GPA <7 THEN 'C' WHEN GPA >=4 AND GPA < 5.5 THEN 'D' ELSE 'F' END"))  grade\_df.show() |

* Vẽ biểu đồ cột mô tả số sinh viên được xếp theo các mức thành tích học tập

|  |
| --- |
| # prompt: visualize how many students in each group  # Failure group who has GPA under 5  # Satisfied group who has GPA from 5 to 7.9  # Good group who has GPA from 8 to 8.9  # Excellence group who has GPA greater or equal 9  # Count the number of students in each group  A\_count = grade\_df.filter(grade\_df["grade"] == 'A').count()  B\_count = grade\_df.filter(grade\_df["grade"] == 'B').count()  C\_count = grade\_df.filter(grade\_df["grade"] == 'C').count()  D\_count = grade\_df.filter(grade\_df["grade"] == 'D').count()  F\_count = grade\_df.filter(grade\_df["grade"] == 'F').count()  # Create a bar chart  labels = ["A","B","C","D","F"]  values = [A\_count, B\_count, C\_count, D\_count, F\_count]  plt.bar(labels, values)  # Set axis labels and title  plt.xlabel("Group")  plt.ylabel("Number of Students")  plt.title("Number of Students in Each Group")  # Display the bar chart  plt.show() |

* Vẽ đồ thị scatter mô tả mối quan hệ giữa thành tích học tập của sinh viên và khoảng cách từ nhà đến trường

|  |
| --- |
| import matplotlib.pyplot as plt  # Assuming 'df\_encrypted' is a Pandas DataFrame with columns 'GPA' and 'distance'  pdf = GPA\_df.toPandas()  # Extract GPA and distance values into separate lists  gpa\_values = pdf['GPA'].tolist()  distance\_values = pdf['distance'].tolist()  # Create a scatter plot  plt.scatter(distance\_values, gpa\_values)  # Set axis labels and title  plt.xlabel('Distance')  plt.ylabel('GPA')  plt.title('Relationship between GPA and Distance')  # Display the plot  plt.show() |

* Xếp các sinh viên thành các nhóm quãng đường từ nhà đến trường theo thứ tự tăng dần

|  |
| --- |
| distance\_groups =grade\_df.withColumn("distance\_group", expr("CASE WHEN distance <5 THEN 'less than 5km' WHEN distance >=5 AND distance <=10 THEN 'from 5 to 10km' WHEN distance > 10 AND distance <15 THEN 'from 10 to 15km' WHEN distance >= 15 AND distance < 20 THEN 'from 15 to 20km' ELSE 'greater than 20km' END")) |

* Vẽ biểu đồ cột chồng mô tả số sinh viên theo từng mức thành tích và quãng đường đi học của họ

|  |
| --- |
| # prompt: visualize stacked barplot to show how many students have grade A, B, C, D and F in each group distance less than 5km, distance from 5 to 10km, distance from 15 to 20 km and >20km  # Extract the required columns into Pandas DataFrame  pdf = distance\_groups.select('distance\_group', 'grade').toPandas()  # Group the data by distance and grade  grouped\_pdf = pdf.groupby(['distance\_group', 'grade']).size().unstack().fillna(0)  # Create stacked bar plot  grouped\_pdf.plot(kind='bar', stacked=True)  # Set axis labels and title  plt.xlabel('Distance')  plt.ylabel('Number of Students')  plt.title('Number of Students in Each Group by Grade')  plt.xticks(rotation = 45)  # Add legend  plt.legend(title='Grade')  # Display the plot  plt.show() |

Bảng phân công công việc

|  |  |  |
| --- | --- | --- |
| STT | Công việc | Người thực hiện |
| 1 | Liên hệ xây dựng bộ dữ liệu | Nguyễn Công Thọ |
| 2 | Tiền xử lý dữ liệu   * Chỉnh sửa kiểu dữ liệu * Xử lý dữ liệu khuyết thiếu * Mã hoá thông tin cá nhân nhạy cảm | Lê Công Minh |
| 3 | Phân tích sự ảnh hưởng của khoảng cách từ nhà đến trường đến kết quả học tập | Nguyễn Công Thọ |
| 4 | Phân tích sự ảnh hưởng của số lượng anh/chị/em đến kết quả học tập | Lê Minh Phúc |
| 5 | Phân tích sự ảnh hưởng của số lượng bạn bè đến kết quả học tập | Lê Minh Phúc |
| 6 | Phân tích sự ảnh hưởng của mức thu nhập do làm thêm đến kết quả học tập | Lê Công Minh |
| 7 | Viết báo cáo tổng kết | Tất cả thành viên |